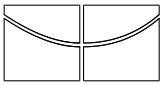

Universidade de Brasília

Instituto de Ciências Exatas

Departamento de Estatística

Estágio supervisionado 2

Inferência Bayesiana na análise de dados de experimentos planejados

por

Rafael Moraes Gazzinelli

Orientador: Prof. Dr. Afrânio Márcio Corrêa Vieira

Dezembro de 2013

Rafael Moraes Gazzinelli

Inferência Bayesiana na análise de dados de experimentos planejados

Relatório apresentado à disciplina Estágio Supervisionado 2, do curso de graduação em Estatística, Instituto de Ciências Exatas, Universidade de Brasília, como parte dos requisitos necessários para o grau de Bacharel em Estatística.

Orientador: Prof. Dr. Afrânio Márcio Corrêa Vieira

Universidade de Brasília

Brasília, Dezembro de 2013

Aos meus pais, Sônia e Euclides (in memorian).

Agradecimentos

Agradeço primeiramente a Deus pela saúde, força e amparo diário.

Ao professor Afrânio Márcio Corrêa Vieira, pela sua orientação, dedicação e inestimável auxílio durante todo o trabalho.

Aos professores, funcionários e amigos do Departamento de Estatística da Universidade de Brasília, pelos ensinamentos, apoio e incentivo no decurso da graduação.

Ao Dr. Marcos A. Gimenes e a Dr. Paula André S. de Vasconcelos Carvalho da Embrapa Recursos Genéticos e Biotecnologia, por ter cedido o conjunto de dados para realização das análises.

À todas as pessoas que me ajudaram e torcem pelo meu progresso.

Ao SAS Institute Brasil por possibilitar a utilização desse software por meio de parceria acadêmica com o Departamento de Estatística da UnB.

Resumo

Inferência Bayesiana na análise de dados de experimentos planejados

Em estudos de experimentos planejados a Estatística Clássica é a base para análise dos experimentos, porém, existem trabalhos publicados no Brasil que já utilizaram métodos Bayesianos. Neste trabalho foi utilizado as duas abordagem para análise de um experimento real fornecido pela Embrapa Recursos Genéticos e Biotecnologia. A vantagem da inferência Bayesiana consiste em obter uma função de densidade contendo toda a informação probabilística a respeito dos parâmetros de interesse. O trabalho, na primeira parte, revisa a literatura Bayesiana, simulação estocástica, método de Monte Carlo via Cadeia de Markov, algoritmos de Metropolis-Hastings, amostrador de Gibbs, introdução aos modelos lineares generalizados e diagnósticos de convergência da cadeias, como Geweke, Raftery-Lewis e Heidelberger-Welch. Em segunda etapa realizou-se um estudo de caso tratado pelas abordagens Clássica e Bayesiana. Os resultados obtidos em ambos os métodos foram comparados e obtiveram os valores estimados dos parâmetros e os intervalos de confiança e credibilidade aproximados. A abordagem Bayesiana mostrou-se eficiente e satisfatória mesmo utilizando distribuições *a priori* vagas. Os dados foram tratados utilizando o software SAS, que se mostrou eficaz e robusto.

Palavras-chaves: Inferência Bayesiana; MCMC; simulação estocástica; algoritmos de Metropolis-Hastings e amostrador de Gibbs; modelos lineares generalizados; diagnósticos de convergência; e *Software SAS*

Lista de Figuras

Lista de Tabelas

Sumário

1 Introdução

Na estatística inferencial um dos interesses é estimar parâmetros, que são valores populacionais desconhecidos. A maneira que a incerteza sobre os parâmetros é tratada é uma das grandes diferenças entre a inferência Clássica e a Bayesiana. A primeira trata os parâmetros como valores fixos e desconhecidos e o estima de maneira analítica; já a segunda, atribui distribuições probabilísticas *a priori* aos parâmetros, para a estimação. Essa é uma das vantagens da abordagem Bayesiana, pois o parâmetro é tratado como uma variável aleatória e, consequentemente, é possível calcular intervalos de credibilidade e estimativas mais precisas com base na distribuição *a posteriori*. A distribuição *a priori* é utilizada para incorporar informações prévias a respeito do parâmetro. A origem dessas informações podem estar em dados de testes já realizados, dados experimentais em diferentes ambientes, experiências pessoais e resultados em itens similares.

