UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
INSTITUTO DE MATEMÁTICA E ESTATÍSTICA

DEPARTAMENTO DE ESTATÍSTICA

BAYESIAN RR: UMA INTERFACE EM SHINY PARA O MODELO LOG-BINOMIAL VIA ABORDAGEM BAYESIANA

GABRIEL DA CUNHA

Porto Alegre
2016

GABRIEL DA CUNHA

BAYESIAN RR: UMA INTERFACE EM SHINY PARA O MODELO LOG-BINOMIAL VIA ABORDAGEM BAYESIANA

Trabalho de Conclusão de Curso submetido como requisito parcial para a obtenção do grau de Bacharel em Estatística.

Orientador Metodológico
Prof.ª. Dra. Vanessa Bielefeldt Leotti

Co-Orientador
Prof.ª Dra. Suzi Alvez Camey

Porto Alegre
2016

Instituto de Matemática e Estatística
Departamento de Estatística

Bayesian RR: uma interface em shiny para o modelo log-binomial via abordagem bayesiana

Gabriel da Cunha

Banca examinadora:

Prof.ª Dra. Vanessa Bielefeldt Leotti
IME/UFRGS

Prof.ª Dra. Suzi Alvez Camey
IME/UFRGS

Prof.ª Dra. Sídia Maria Callegari Jacques
IME/UFRGS

AGRADECIMENTOS

Minhas habilidades para a escrita são mais fortes em comentários sarcásticos ou trocadilhos, então elas perdem o poder depois de um parágrafo, assim esse trabalho surgiu com a ajuda direta ou indireta de algumas pessoas.

Como os guardiões da minha sanidade mental e provedores de momentos de alegria, aos meus *clusters* de amigos: *Handoff*, *Outliers* e Futucada, aqueles que ouviram sobre os aspectos subjetivos do processo com muita paciência: Rodrigo B. Cardoso, Vinícius Bampi e Fernanda Wagner, e àquela que me ajudou a definir a ideia, antes mesmo de sermos amigos: Paula Bracco, obrigado.

Também agradeço às mulheres que conheci no curso de estatística, que dominam e compatilham seu conhecimento com maestria, principalmente minha orientadora Dra. Vanessa Bielefeldt Leotti, que além de tudo me guiou nessas águas turvas, onde minha canoa de trocadilhos é furada. À Dra. Suzi Camey, que ajudou a tornar o projeto mais dinâmico e amigável para os usuários do programa, e à Dra. Márcia Barbian pela amizade e *inputs* gratuitos durante cafés.

Por último, obrigado à minha família, grande ou pequena, presente ou ausente. Em especial para minha irmã Flaviani da Cunha e prima Fernanda Bernardes, pelas ajudas textuais e ouvidos disponíveis.

Para minhas pessoas favoritas: Flávio, Rosinha, Mosa e Fá.

“Wiggle your big toe.”
(Beatrix Kiddo)

RESUMO

Entre as medidas de associação amplamente utilizadas em estudos epidemiológicos, o risco relativo é recomendado em relação à razão de chances. Através da abordagem frequentista, os métodos para estimar o risco relativo, como o modelo log-binomial ou o Poisson robusto, podem apresentar problemas de convergência ou produzir probabilidades acima de 1, no entanto, a abordagem bayesiana consegue ultrapassar esses obstáculos. Essa abordagem pode estar sendo subutilizada, pois é feita através da linguagem de programação em R, que necessita que o usuário tenha habilidades de programação. Neste trabalho, foi proposta a implementação de uma interface visual criada a partir do pacote Shiny, levando em consideração a contribuição de autores para estimar risco relativo por abordagem bayesiana. O Bayesian RR permite que, usuários sem conhecimento de programação, possam estimar o risco relativo para dados independentes com desfecho dicotômico. Ele é acompanhado por um guia passo-a-passo com os itens: carregar o banco, seleção do modelo, configuração do MCMC e resultados.

Palavras-chave: Risco relativo. MCMC. Bayesiana. Shiny. Log-binomial. Interface.

ABSTRACT

Among measures of association widely used in epidemiological studies, the relative risk is recommended over the odds ratio. Through the frequentist approach, the methods to estimate relative risk, like the log-binomial model and the robust Poisson, can result in convergence problems or produce probabilities greater than 1, however, the Bayesian approach can overcome those obstacles. This approach may be underused because it's done through the R programming language, which requires reasonable programing skills from the user. This paper proposes the implementation of such approach by creating of a visual interface build with the Shiny package and recent contributions of authors for the use of relative risk estimation by the Bayesian approach. The creation of the Bayesian RR allows that users without programming skills to estimate relative risk for independent data and binary outcomes. A step-by-step tutorial of its use (data upload, model selection, MCMC configuration and results) is presented also.

Keywords: Relative Risk. MCMC. Bayesian. Shiny. Log-binomial. Interface.

LISTA DE FIGURAS

SUMÁRIO

Este artigo será submetido à revista "*Journal of Biomedical Informatics*"

1 INTRODUÇÃO

Uma medida de associação amplamente utilizada em estudos epidemiológicos longitudinais é o risco relativo (RR), uma razão de probabilidades cuja interpretação é mais clara que a da razão de chances (RC), outra medida de associação comum em epidemiologia. DAVIES; CROMBIE; TAVAKOLI, 1998; LEE, 1994, entre outros autores recomendam o uso do RR ao invés da RC.

Modelos estatísticos alternativos à regressão logística, como o modelo log-binomial ou o modelo Poisson com variância robusta, podem ser utilizados para a estimação do RR. Utilizando a abordagem da estatística frequentista, entretanto, existe a possibilidade de não ocorrer convergência do modelo log-binomial (BARROS; HIRAKATA, 2003). Recomenda-se então utilizar o modelo Poisson (SPIEGELMAN; HERTZMARK, 2005), mas este ocasionalmente pode estimar probabilidades fora do intervalo 0 a 1 (BARROS; HIRAKATA, 2003).

A abordagem bayesiana como tentativa de ultrapassar os problemas de convergência do modelo log-binomial frequentista foi primeiramente apresentada por Chu e Cole (2010). Posteriormente, Torman e Camey (2015) apontaram vantagens da abordagem bayesiana, utilizando a linguagem de programação estatística R (R PROJECT, 2016), como solução unificadora para a estimação de RR em desfechos dicotômicos e politômicos. O método bayesiano necessita utilizar uma simulação do tipo *Markov Chain Monte Carlo* (MCMC) para aproximar a distribuição da posteriori. No entanto, programas tradicionalmente usados para MCMC, como WinBUGS, podem apresentar erros Monte Carlo grande, provenientes de autocorrelação ao gerar as simulações. Recentemente, Salmerón, Cano e Chirlaque (2015) propuseram uma rotina escrita puramente na linguagem R, que se apresentou eficaz na redução de erro de Monte Carlo e ainda na redução do tempo de processamento em comparação ao WinBUGS.

Apesar de estes avanços demonstrarem que a abordagem bayesiana é uma alternativa sólida e eficaz para o ajuste do modelo log-binomial, existe a necessidade de conhecimento específico de programação, como a linguagem R, o que pode ser um impedimento ao uso abrangente desse método. Pensando nessa barreira, a ideia central deste artigo é apresentar uma solução que diminua o espaço entre o usuário final, aquele que pode se beneficiar do uso desse tipo de análise, e a aplicação da técnica, através da criação de uma interface visual que

permita que, mesmo sem ter conhecimentos de programação, o usuário possa utilizar a solução bayesiana para estimar risco relativo por meio de um ambiente simples e intuitivo.