Na análise de experimentos planejados é comum ter informações *a priori*, seja por variedades similares, seja pelo conhecimento especializado do pesquisador. Existem alguns trabalhos publicados no Brasil, em periódicos das ciências agrárias, que já utilizaram métodos Bayesianos, como [Barbosa(2005)], [Boligon and Alburquerque(2010)] e [de Aquino(2008)], que utilizaram distribuições *a priori* vagas ou não informativas. Neste trabalho, o objetivo é realizar um estudo de caso e comparar a abordagem Clássica e Bayesiana.

Para isso serão revisados os seguintes assuntos na literatura: inferência Bayesiana, simulação estocástica, método de Monte Carlo via Cadeia de Markov, algoritmos de Metropolis-Hastings, amostrador de Gibbs, diagnósticos de convergência e introdução aos modelos lineares generalizados. Como também um estudo de caso para comparar a abordagem Clássica e Bayesiana, com base em dados que foram selecionados de um experimentos real, a respeito do resveratrol, desenvolvidos e fornecidos pela EMBRAPA Recursos Genéticos e Biotecnologia.

2 Metodologia

2.1 Introdução a Inferência Bayesiana

A seguir será introduzido o Formula de Bayes, a Inferência Bayesiana e o conceito de distribuições *a priori* conjugadas.

2.1.1 Formula de Bayes

Seja $A_1, A_2, ..., A_i$ variáveis aleatórias de uma partição do espaço amostral Ω. Seja B um conjunto qualquer. Então para $i = 1, 2, ..., n$, tem-se

$$P(A_i|B) = \frac{P(B|A_i)P(A_i)}{P(B)} = \frac{P(B|A_i)P(A_i)}{\sum_{i=1}^{n} P(B|A_i)P(A_i)} \tag{2.1}$$

Dado "$\propto$"como indicação de proporcionalidade, portanto

$$P(A_i|B) = \frac{P(B|A_i)P(A_i)}{P(B)} \propto P(B|A_i)P(A_i) \tag{2.2}$$

Essa equação acima é conhecida como regra de Bayes, introduzida por Pierre-Simon de Laplace, para maiores detalhes ver [Ntzoufras(2009)].

2.1.2 Inferência Bayesiana

A abordagem Bayesiana leva em consideração o fato de que o investigador tem algum conhecimento sobre o parâmetro θ e que esse pode ser incorporado à análise. Já um estatístico clássico não admite essa informação, porque ela não foi observada no estudo, portanto, não está sujeita a verificação empírica. A abordagem Bayesiana incorpora esta informação na análise por meio de uma função densidade $f(\theta)$. Ver [Gamerman and Lopes(2006), p. 42].

Para análise Bayesiana é necessário conhecer a função de verossimilhança $f(y|\theta)$ e a distribuição *a priori* do parâmetro, $f(\theta)$. A função de verossimilhança pode ser

obtida de uma função de θ, e assim tem-se,

$$L(\theta|\mathbf{y}) = f(\mathbf{y}|\theta) = f(y_1;\theta) \times ... \times f(y_n;\theta) = \prod_{i=1}^{n} f(y_i|\theta) \tag{2.3}$$

sendo que nesta verossimilhança assume-se independência entre as observações. $f(\theta)$ é também chamada de densidade *a priori* por conter a distribuição de probabilidade de θ antes de observar os valores da amostra.

distribuição *a posteriori* de θ é obtida pela formula de Bayes (2.2):

$$f(\theta|y) = \frac{f(y|\theta)f(\theta)}{f(y)} \propto f(y|\theta)f(\theta) \tag{2.4}$$

onde, $f(y) = \int f(y|\theta)f(\theta)d\theta$. Como $f(\theta)$ é uma densidade de θ, os valores observados de y são simplesmente uma constante, assim como $f(y)$. A função $f(\theta|y)$ é denominada função densidade *a posteriori* de θ, após o conhecimento da amostra.

2.1.3 Prioris Conjugadas

A informação *a priori* é o conhecimento que se tem sobre θ antes da realização do experimento. Na ausência desse conhecimento adota-se uma priori não informativa ou se faz uma eliciação da priori. Para tal é necessário que essa tenha uma distribuição sobre o parâmetro θ.

Seja uma distribuição *a priori* $D(\alpha)$ pertence à família de distribuição D com parâmetro α e distribuição conjugada $f(y|\theta)$. Sabendo que a distribuição *a posteriori* $f(\theta|y)$ é da mesma família de distribuição D, portanto

$$\text{Se } \theta \sim D(\alpha) \text{ então } \theta|y \sim D(\tilde{\alpha}) \tag{2.5}$$

Em que α e $\tilde{\alpha}$ são os parâmetros da distribuição *a priori* e da *posteriori* de D, respectivamente. A abordagem da distribuição *a priori* conjugada facilita a análise do ponto de vista analítico por ter a propriedade de resultar *a posteriori* na mesma família de distribuição. Com essa propriedade, a atualização do conhecimento que se tem de θ envolve apenas a mudança dos novos parâmetros, também chamados de *hiperparâmetros*. A tabela das prioris conjugadas encontra-se no anexo 7.5.

Exemplo:[Ntzoufras(2009)] Gols marcados pela seleção de futebol da Grécia na Euro 2004 (dados Poisson).

A distribuição de Poisson é amplamente utilizada para modelar esses dados, apesar de um ligeira sobredispersão ser observada em [Karlis and Ntzoufras(2000)]. Neste exemplo, considere a pontuação final da seleção da Grécia na Euro 2004, onde a Grécia surpreendentemente venceu a competição, ver tabela 7.5. O objetivo é estimar a distribuição *a posteriori* do número esperado de gols marcados e sofridos pela Grécia, bem como o número total de gols marcados pelas equipes oponentes e da Grécia (este último é frequentemente utilizado para fins de apostas).

Tabela 2.1: Jogos da Grécia na competição Euro 2004

	Oponentes	Gols marcados		Total
		A favor	Contra	
1	Portugal	2	1	3
2	Espanha	1	1	2
3	Rússia	1	2	3
4	Franca	1	0	1
5	República Checa	0 (1)	0 (0)	0 (1)*
6	Portugal	1	0	1
	Soma	6	4	10
	Media amostral $\bar{x}$	1	0,67	1,67

*tempo normal 0x0, após tempo extra de 15m 1x0

Para modelar os gols marcados e sofridos pela Grécia (denotado por y_i^m e y_i^s, respectivamente) e o número total de gols ($y_i^t = y_i^m + y_i^s$), será usado a distribuição de Poisson. Assim, podemos escrever

$$y_i^k \sim Poisson(\theta^k), \text{ para } k \in \{m, s, t\} \text{ e } i = 1, 2, ..., 6$$

Onde θ^m e θ^s são os números esperados de gols marcados e sofridos pela Grécia respectivamente, e θ^t é o número total esperado de gols marcados em cada jogo.

Considerando uma distribuição *a priori* não informativa, então é usado uma $gama(0,001; 0,001)$ *a priori*, com média igual a um e variância igual a mil. Sabendo que a distribuição *a posteriori* conjugada é dada por $\theta|y \sim gama(n\bar{y} + a, n + b)$

(Tabela 7.5). Então as distribuições *a posteriori* para gols marcados, contra e total são expressos a seguir, com a média e o erro padrão respectivos:

$$\theta^m|y \sim gama(6,001; 6,001)$$
$$\theta^s|y \sim gama(4,001; 6,001)$$
$$\theta^t|y \sim gama(10,001; 6,001)$$

$$E(\theta^m|y) = \frac{6,001}{6,001} = 1,000,\ SD(\theta^m|y) = \frac{6,001}{6,001^2} = 0,166$$
$$E(\theta^s|y) = \frac{4,001}{6,001} = 0,667,\ SD(\theta^s|y) = \frac{}{4,001^2} = 0,111$$
$$E(\theta^t|y) = \frac{10,001}{6,001} = 1,667,\ SD(\theta^t|y) = \frac{10,001}{6,001^2} = 0,278$$