A construção desse aplicativo foi realizada principalmente usando o pacote `shiny` (CHANG, 2016a), idealizado para o desenvolvimento de aplicações em sintaxe R através da web. O pacote permite a construção de uma rotina que é traduzida para HTML, onde as interações com o usuário podem ser registradas como comandos que o console R utiliza para preencher o algoritmo. Esse algoritmo realiza a análise estatística sem que seja necessário programar o código diretamente, facilitando o processo de análise. Outro benefício apresentado pela utilização do pacote é a possibilidade de manter a interface em um repositório online, onde qualquer pessoa com um dispositivo compatível e internet possa utilizá-lo.

Neste artigo será detalhada a utilização do modelo log-binomial por meio da abordagem bayesiana, a proposta alternativa de MCMC em linguagem R, o pacote `shiny` e outros que compõem a construção da interface, assim como uma demonstração de como utilizar a aplicação desenvolvida com o objetivo de simplificar e difundir o uso do método bayesiano para estimar risco relativo.

2 LOG-BINOMIAL

O modelo log-binomial é um modelo linear generalizado com uma função de ligação log e resposta binomial. Inicialmente proposto por Wacholder (1986), o modelo surgiu como uma alternativa ao uso da regressão logística, com o objetivo de estimar o risco relativo, onde a variável dependente é o desfecho dicotômico de interesse.

O modelo log-binomial é representado por:

$$\theta_i = P(Y_i = 1 \mid X_{1i}, \ldots, X_{ki}) = exp(\beta_0 + \beta_1 X_{1i} + \ldots + \beta_k X_{ki}), i = 1, \ldots, n$$

onde Y_i é o desfecho binário para a i-ésima unidade, X_{1i}, …, X_{ki} são os valores dos k preditores, β_0 é o intercepto do modelo e β_1, ..., β_k são os coeficientes dos k preditores. O índice i varia de 1 até n, que é o tamanho da amostra, e exp(β_1), …, exp(β_k) são os RRs

associados a cada um dos preditores presentes no modelo. Apesar de atraente, este modelo pode apresentar problemas de convergência se a estimação dos parâmetros foi feita pela abordagem frequentista.

Encontram-se na literatura contribuições de autores como Torman e Camey (2015), onde é demonstrada a eficiência da abordagem bayesiana como maneira de contornar o problema de convergência do modelo log-binomial. A abordagem bayesiana é fundamentada nas distribuições de probabilidades dos parâmetros desconhecidos, onde informação anterior sobre esses parâmetros pode ou não ser incorporada através das distribuições à priori. O processo bayesiano utiliza a informação obtida em uma amostra observada para a função de verossimilhança e a priori, produzindo assim, a distribuição a posteriori, a partir da qual são feitas todas as inferências da análise. Como a distribuição a posteriori do modelo log-binomial não pode ser obtida de forma analítica, utilizam-se métodos computacionais como o MCMC para, de forma numérica, chegar à uma aproximação.

3 MCMC

A utilização de MCMC, para computar distribuições a posteriori apresenta uma solução para a barreira analítica frequentemente encontrada na abordagem bayesiana. No seu uso é definido um algoritmo para a geração de valores simulados que, após certo número de iterações, serão considerados como provenientes da posteriori de interesse. Neste artigo será exemplificado o uso do MCMC através de interpretações dos resultados gerados para mais detalhes sobre MCMC, recomenda-se a leitura de Gilks, Richardson e Spiegelhalter (1996).

Programas como o WinBUGS, ou sua variante em código aberto OpenBUGS, utilizam um esquema MCMC baseado em amostragem de Gibbs para as simulações. No entanto, para o modelo log-binomial, tal esquema pode gerar erros Monte Carlo grandes, onde os valores simulados apresentam autocorrelação que, ao final do processo, prejudicam a aproximação numérica. Partindo desse problema, Salméron, Cano e Chirlaque (2015) propuseram uma solução que visa diminuir o erro de Monte Carlo, consequentemente aumentando a precisão da aproximação. Para isso os autores sugerem uma reparametrização utilizando o modelo Poisson, onde os parâmetros originais β do modelo log-binomial são substituídos por outros θ cuja matriz de covariâncias é aproximadamente uma matriz identidade, reduzindo assim, a autocorrelação dos valores simulados. Além disso, foi criado um algoritmo MCMC específico, baseado em Metropolis-Hastings com distribuição Cauchy truncada, para simular

a posteriori, onde é levada em consideração a troca de parâmetros proposta. Através de testes com dados reais e simulados, os autores concluiram que a proposta apresentada era mais eficaz não apenas para a redução da autocorrelação mas também no tempo de processamento.

4 SHINY

O `shiny`, concebido em 2012, é um pacote que permite a construção de aplicações de web através da linguagem R, sem requerer qualquer tipo de conhecimento em HTML, *javascript* ou outras linguagens comumente utilizadas na internet. Apesar de ser relativamente recente, já existe uma gama de aplicativos e ferramentas estatísticas desenvolvidas que empregam o pacote e seus derivados, como o `shinydashboard` (CHANG, 2016b), que permite a criação de painéis estruturados em HTML.

Ele é utilizado principalmente como meio de visualização de dados ao vivo, onde o usuário pode manipular pequenas configurações e obter um retorno, geralmente por meio de gráficos ou tabelas, após as intervenções realizadas na interface. Diversos artigos exploram a implementação de mecanismos de visualização usando R e `shiny` em áreas como farmacologia (WOJCIECHOWSKI; HOPKINS; UPTON, 2015), psicologia (ELLIS; MERDIAN, 2015), agronomia (PERONDI et al, 2015) e análise geoespacial (JAHANSHIRI; SHARIFF, 2014).

Usufruindo de propriedades comuns na internet atual, como layout fluido ou responsivo, que permite que a interface seja adaptável ao formato do dispositivo físico em que é acessado, o pacote permite ampla implementação e utilização sem precisar de grandes arquivos, devido à simplicidade e compatibilidade do HTML e por ser totalmente concebido em sintaxe R, que é basicamente um arquivo de texto.

O aplicativo é composto de duas partes principais: a primeira é a interface para o usuário (*user interface - ui*), em que são definidas as propriedades de layout, podendo incluir arquivos externos, como imagens ou .CSS (*Cascading Style Sheets*) que são regras de formatação do HTML, e elementos do ambiente, como botões, menus, áreas para gráficos, áreas para textos e controles em geral manipulados pelo usuário. A segunda parte diz respeito aos processos que serão executados pelo servidor (*server*) com ligação ao console R; as sintaxes e algoritmos utilizados ao executar o programa podem ser carregados nesta seção. A aplicação dessas rotinas depende das definições do usuário feitas na interface. Por exemplo,

ao se escolher uma variável para a qual será criado um gráfico, é disparado um comando de mudança para o servidor que irá retornar o resultado para a área de visualização na interface.

Por padrão (*default*), todas as interações entre programa e usuário são automaticamente processadas. No entanto, dependendo do grau de complexidade do aplicativo, utiliza-se uma propriedade do pacote chamada *reactive*, que permite definir todas as ligações entre a ação, execução e a exibição dos resultados. A sua utilização permite que as ações sejam executadas de forma ordenada e respeitando os comandos do usuário, principalmente em casos onde o método vai além da demonstração de gráficos simples.

Outra vantagem da utilização do `shiny` é a possibilidade de publicar de forma gratuita o aplicativo desenvolvido. Para usuários cadastrados no *shinyapps.io*[1], é possível hospedar a sua interface no repositório permanente. Além disso, como o pacote é uma iniciativa do RStudio, pode-se realizar sua publicação de forma rápida e eficaz através do próprio programa.

5 BAYESIAN RR

O aplicativo Bayesian RR[2] é uma interface visual desenvolvida na linguagem R, utilizando os pacotes `shiny` e `shinydashboard` para a construção do ambiente e sua tradução para HTML. O projeto foi concebido para simplificar a maneira do usuário realizar uma análise bayesiana para estimar RR, de forma que sua utilização não necessita de domínio de programação. O processo de análise foi dividido em quatro estágios, sendo o primeiro o envio dos dados que serão analisados, seguido da construção do modelo pelo usuário, definição dos requisitos do MCMC e o armazenamento dos resultados obtidos.