A representação gráfica das distribuições *a posteriori* pode ser vista na figura a seguir

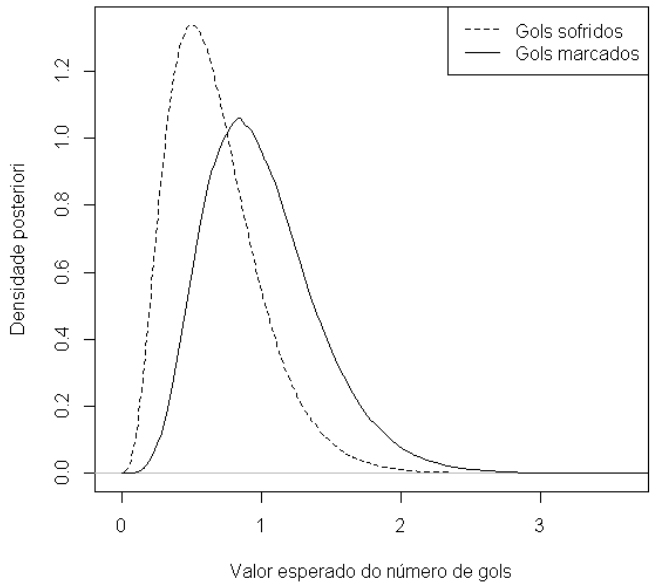

Figura 2.1: Distribuições posteriores do número médio de gols marcados e sofridos pela Grécia no Euro 2004; Priori com baixa informação.

Agora, usando distribuição com *a priori* informativa. Antes da competição Euro 2004, informações a priori estavam disponível a partir da fase de qualificação dos

grupo da competição. Embora a qualidade dos adversários não fosse tão alta quanto na Euro 2004, pode-se extrair informações a partir desses jogos e especificar uma distribuição *a priori* mais informativa do que o utilizada anteriormente.

Com bases nessas informações, a seleção grega foi a primeira em seu grupo que marcou oito gols e levou quatro em apenas oito jogos. A partir desta informação, pode-se construir a distribuição *a posteriori*, após a conclusão da fase de qualificação. Assim, de acordo com *a posteriori* conjugada dada por $\theta|y \sim gama(n\bar{y} + a, n + b)$ (Tabela 7.5), a nossa distribuição *a posteriori*, após considerar estes jogos, é uma $gama(8, 8)$ para gols a favor e uma $gama(4, 8)$ para gols contra a Grécia. Embora essas distribuições *a posteriori* possam ser usadas diretamente como distribuições *a priori* para a análise dos dados no Euro 2004, propõem-se uma pequena modificação da distribuição *a priori*, aumentando a variância, a fim de refletir a incerteza adicional, pois esta informação resultou em diferentes condições. As distribuições *a priori* assumem as mesmas médias (1 e 0,5), mas a variância é multiplicada pelo tamanho da amostra dos dados anteriores. Portanto, as distribuições *a priori* são, respectivamente, $gama(1, 1)$ e $gama(1, 2)$ para gols marcados a favor e contra a Grécia. De acordo com *a posteriori* conjugada, agora a distribuição posterior para os gols marcados pela Grécia é $\theta^m|y \sim gama(7, 7)$, com média 1 e desvio padrão $0,38$ gols por jogo. Enquanto o posterior para os gols sofridos pela Grécia é $\theta^s|y \sim gama(5, 8)$ com média $0,625$ e desvio padrão $0,078$ por jogo. Estas distribuições posteriores são dadas a seguir pela figura 2.2.