Para compor a porção do aplicativo que realiza o MCMC, foi utilizado o script em R desenvolvido e disponibilizado por Salméron, Cano e Chirlaque (2015), com uma correção que substitui o modelo Poisson, descrito originalmente, pelo modelo Poisson com variância robusta. Com essa alteração foi necessário a inclusão do pacote `sandwich` (LUMLEY; ZEILEIS, 2015). Para a análise de diagnóstico do MCMC e sumário dos resultados, o pacote `coda` (PLUMMER et al., 2015) foi adicionado, enquanto o pacote `mcmcplots` (CURTIS et al, 2015) foi usada para a criação dos gráficos. Finalmente, para possibilitar que o usuário realize o *download* dos resultados obtidos, utilizou-se o pacote `rmarkdown` (ALLAIRE et

[1] Disponível em: <https://www.shinyapps.io/>.
[2] Disponível em: <http://goo.gl/Da1Ufr>.

al., 2016) e um arquivo .Rmd auxiliar para a construção do relatório que pode ser salvo tanto no formato .PDF como no formato .DOC. O relatório contém o modelo, definições do MCMC, gráficos e resultados gerados durante a sessão de uso do aplicativo.

Algumas medidas foram tomadas para incentivar o fluxo de uso da interface, dado que a navegação no menu lateral não segue uma ordem predefinida, algumas etapas só são permitidas caso o usuário tenha realizado o seu pré-requisito. Qualquer etapa do processo só pode ser realizada após o envio do banco de dados para o programa, assim como a execução da análise só é permitida após todos os parâmetros e configurações terem sido definidos. Essas medidas têm como objetivo evitar que o processamento seja iniciado com qualquer atributo vital faltando. Além disso, todas as etapas contam com um quadro auxiliar de ajuda sobre como utilizar aquele setor do ambiente.

O programa foi concebido para dados independentes com desfecho binário, sem a possibilidade de definir conhecimento anterior dos parâmetros, isto é, apenas com prioris não-informativas, e com geração de apenas uma cadeia MCMC. Essas limitações são provenientes da sintaxe de Salméron, Cano e Chirlaque (2015). Tendo em vista o vocabulário comum em estudos e aplicações de análises dessa natureza, assim como desejando maior alcance de sua utilização, todo o programa foi concebido em inglês.

Para ilustrar a utilização do programa, será apresentado um guia passo-a-passo, utilizando um banco de dados[3] adaptado de Hosmer e Lemeshow (1989). Este banco com 567 registros é referente à dados de crianças recém-nascidas e o objetivo é avaliar a associação entre nascer com baixo peso e vários confundidores e preditores protenciais. O banco contém 14 variáveis, detalhadas no anexo A, e o desfecho de interesse é a variável “bpn” (baixo peso ao nascer).

5.1 Carregar banco de dados

Na tela inicial, ou primeira aba do menu (*File Upload*), se faz o *upload* do arquivo do banco de dados. O formato aceito é o .CSV (*Comma-separated values*), escolhido pela simplicidade e alta compatibilidade. O arquivo deve ser uma tabela simples contendo as variáveis nas colunas e as observações nas linhas (Figura 1), sendo que o desfecho de interesse pode assumir apenas valores 0 ou 1 e os preditores categóricos devem ser em formato *string* ou codificados para a representação binária das categorias (variável *dummy*).

[3] Disponível em: <https://goo.gl/Aa6EAy>.

Recomenda-se que a codificação do arquivo seja UTF-8 (*8-bit Unicode Transformation Format*).

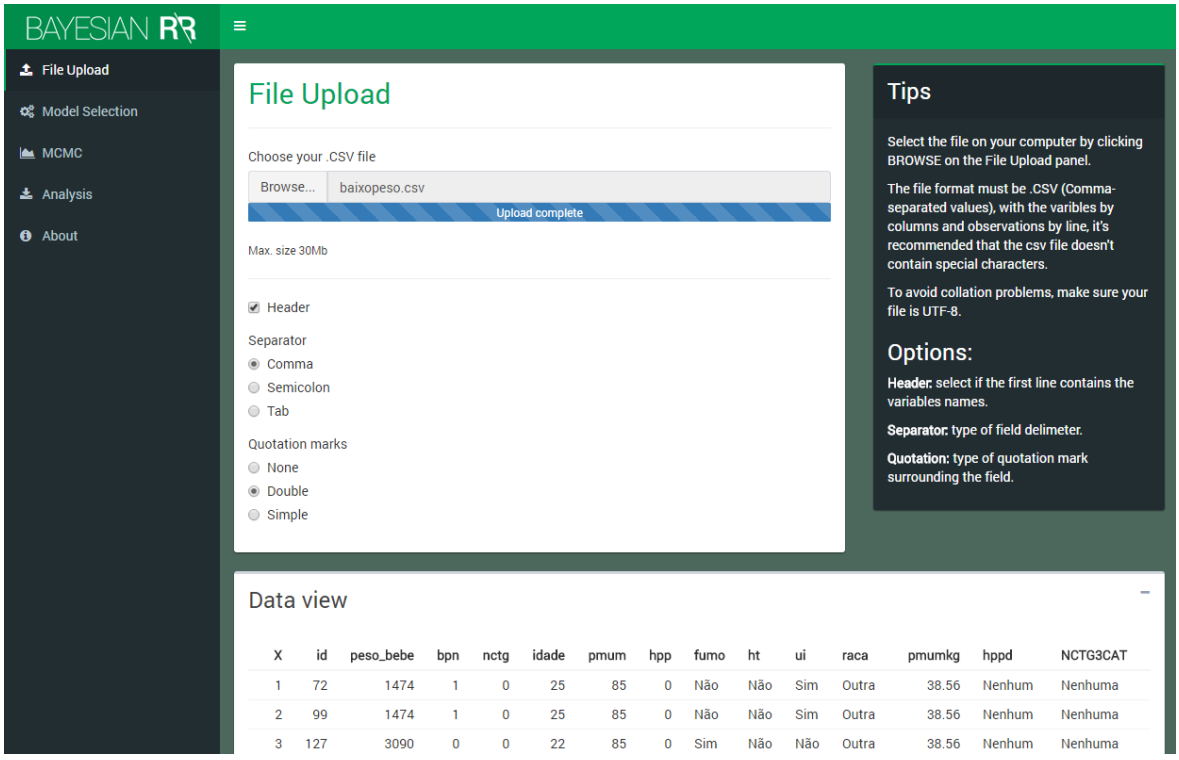
Figura 1. Banco de dados baixopeso.csv, primeiras e últimas linhas

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1		id	peso_beb	bpn	nctg	idade	pmum	hpp	fumo	ht	ui	raca	pmumkg	hppd	NCTG3CAT
2	1	72	1474	1	0	25	85	0	Não	Não	Sim	Outra	38.56	Nenhum	Nenhuma
3	2	99	1474	1	0	25	85	0	Não	Não	Sim	Outra	38.56	Nenhum	Nenhuma
4	3	127	3090	0	0	22	85	0	Sim	Não	Não	Outra	38.56	Nenhum	Nenhuma
5	4	158	1474	1	0	25	85	0	Não	Não	Sim	Outra	38.56	Nenhum	Nenhuma
6	5	337	3090	0	0	22	85	0	Sim	Não	Não	Outra	38.56	Nenhum	Nenhuma
564	563	122	3790	0	0	25	241	0	Não	Sim	Não	Negra	109.32	Nenhum	Nenhuma
565	564	190	3790	0	0	25	241	0	Não	Sim	Não	Negra	109.32	Nenhum	Nenhuma
566	565	522	3790	0	0	25	241	0	Não	Sim	Não	Negra	109.32	Nenhum	Nenhuma
567	566	32	3303	0	6	28	250	0	Sim	Não	Não	Outra	113.4	Nenhum	Duas ou mais
568	567	53	3303	0	6	28	250	0	Sim	Não	Não	Outra	113.4	Nenhum	Duas ou mais

Fonte: Elaborada pelo autor.

Para carregar o arquivo no programa, deve-se clicar no botão "*Browse*", aguardar a abertura de uma janela de navegação, escolher o arquivo desejado e clicar em "Abrir". Uma visualização será disponibilizada em "*Data View*" (Figura 2). Caso o arquivo pareça desconfigurado, é possível trocar as opções na leitura do arquivo, indicando se ele contém cabeçalho com os nomes das variáveis (*Header*) ou não, o tipo de separação das colunas - vírgula (*Comma*), ponto-e-vírgula (*Semicolon*) e espaçamento (*Tab*), e a marcação de *strings* - nenhum (*None*), aspas duplas (*Double*) e aspas simples (*Simple*).

Figura 2. Aba "*File Upload*" com banco carregado



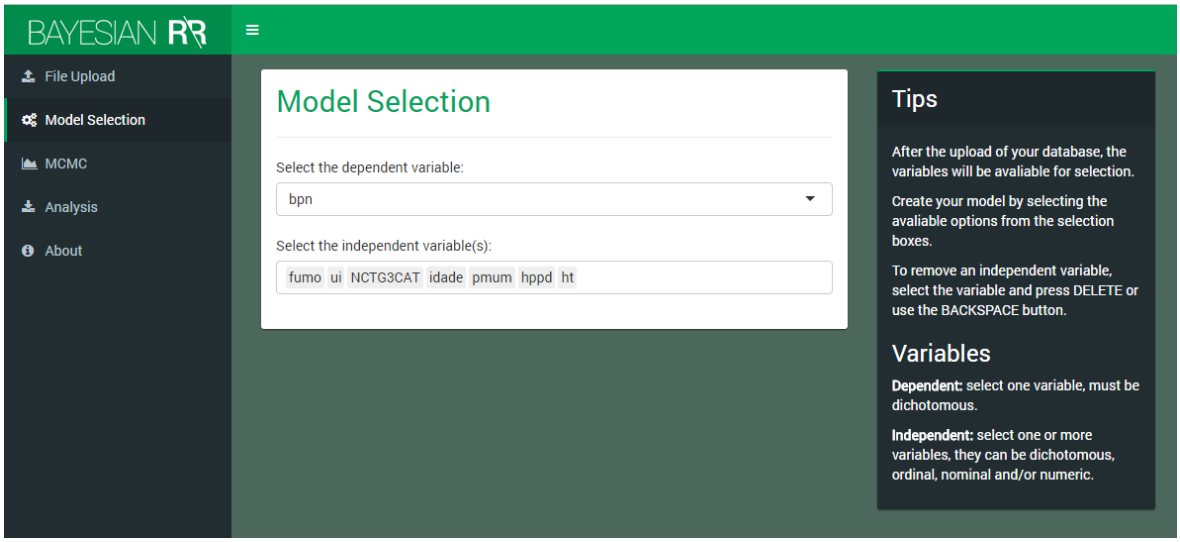
Fonte: Elaborada pelo autor.

5.2 Seleção do modelo

A segunda etapa a ser realizada é a definição do modelo na aba "*Model Selection*" (Figura 3). Após carregar o banco de dados no aplicativo, as variáveis automaticamente ficarão disponíveis para seleção, de acordo com as colunas do arquivo. Então se deve selecionar o desfecho dentre as opções que ficam listadas em uma caixa de seleção clicável. No caso do exemplo escolhemos a variável "bpn" que indica a presença (bpn=1) ou ausência (bpn=0) do baixo peso ao nascer.

Para a seleção das variáveis independentes, que irão compor o modelo, é utilizado o mesmo processo de seleçã, podendo ser selecionadas várias variávias. Caso seja necessário excluir algum dos preditores escolhidos, basta selecionar a varável no modelo e apertar a tecla "*Delete*" ou utilizar a tecla "*Backspace*". Neste exemplo, foram adicionadas as seguintes variáveis: "fumo", "ui", "NCTG3CAT", "idade", "pmum", "hppd" e "ht".

Figura 3. Aba "*Model Selection*" com variáveis selecionadas



Fonte: Elaborada pelo autor.

5.3 Configuração do MCMC

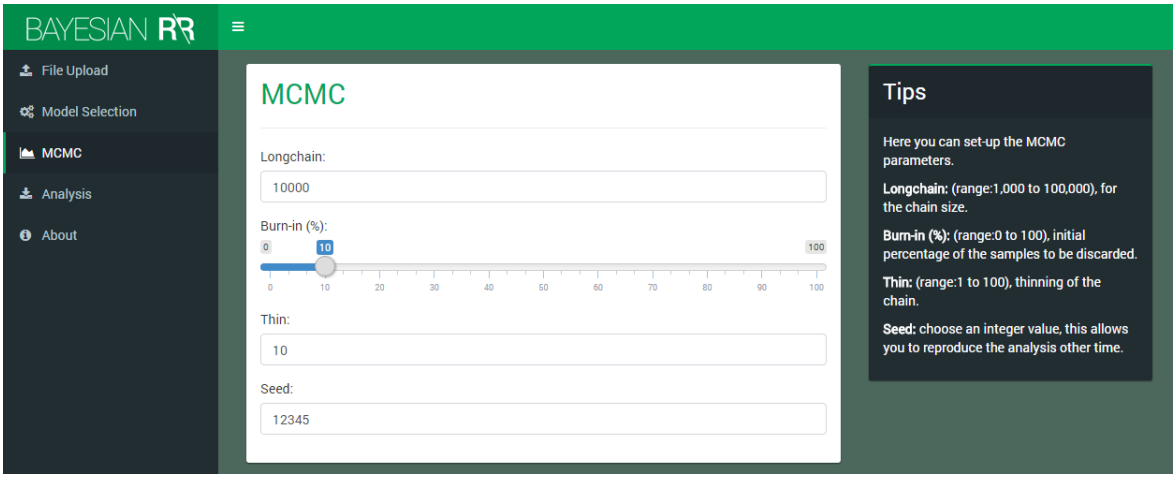
O último passo antes de realizar a análise é a configuração do MCMC. Na aba "MCMC" do aplicativo (Figura 4) é possível definir todos os itens necessários para a geração da cadeia de simulações. No campo *longchain* se define o tamanho da cadeia a ser executada, com valores variando entre 1.000 e 100.000. O *burn-in* indica o número de iterações iniciais que serão descartadas da cadeia para obtenção das estimativas. Esse item é utilizado para

remover valores iniciais da cadeia antes das operações para obter a convergência. Neste campo pode ser definida a porcentagem do *longchain* que será utilizada, deslizando o botão na régua disponível para tal.

Caso seja necessário, recomenda-se a utilização do *thin* para casos onde permanece algum grau de autocorrelação na cadeia. O *thin* pode ser definido entre 1 e 100, e implica que só será utilizado para obtenção de estimativas um valor a cada bloco de tamanho definido pelo *thin*. Por exemplo, se o *thin* for definido como 10, somente serão utilizados os valores da cadeia dos passos 10, 20, 30,... e assim por diante, sendo descartados os demais passos.

Nessa aba também é possível definir uma semente (*seed*) para a análise. O padrão é não ter uma semente definida durante a sessão do ambiente R, mas, ao indicar uma, é possível repetir as estimativas geradas na análise em outro momento, ou caso a sessão seja reiniciada, usando-se o mesmo valor como semente. Para esse campo pode ser digitado qualquer número inteiro positivo. Seguindo com o exemplo usamos a seguinte configuração do MCMC: *longchain* (10.000), *burn-in* (1.000), *thin* (10) e *seed* (12345).