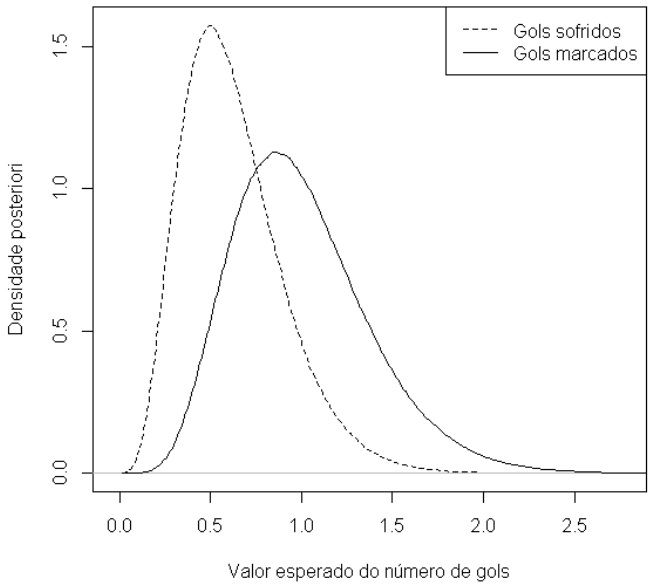

Figura 2.2: Distribuições posteriores do número médio de gols marcados e sofridos pela Grécia no Euro 2004, incluindo prioris informativas da qualificação dos jogos de grupo.

Na análise acima, foi excluído o gol no tempo extra jogado no jogo contra a República Checa. Isto pode ser facilmente incorporado, considerando o fator de tempo. Neste caso, podemos supor que $y_i \sim Poisson(t_i\theta)$, onde t_i é igual a um jogo normal de 90 minutos, $t_i = 1 + \frac{15}{90} = 1,167$ para jogos com 15 minutos de tempo extra e $t_i = 1 + \frac{30}{90} = 1,33$ para jogos com 30 minutos de tempo extra. Após a inclusão do tempo extra adicional, a distribuição *a posteriori* é

$$\theta|y \sim gama\left(\sum_{i=1}^{n} t_i y_i + a, \sum_{i=1}^{n} t_i + b\right)$$

Utilizando os dados do nosso exemplo, as distribuições posteriores são ligeiramente alteradas para

$$\theta^m|y \sim gama(7,168; 6,618) \text{ e } \theta^s|y \sim gama(4,001; 6,618)$$

no caso não informativo. No caso com distribuições *a priori* informativas, tem-se

$$\theta^m|y \sim gama(8,168; 7,618) \text{ e } \theta^s|y \sim gama(5,001; 8,618)$$

2.1.4 Intervalos Bayesianos

Diferentemente do que ocorre no intervalo de confiança clássico onde o parâmetro é considerado um valor fixo e desconhecido e o intervalo é considerado aleatório, na abordagem Bayesiana ocorre o contrário, o parâmetro é considerado aleatório (tem uma distribuição de valores possíveis) e o intervalo fixo (não depende de uma amostra aleatória).

A interpretação na abordagem Bayesiana de um intervalo de credibilidade de $100(1-\alpha)\%$ significa que a probabilidade *a posteriori* de que o parâmetro esteja no intervalo é de $100(1-\alpha)\%$. Já na abordagem Clássica, um intervalo de confiança de $100(1-\alpha)\%$ significa dizer que $100(1-\alpha)\%$ dos intervalos de confiança calculados incluirão o verdadeiro valor do parâmetro, visto que ele é fixo e desconhecido. Mas ambas as abordagens são usados para fins similiares.

Seja $f(\theta|y)$ a função *a posteriori* marginal da distribuição θ e $100(1-\alpha)\%$ o intervalo de credibilidade, para bicaudal $f(\theta^{\frac{\alpha}{2}}|y) = \frac{\alpha}{2}$ e $f(\theta^{1-\frac{\alpha}{2}}|y) = 1-\frac{\alpha}{2}$. O intervalo é obtido calculando os percentis empíricos de $\frac{\alpha}{2}$ e $1-\frac{\alpha}{2}$ da distribuição *a posteriori*.

Outro intervalo de credibilidade frequentemente usado na abordagem Bayesiana é o intervalo HPD (Highest posterior density). O intervalo HPD é necessário satisfazer duas condições: a probabilidade da distribuição *a posteriori* pertence a região $100(1-\alpha)\%$ e a densidade para todo ponto dentro do intervalo é igual ou maior do que para todo ponto não pertencente a ela, considerando o menor intervalo possível. para mais detalhes veja [Box and Tiao(1992)].