Figura 4. Aba “MCMC” com parâmetros definidos



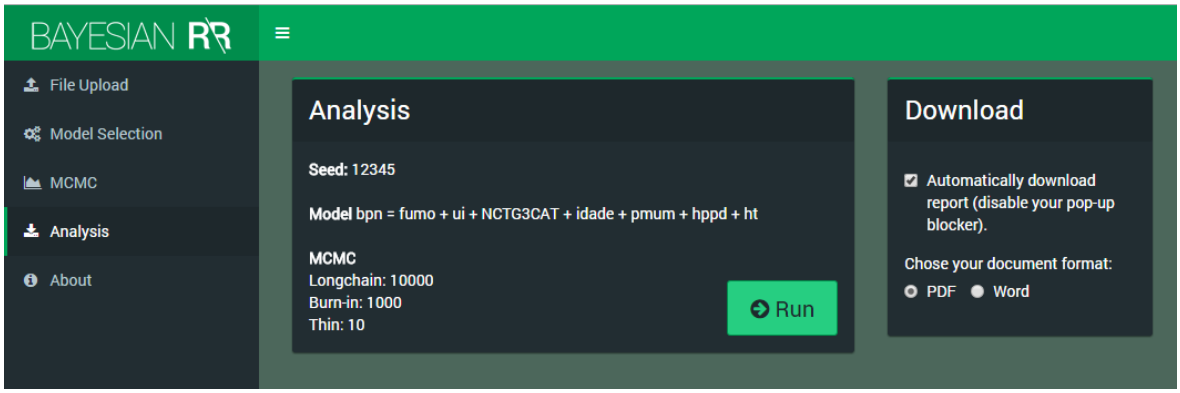
Fonte: Elaborada pelo autor.

5.4 Realização e resultados da análise

Na etapa final, ao selecionar a aba *“Analysis”*, pode-se conferir antes da análise o modelo escolhido na aba “*Model Selection”* e todas as definições da aba “MCMC” (Figura 5). Também é possível optar por realizar o download automático do relatório em dois formatos: .PDF e .DOC. Se essa opção estiver selecionada, o programa abrirá uma nova janela do navegador para salvar o relatório no formato preferencial. É importante que o bloqueador de

pop-up do navegador esteja desativado, permitindo que o download ocorra automaticamente. Se o download automático não estiver selecionado, haverá a opção para fazê-lo manualmente.

Figura 5. Aba "*Analysis*" antes de rodar análise



Fonte: Elaborada pelo autor.

Após a confirmação das configurações, basta clicar no botão "*Run analysis*" para iniciar a análise estatística. Enquanto estiver processando, o programa irá mostrar uma mensagem de espera no canto inferior direito da janela. No momento que a análise for finalizada, os resultados serão demonstrados na mesma página. Entre os resultados estão os seguintes gráficos de diagnóstico do MCMC: *trace* da cadeia, densidade e autocorrelação para os coeficientes do modelo. Também se encontram os tamanhos efetivos da cadeia, as estimativas pontuais e intervalos de credibilidade para os coeficientes e para os riscos relativos calculados. Além disso, são apresentados dois testes de convergência (Geweke e Heidelberg-Welch), que serão descritos na próxima seção. Estes resultados serão exemplificados a seguir.

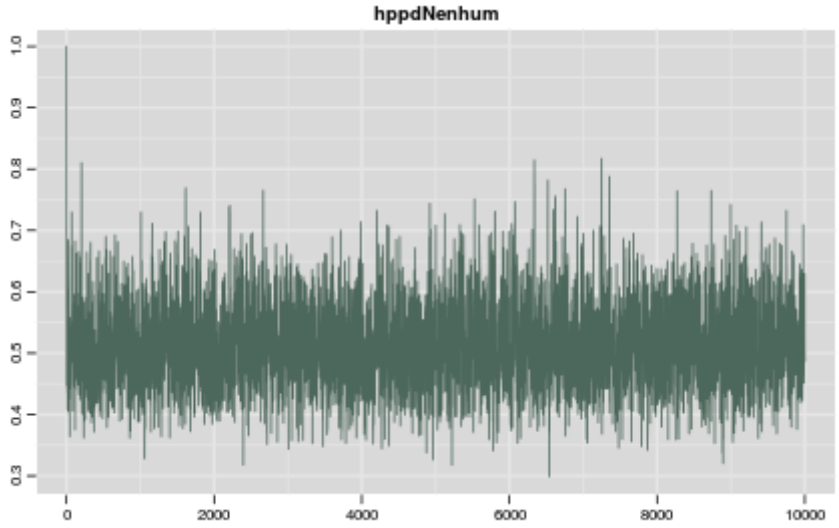
5.5 Interpretação dos resultados

Os gráficos de *trace* dos resultados mostram o comportamento das simulações geradas para cada parâmetro do modelo. As cadeias foram criadas com o tamanho definido pelo *longchain* (10.000), sendo cada gráfico representativo ou do parâmetro do intercepto (título "(*Intercept*)") ou do parâmetro de um dos preditores do modelo criado.

No exemplo utilizado, nota-se que as cadeias apresentam, de forma visual, um comportamento estável ao longo das iterações. No caso do coeficiente do preditor "hppdNenhum" (Figura 6) é notável uma diferença no início da cadeia, mas o *burn-in* definido acaba corrigindo o problema de forma conservadora. A partir do ponto 1.000, a

simulação parece estar explorando a distribuição de forma equilibrada, indicando um bom gráfico de *trace*. As outras variáveis apresentam as mesmas propriedades, mesmo sem a remoção dos valores iniciais (*burn-in*).

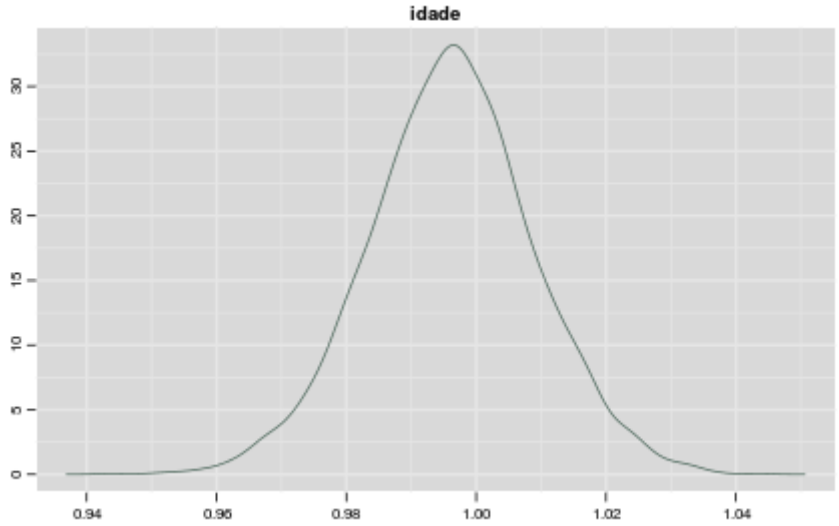
Figura 6. Gráfico de *trace* para o preditor “hppdNenhum”



Fonte: Elaborada pelo autor.

Nos gráficos de densidade dos resultados, podemos observar a distribuição a posteriori, estimada não-parametricamente, para cada um dos coeficientes do modelo. Distribuições com formatos não muito suaves podem indicar problemas nas definições do MCMC. A interpretação é como a de qualquer gráfico de densidade, onde os valores com maior ocorrência apresentam um volume maior abaixo da curva. No caso do preditor “idade” (Figura 7), observamos que os valores entre 0,99 e 1 são os de maior probabilidade a posteriori.

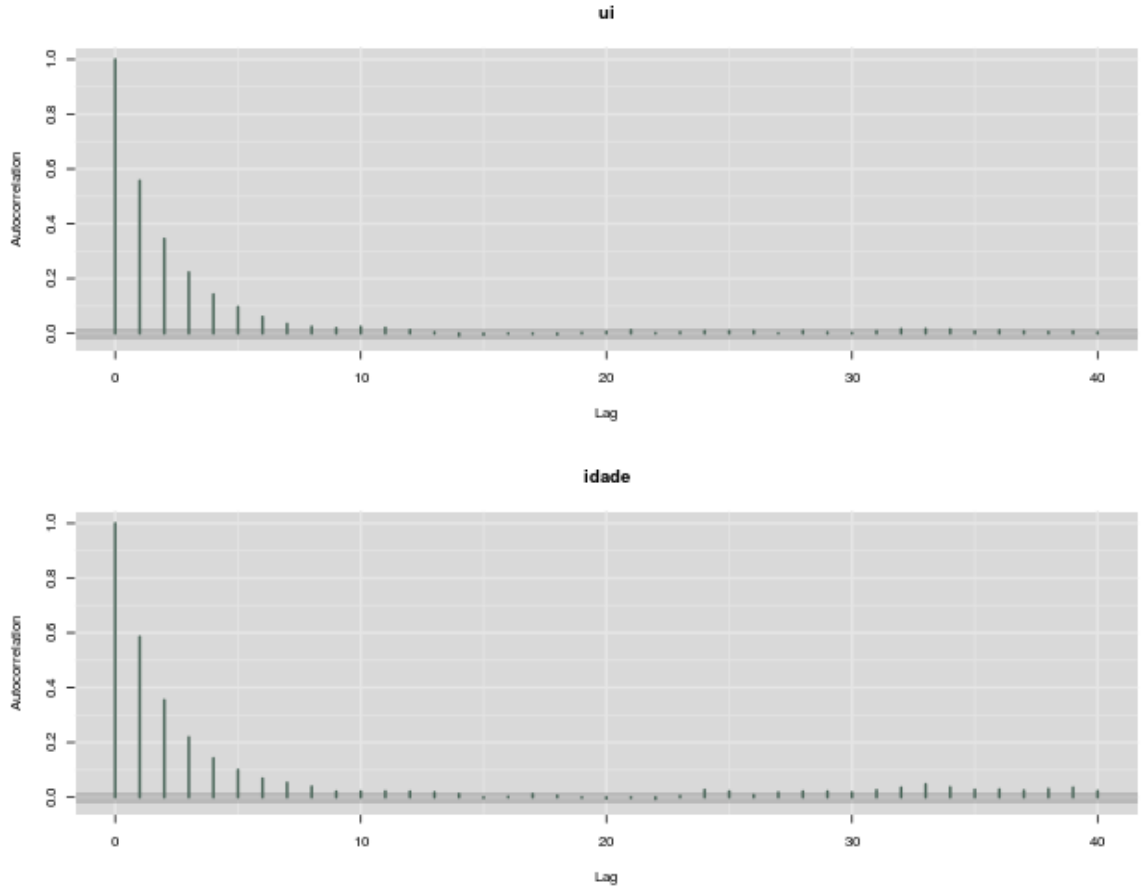
Figura 7. Gráfico de densidade para o preditor “idade”



Fonte: Elaborada pelo autor.

Os gráficos de autocorrelação ajudam a avaliar a independência entre os valores gerados, informação importante para o diagnóstico do MCMC, já que a autocorrelação deve ser inexistente. Estes gráficos indicam o tamanho do *thin* a ser utilizado. Para o nosso exemplo podemos, observar que o valor escolhido para o *thin*, igual a 10, será o suficiente para garantir essa propriedade, pois as autocorrelações estão próximas a zero a partir desse ponto. Esta constatação foi observada para todos os preditores escolhidos. A Figura 8 exibe os gráficos de autocorrelação para os preditores “ui” e “idade”, como ilustração.

Figura 8. Gráficos de autocorrelação para os preditores “ui” e “idade”



Fonte: Elaborada pelo autor.

As estimativas geradas dos parâmetros do modelo gerado estão localizadas no bloco “*Coefficients*” (Figura 9). Inicialmente, em *“Effective Size”*, está o tamanho efetivo da cadeia, que é o cálculo do número de valores simulados, para cada preditor, que não apresentam autocorrelação. Idealmente os valores devem estar próximos do número indicado em “*Sample size per chain*”, que aparece no “*Summary*”. As estimativas pontuais e intervalares são listadas no “*Summary*”. Nesse bloco, já foi removido o valor do *burn-in* e aplicado o *thin*, como indicado no texto inicial do “*Summary*”. As estimativas pontuais da média (coluna *Mean*) e desvio-padrão (coluna *SD*) encontram-se na primeira parte do “*Summary*”, enquanto a mediana (coluna 50%) pode ser encontrada na segunda parte, juntamente com o intervalo de 95% de credibilidade, obtido pelos valores nas colunas 2.5% e 97.5%. Após a aplicação do *thin*, se os valores do *Naive SE* e *Time-series SE* forem próximos, há indícios de que a autocorrelação remanescente foi removida satisfatoriamente. Chu e Cole (2010) recomendam que a mediana seja utilizada como estimativa pontual dos parâmetros do modelo log-binomial.

Figura 9. Resultados dos coeficientes do modelo

Coefficients

Effective Size

(Intercept)	fumoSim	uiSim	NCTG3CATNenhuma	NCTG3CATUma
781.5934	900.0000	900.0000	610.8600	900.0000
idade	pmum	hppdNenhum	htSim	
634.3963	900.0000	1008.1886	795.7988	

Summary

```
Iterations = 1001:9991
Thinning interval = 10
Number of chains = 1
Sample size per chain = 900

1. Empirical mean and standard deviation for each variable,
   plus standard error of the mean:

                    Mean       SD  Naive SE Time-series SE
(Intercept)     -0.343563 0.527485 1.758e-02      1.887e-02
fumoSim          0.204892 0.136336 4.545e-03      4.545e-03
uiSim            0.284015 0.131685 4.390e-03      4.390e-03
NCTG3CATNenhuma  0.206186 0.190398 6.347e-03      7.704e-03
NCTG3CATUma     -0.099161 0.207453 6.915e-03      6.915e-03
idade           -0.003312 0.013272 4.424e-04      5.269e-04
pmum            -0.004389 0.002159 7.195e-05      7.195e-05
hppdNenhum      -0.677076 0.126420 4.214e-03      3.981e-03
htSim            0.531441 0.185575 6.186e-03      6.578e-03

2. Quantiles for each variable:

                    2.5%       25%       50%       75%      97.5%
(Intercept)     -1.356729 -0.685912 -0.328899  0.005044  0.6957661
fumoSim         -0.062134  0.108329  0.201947  0.300626  0.4678310
uiSim            0.020988  0.197807  0.285017  0.367813  0.5424138
NCTG3CATNenhuma -0.164573  0.081069  0.202501  0.335861  0.5776978
NCTG3CATUma     -0.505097 -0.237759 -0.107830  0.044527  0.2963726
idade           -0.032472 -0.011332 -0.003187  0.005224  0.0223719
pmum            -0.008917 -0.005913 -0.004266 -0.002859 -0.0003095
hppdNenhum      -0.930596 -0.757909 -0.674842 -0.594961 -0.4337977
htSim            0.156222  0.411674  0.533240  0.657580  0.8738195
```

Fonte: Elaborada pelo autor.

As informações do bloco "*Relative Risk*" (Figura 10) seguem o mesmo padrão do bloco "*Coefficients*", mas nesse caso estão apresentadas as exponenciais dos coeficientes, ou seja, os valores de RR para cada um dos preditores. Para o coeficiente "uiSim", o intervalo de credibilidade não contém o valor 1, assim fornecendo evidências da associação do preditor com o desfecho neste modelo. Nesse caso, utilizando a mediana como estimador do RR, tem-se que o risco do bebê nascer com baixo peso para mulheres com irritabilidade uterina é 32,98% maior do que o risco para mulheres que não apresentam essa característica, quando os valores dos outros sete preditores forem os mesmos (ou controlando pelos demais sete preditores).