2.2 Métodos Monte Carlo via Cadeia de Markov na abordagem Bayesiana

Nesta seção, será introduzido o conceito básico de simulação, cadeia de Markov, geração de variáveis aleatórias discretas e contínuas e o método de Monte Carlo via Cadeia de Markov.

2.2.1 Simulação

O estudo de simulação surgiu para resolver problemas envolvendo modelos probabilísticos que não podem ser resolvidos analiticamente. O princípio é gerar amostras de variáveis aleatórias para representar o comportamento de fenômenos não determinísticos.

Simulação de variáveis aleatórias discretas

Existem várias técnicas para se gerar números aleatórios, antigamente eles eram gerados manualmente ou por meios mecânicos como roletas e dados. Atualmente são usados computadores para gerar números pseudo-aleatórios. Os números pseudo-aleatórios têm esse nome porque são gerados por mecanismos determinísticos que simulam números gerados de alguma distribuição de probabilidade.

Para gerar variáveis aleatórias discretas, independente da distribuição, é usada como base a distribuição uniforme com os parâmetros [0,1] e será denotada por $u \sim U(0,1)$.

Seja X uma variável aleatória com a seguinte distribuição de probabilidades:

$$\begin{aligned} P(X \leqslant x_{(i)}) &= P(X = x_{(1)}) + P(X = x_{(2)}) + ... + P(X = x_{(k)}) \\ p_i &= p_1 + p_2 + ... + p_k \end{aligned}$$

Em que $\sum_i p_i = 1$ e o intervalo $[0,1]$ é dividido em k intervalos $I_1, ..., I_k$ com $I_i = (F_{i-1}, F_i]$ onde $F_0 = 0$ e $F_i = p_1 + ... + p_i,\ i = 1, ..., k$.

Gera-se o número aleatório u e verifica-se em qual intervalo I_i pertence u. O método gera valores da distribuição de X, já que

$$P(X = x_i) = P(u \in I_i) = F_i - F_{i-1} = p_i$$

Para a distribuição uniforme discreta, basta dividir as unidades de intervalo em subintervalos k de largura igual a $1/k$ e $F_i = i/k$. A função de distribuição acumulada (f.d.a) de X é dada por:

$$X = \begin{cases} x_1, & u < p_1 \\ x_2, & p_1 \leqslant u < p_1 + p_2 \\ \vdots & \vdots \\ x_k, & p_1 + ... + p_{k-1} \leqslant u < p_1 + ... + p_k \end{cases}$$

Distribuição Bernoulli

Se X segue uma distribuição de Bernoulli $X \sim bern(p)$, com probabilidade de sucesso dada por p. As probabilidades $P(X = 1) = p$ e $P(X = 0) = 1 - p$. $0 < p < 1$. Com função de probabilidade dada por:

$$f(i) = P(X = i) = p(1 - p)^{1-x} \tag{2.6}$$

Neste caso $k = 2, x_1 = 1, x_2 = 0$ e $F_1 = p$. Para gerar valores dessa distribuição o intervalo será divido em duas partes:

Se $u \leqslant p$, então x receberá o valor 1;
Se $u > p$, então x receberá o valor 0.

Distribuição Binomial

Seja Y uma variável aleatória com distribuição Binomial, denotada por $bin(n, p)$. Com função de probabilidade dada por:

$$f(i) = P(Y = i) = \binom{n}{i} p^i (1 - p)^{n-i}, i = 0, 1, ..., n \tag{2.7}$$

Então Y é a soma do número n de sucessos em um experimento de Bernoulli, com probabilidade de sucesso $P(X = 1) = p$. Assim, é mais fácil fazer uma amostra X da distribuição de Bernoulli, pois $X_1, ..., X_n$ são independe e identicamente distribuídos, portanto

$$Y = \sum_i X_i \sim bin(n, p) \tag{2.8}$$

Os valores de x da binomial são obtidos por uma amostra $u_1, ..., u_n$ de uma distribuição $U(0, 1)$ contando o número de x destes n valores gerados, em que $x \leqslant p$.