Figura 10. Resultados dos riscos relativos do modelo

Relative Risk

Effective Size

```
     (Intercept)         fumoSim           uiSim NCTG3CATNenhuma     NCTG3CATUma
        563.8420        900.0000        900.0000        622.9727        900.0000
           idade            pmum      hppdNenhum           htSim
        634.7565        900.0000       1001.7819        802.1494
```

Summary

```
Iterations = 1001:9991
Thinning interval = 10
Number of chains = 1
Sample size per chain = 900

1. Empirical mean and standard deviation for each variable,
   plus standard error of the mean:

                  Mean       SD  Naive SE Time-series SE
(Intercept)     0.8149 0.464462 1.548e-02      1.956e-02
fumoSim         1.2388 0.169396 5.647e-03      5.647e-03
uiSim           1.3400 0.177595 5.920e-03      5.920e-03
NCTG3CATNenhuma 1.2514 0.239866 7.996e-03      9.610e-03
NCTG3CATUma     0.9252 0.193225 6.441e-03      6.441e-03
idade           0.9968 0.013207 4.402e-04      5.242e-04
pmum            0.9956 0.002149 7.162e-05      7.162e-05
hppdNenhum      0.5122 0.064695 2.156e-03      2.044e-03
htSim           1.7305 0.317095 1.057e-02      1.120e-02

2. Quantiles for each variable:

                  2.5%    25%    50%    75%  97.5%
(Intercept)     0.2575 0.5036 0.7197 1.0051 2.0052
fumoSim         0.9398 1.1144 1.2238 1.3507 1.5965
uiSim           1.0212 1.2187 1.3298 1.4446 1.7202
NCTG3CATNenhuma 0.8483 1.0844 1.2245 1.3991 1.7819
NCTG3CATUma     0.6034 0.7884 0.8978 1.0455 1.3450
idade           0.9680 0.9887 0.9968 1.0052 1.0226
pmum            0.9911 0.9941 0.9957 0.9971 0.9997
hppdNenhum      0.3943 0.4686 0.5092 0.5516 0.6480
htSim           1.1691 1.5093 1.7044 1.9301 2.3960
```

Fonte: Elaborada pelo autor.

O *output* dos resultados mostram ainda com dois testes de diagnóstico de convergência do método MCMC que fazem parte do pacote `coda` (PLUMMER et al., 2015). Inicialmente os blocos com os testes de convergência estão minimizados; para expandir, basta clicar no símbolo “+” no canto superior direito de cada bloco.

O primeiro teste é o Geweke (Figura 11). O teste consiste em selecionar duas frações não sobrepostas da cadeia gerada para cada parâmetro, os primeiros 10% e os últimos 50%, e realizar uma comparação das médias dessas duas frações através de um teste de hipóteses. A estatística de teste apresentada é uma estatística Z que tem distribuição normal padronizada. No exemplo, para nenhum parâmetro há diferença significativa a 5% (valor crítico 1,96) entre resultados das duas frações (todos valores $Z < 1{,}96$), ou seja, por este teste a convergência parece estar adequada.

Figura 11. Teste de diagnóstico de convergência Geweke

Fonte: Elaborada pelo autor.

O segundo teste de diagnóstico de convergência disponível nos resultados é o de Heidelberger e Welch (Figura 12), que por definição da função do pacote `coda` utiliza parâmetros *eps* = 0.1 e significância 5%. Esse teste é dividido em duas partes. Na primeira calcula-se uma estatística que testa a hipótese nula de que a cadeia de Markov é estacionária. Para isto são feitos descartes percentuais de valores iniciais da cadeia e procede-se o teste até a hipótese não ser rejeitada ou alcançar 50% de descarte. Se for necessário chegar aos 50% de descarte, existe indício de que uma cadeia MCMC maior deve ser gerada. Todos os preditores do exemplo passaram na primeira parte do teste (*"passed"*, coluna "*Stationarity test*"). Também na coluna ao lado está o valor-p para cada um dos preditores (coluna *p-value*).

Na segunda parte do teste, é calculado o intervalo de confiança de 95% para a média com a porção da cadeia que foi considerada estacionária e metade da amplitude desse intervalo é comparado com a estimativa pontual da média. Caso a razão entre esses valores seja menor que *eps*, se considera que há convergência (*"passed"*). Se o valor for maior, recomenda-se aumentar o tamanho da cadeia. Os valores utilizados para a razão *eps* podem ser encontrados nas colunas *"Mean"* e *"Halfwidth"*, para o exemplo todos os preditores passaram no teste de Heidelberger e Welch.

Figura 12. Teste de diagnóstico de convergência Heidelberger e Welch

Heidelberger and Welch

Convergence diagnostic

	Stationarity test	start iteration	p-value
(Intercept)	passed	1	0.2826
fumoSim	passed	1	0.7098
uiSim	passed	1	0.4859
NCTG3CATNenhuma	passed	1	0.8770
NCTG3CATUma	passed	1	0.5048
idade	passed	1	0.0817
pmum	passed	1	0.1905
hppdNenhum	passed	1	0.7358
htSim	passed	1	0.6101

	Halfwidth test	Mean	Halfwidth
(Intercept)	passed	0.833	1.59e-02
fumoSim	passed	1.242	6.73e-03
uiSim	passed	1.344	7.05e-03
NCTG3CATNenhuma	passed	1.232	8.72e-03
NCTG3CATUma	passed	0.915	7.65e-03
idade	passed	0.996	5.12e-04
pmum	passed	0.996	9.03e-05
hppdNenhum	passed	0.512	2.42e-03
htSim	passed	1.748	1.35e-02

Fonte: Elaborada pelo autor.

5.6 Opções *post hoc*

Após finalizar a análise, existem opções adicionais para o usuário. Como salvar o relatório exposto na página, selecionando entre as opções de arquivo disponíveis e clicando no botão "*Download"*. Além disso, se os resultados não forem satisfatórios, a análise pode ser reconfigurada fazendo alterações na aba "*Model Selection*" e também na aba "*MCMC*". Uma vez redefinida alguma propriedade, é necessário voltar à aba "*Analysis*" e executar a análise novamente. Para encerrar o Bayesian RR, basta fechar o navegador. Após a finalização do programa, todos os dados temporários como os resultados e banco de dados são automaticamente excluídos do servidor.

6 CONCLUSÕES

Neste trabalho foi proposta e desenvolvida uma solução visual, interativa e amigável para a aplicação da abordagem bayesiana na estimação do risco relativo, para dados independentes e desfecho binário com a intenção de facilitar e difundir esse método estatístico. A construção desta interface foi possível utilizando a linguagem de programação R e pacotes de desenvolvimento e análise, baseado em inovações propostas recentes (TORMAN; CAMEY, 2015 e SALMERÓN; CANO; CHIRLAQUE, 2015).

Entre as vantagens dessa proposta está a facilidade de uso da interface construída, já que o usuário não necessita ter qualquer noção de programação para executar uma análise feita em R, onde tradicionalmente qualquer ação é feita através de linhas de comando. Além disso, essa linguagem de programação é aberta, o que facilita na sua disseminação e permite que outras pessoas contribuam para o melhoramento do console, pacotes e programas construídos a partir dela, sem o custo geralmente alto embutido na aquisição de softwares estatísticos.

Outra vantagem desta solução é o fato de que, para seu uso, é necessário somente acesso à internet e um dispositivo eletrônico que permita o envio do arquivo que contém o banco de dados. Devido à natureza do programa base utilizado, a execução do aplicativo se demonstrou eficaz mesmo para cadeias com número elevado de iterações, que exigem um maior tempo de processamento. Não foi de interesse comparar a diferença do tempo de execução do programa localmente ou remotamente, pois um fator de influência no tempo de execução é a velocidade de conexão do usuário.