Simulação de variáveis aleatórias continuas

Pode-se gerar algumas variáveis aleatórias contínuas a partir de um gerador de números aleatórios e, a da função inversa da distribuição acumulada.

Seja a v.a. contínua X, com f.d.a. $F(X)$ e o número aleatório u produzido pelo gerador. Então a variável aleatória X será a $F^{-1}(u)$.

Deste modo pode-se gerar variáveis aleatórias de funções de densidade $f(x)$ apenas achando sua $F^{-1}(X)$.

Se X for uma v.a. com f.d.a. F, então a v.a. $U = F(X)$ tem distribuição uniforme no intervalo [0,1]. Então para uma distribuição $G(u)$ tem-se,

$$G(u) = P(U \leqslant u) = P(F(X) \leqslant u) = P(X \leqslant F^{-1}(u)) = F(F^{-1}(u)) = u \qquad (2.9)$$

Distribuição Normal

Se X segue uma distribuição Normal, denotada por $X \sim N(\mu, \sigma^2)$, com função de densidade dada por

$$f(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}} \qquad (2.10)$$

Para gerar qualquer variável aleatória com distribuição normal é preciso apenas gerar uma v.a. normal padrão $N(0, 1)$ e padronizá-la, utilizando:

$$y = \frac{x - \mu}{\sigma} \qquad (2.11)$$

Então se é gerado um número aleatório y_1 de $Y \sim N(0, 1)$ e queremos uma variável aleatória de $X \sim N(\mu, \sigma^2)$, basta usar a relação $x_1 = \mu + \sigma y_1$.

Para se gerar uma $N(0, 1)$ a partir de uma amostra $u_1, ..., u_n$, da distribuição uniforme $U(0, 1)$ utiliza-se

$$X = \sqrt{n}\left(\frac{\bar{u} - \frac{1}{2}}{\frac{1}{\sqrt{2}}}\right) \sim N(0, 1) \qquad (2.12)$$

Onde $\bar{u} = \frac{1}{n}\sum_{i=1}^{n} u_1$. Para maiores detalhes veja [Bussab and Moretin(2004)].

Outra forma mais eficiente para algoritmos computacionais foi apresenta por [Box and Muller(1958)] em que se gera duas variáveis aleatórias com distribuição

$N(0,1)$ e independentes por meio de transformações

$$x_1 = \sqrt{-2log(u_1)}cos(2\pi u_2)$$
$$x_2 = \sqrt{-2log(u_1)}sen(2\pi u_2)$$

Onde u_1 e u_2 são v.a.'s com distribuição uniforme em [0,1].

Distribuição Gama

Seja X uma variável aleatória com distribuição gama $X \sim G(\alpha, \beta)$. Com função de densidade dada por:

$$f(x;\alpha,\beta) = \frac{\beta^\alpha}{\Gamma(\alpha)}x^{\alpha-1}e^{-\beta x}, x > 0 \tag{2.13}$$

Para gerar variáveis aleatórias da gama pode-se fazer algumas transformações para outros modelos que tornam os cálculos mais fáceis. Pode-se gerar variáveis aleatórias da distribuição gama extraindo amostras de uma distribuição exponencial, $Z \sim Exp(\lambda)$, de tamanho n, onde $n = \alpha$ inteiros e somando os valores da amostra. Logo,

$$X = \sum_{i=1}^{n} Z_i = z_1 + ... + z_n \sim G(n, \lambda)$$

Se $Y_1, ..., Y_n$ são amostras de uma distribuição normal, $N(\mu, \sigma^2)$, existe a seguinte relação

$$X = Y_1^2 + ... + Y_n^2 \sim \chi_n^2 \tag{2.14}$$

E, consequentemente, a soma de $Y_1^2, ..., Y_n^2$ tende para uma distribuição qui-quadrado (χ_n^2) com n graus de liberdade, é uma gama com os parâmetros $\alpha = \frac{n}{2}$ e $\beta = \frac{1}{2}$, isto é, $\chi_n^2 \sim G\left(\frac{n}{2}, \frac{1}{2}\right)$. Assim, para gerar variáveis aleatórias da distribuição gama é necessário apenas gerar números aleatórios de uma distribuição normal padrão com amostra de tamanho n e fazer o $\sum_{i=1}^{n} X^2$, então tem-se

$$X = \sum_{i=1}^{n} Y^2 \sim G\left(\frac{n}{2}, \frac{1}{2}\right) \tag{2.15}$$