O maior, e talvez único, problema que pudemos observar nesta solução, foi a possibilidade de desconexão do servidor durante o uso da ferramenta. Esse fato pode decorrer tanto de problemas da conexão do usuário, como do servidor remoto gratuito utilizado para hospedar o aplicativo. Caso a desconexão do servidor seja frequente, é recomendado que o usuário realize o download dos arquivos[4] que compõem a interface e a execute no console R de seu computador local, as instruções de como proceder encontram-se junto aos arquivos. O Bayesian RR pode ser expandido para incorporar a análise de dados que tenham dependência e/ou desfecho politômico, bem como agregar outros aspectos da análise, por exemplo, diagnósticos de relação linear (nos parâmetros β) entre preditores quantitativos e o desfecho.

[4] Disponível em: <https://goo.gl/PXuR8b>.

REFERÊNCIAS

ALLAIRE, JJ et al. **Rmarkdown**: Dynamic Documents for R. Version: 1.1. Boston, 2016. Disponível em: <https://CRAN.R-project.org/package=rmarkdown>. Acesso em: 30 nov. 2016.

BARROS, Aluísio J.D.; HIRAKATA, Vânia N. Alternatives for logistic regression in cross-sectional studies: an empirical comparison of models that directly estimate the prevalence ratio. **BMC Medical Research Methodology**. Londres, v. 03, n. 01, p. 01, 2003.

CHANG, Winston. **Shiny**: Web Application Framework for R. Version: 0.14.2. Boston, 2016a. Disponível em: <https://CRAN.R-project.org/package=shiny>. Acesso em: 30 nov. 2016.

_____. **Shinydashboard**: Create Dashboards with 'Shiny'. Version: 0.5.3. Boston, 2016b. Disponível em: <https://CRAN.R-project.org/package=shinydashboard>. Acesso em: 30 nov. 2016.

CHU, Haitao; COLE, Stephen R. Estimation of risk ratios in cohort studies with common outcomes: a Bayesian approach. **Epidemiology**. Atlanta, v. 21, n. 06, p. 855-862, 2010.

CURTIS, S. M et al. **Mcmcplots**: Create Plots from MCMC Output. Version: 0.4.2. 2015. [S.l.], 2015. Disponível em: <https://CRAN.R-project.org/package=mcmcplots>. Acesso em: 30 nov. 2016.

DAVIES, Huw T. O.; CROMBIE, Iain K.; TAVAKOLI, Manouche. When can odds ratios mislead? **British Medical Journal**. Londres, v. 317, n. 7166, p. 1155-1157, 1998.

ELLIS, David A.; MERDIAN, Hannah L. Thinking outside the box: developing dynamic data visualizations for psychology with Shiny. **Frontiers in Psychology.** Pully, v. 06, 2015.

GILKS, WR; RICHARDSON, S, SPIEGELHALTER, DJ. (Ed.). **Markov chain Monte Carlo in practice**. Londres: Chapman & Hall, 1996.

HOSMER, David W.; LEMESHOW, Stanley. **Applied logistic regression**. New York: John Wiley & Sons, 1989.

JAHANSHIRI, Ebrahim; SHARIFF, Abdul R. M. Developing web-based data analysis tools for precision farming using R and Shiny. **IOP Conference Series: Earth and Environmental Science**. Bristol, v. 20, n. 01, 2014.

LEE, James. Odds ratio or relative risk for cross-sectional data? **International Journal of Epidemiology**. Londres, v. 23, n. 01, p. 201-203, 1994.

LUMLEY, Thomas; ZEILEIS, Achim. **Sandwich**: Robust Covariance Matrix Estimators. Version: 2.3-4. Boston, 2015. Disponível em: <https://CRAN.R-project.org/package=sandwich>. Acesso em: 30 nov. 2016.

PERONDI, Daniel et al. Uma solução computacional de aquisição, tratamento, armazenamento, disponibilização e apresentação de dados meteorológicos. In: CONGRESSO

BRASILEIRO DE AGROINFORMÁTICA, 10., 2015, Ponta Grossa. **Anais eletrônicos...** Ponta Grossa:
Universidade Estadual de Ponta Grossa (UEPG), 2015. Disponível em: <http://eventos.uepg.br/sbiagro/2015/anais/SBIAgro2015/analcompleto.htm>. Acesso em: 30 nov. 2016.

PLUMMER, Martyn et al. **Coda**: Output Analysis and Diagnostics for MCMC. Version: 0.18-1. [S.l.], 2015. Disponível em: <https://CRAN.R-project.org/package=coda>. Acesso em: 30 nov. 2016.

R PROJECT. **The R Project for Statistical Computing.** [S.l.]. Disponível em: <https://www.r-project.org/>. Acesso em: 30 nov. 2016.

SALMERÓN, Diego; CANO, Juan A.; CHIRLAQUE, María D. Reducing Monte Carlo error in the Bayesian estimation of risk ratios using log-binomial regression models. **Statistics in Medicine.** Chichester, v. 34, n. 19, p. 2755-2767, 2015.

SPIEGELMAN, Donna; HERTZMARK, Ellen. Easy SAS calculations for risk or prevalence ratios and differences. **American Journal of Epidemiology**. Baltimore, v. 162, n. 03, p. 199-200, 2005.

TORMAN, Vanessa B. L.; CAMEY, Suzi A. Bayesian models as a unified approach to estimate relative risk (or prevalence ratio) in binary and polytomous outcomes. **Emerging Themes in Epidemiology.** Londres, v. 12, n. 01, p. 01, 2015.

WACHOLDER, Sholom. Binomial regression in GLIM: estimating risk ratios and risk differences. **American Journal of Epidemiology.** Baltimore, v. 123, n. 01, p. 174-184, 1986.

WOJCIECHOWSKI, J.; HOPKINS, A. M.; UPTON, R. N. Interactive pharmacometric applications using R and the shiny package. **CPT: Pharmacometrics & Systems Pharmacology**. New York, v. 04, n. 03, p. 146-159, 2015.

ANEXOS A – DICIONÁRIO DO BANCO BAIXO PESO

Nº	VARIÁVEL	NOME DA VARIÁVEL	RÓTULO DOS VALORES
1.	Identificação do recém-nascido	ID	-
2.	Peso do nascido (em gramas)	peso_bebe	Quantitativa
3.	Baixo peso ao nascer	bpn	0 = Não 1 = Sim
4.	Número de consultas da mãe no 1º trimestre da gestação	nctg	Quantitativa
5.	Idade da mãe (em anos)	idade	Quantitativa
6.	Peso da mãe no último período menstrual (em libras)	pnum	Quantitativa
7.	História/número de partos prematuros	hpp	Quantitativa
8.	Hábito de fumar na gestação	fumo	0 = Não 1 = Sim
9.	História de hipertensão	ht	0 = Não 1 = Sim
10.	Presença de irritabilidade uterina	ui	0 = Não 1 = Sim
11.	Raça/cor da mãe	raca	1 = Branca 2 = Negra 3 = Outra
12.	Peso da mãe no último período menstrual (em quilos)	pnumkg	Quantitativa
13.	Variável dicotômica indicadora da experiência (ou não) de parto prematuro prévio	hppd	Se hpp = 0 → hppd = 0 (Nenhum) Se hpp > 0 → hppd = 1 (1 ou mais)
14.	Variável categórica representando a frequência de visitas ao médico no primeiro trimestre da gestação, 3 categorias, construída a partir de NCTG	NCTG3CAT	Se nctg = 0 → NCTG3CAT = 1 (Nenhum) Se nctg = 1 → NCTG3CAT = 2 (Um) Se nctg ≥2 → NCTG3CAT = 3 (Duas ou mais)