2.2.2 Cadeia de Markov

Uma cadeia de Markov é um processo estocástico $\{X_t : t \in T\}$ tal que a distribuição de X_t para todos os valores anteriores $X_{t-1}, X_{t-2}, ..., X_0$ depende unicamente do seu anterior X_{t-1}. Isto é, para qualquer subconjunto A, tem-se

$$P(X_t \in A|X_0, ..., X_{t-1}) = P(X_t \in A|X_{t-1}) \tag{2.16}$$

Se $t \to \infty$ a distribuição X_t converge para uma distribuição de equilíbrio, no qual os valores finais são independentes do valores iniciais da cadeia.

Monte Carlo via Cadeias de Markov (MCMC) são métodos de simulação baseados em Cadeias de Markov ergódicas, cuja a distribuição estacionária do processo estocástico é a distribuição *a posteriori* de interesse. Maiores detalhes serão vistos na subseção 2.2.6.

Os métodos MCMC requerem ainda que a cadeia seja irredutível, aperiódica e não recorrente. Isto é, cada estado da cadeia pode ser atingido a partir de qualquer outro em um número finito de iterações, as probabilidades de transição de um estado para outro são invariantes e não existe estados absorventes. Os algoritmos que serão vistos aqui satisfazem a estas condições.

2.2.3 Integração de Monte Carlo

O nome Monte Carlo foi sugerido por Nicholas Metropolis inspirado em um tio de Ulam [Metropolis and Ulam(1949)], que sempre pegava dinheiro emprestado com parentes para ir até Monte Carlo jogar. Monte Carlo é uma cidade no Principado de Mônaco. A contribuição de Ulam foi a de reconhecer o potencial dos computadores eletrônicos recém inventados para automatizar as amostragens.

Existem vários métodos na literatura para resolver integrais, muitos baseados em aproximações e métodos computacionais intensivos. Um método de aproximação é o método de Monte Carlo, que será explicado a seguir. Seja uma integral

$$I = \int_D g(x)dx$$

Sabe-se que a esperança de uma função aleatória $g(x)$ é o valor médio dessa função, isto é, o conjunto de amostras pertencentes ao domínio da função e distribuídos com probabilidade $f(x)$. Portanto,

$$E(g(x)) = \int_D g(x)f(x)dx \tag{2.17}$$

Pela *lei dos grandes números* a esperança de uma função é aproximado pela média de um número de amostras aleatórias, pois ela converge quando tende ao infinito. Ou seja,

$$E(g(x)) \approx \frac{1}{T}\sum_{t=1}^{T} g(x_i),\ T \to \infty \tag{2.18}$$

Igualando as expressões 2.17 com a 2.18 tem-se

$$\begin{aligned} \int_D g(x)f(x)dx &\approx \frac{1}{T}\sum_{t=1}^{T} g(x_t) \\ \int_D \frac{g(x)f(x)}{f(x)}dx &\approx \frac{1}{T}\sum_{t=1}^{T} \frac{g(x_t)}{f(x_t)} \\ \int_D g(x)dx &\approx \frac{1}{T}\sum_{t=1}^{T} \frac{g(x_t)}{f(x_t)} \end{aligned} \tag{2.19}$$

Percebe-se que para integrar $g(x)$ em um domínio D basta gerar variáveis aleatórias $X_1, X_2, ..., X_t$ da função de densidade $f(x)$ e calcular a média da amostra. Assim, a integração de Monte Carlo consiste em estimar $\hat{I}$ da seguinte maneira:

$$\hat{I} = \frac{1}{T}\sum_{t=1}^{T} \frac{g(x_t)}{f(x_t)} \tag{2.20}$$

Esse método é aplicado para vários problemas em Inferência Bayesiana, pois *a posteriori* pode ser complicada e a única forma de amostrar valores dela é usando métodos estocásticos como MCMC. A geração de valores para a distribuição *a posteriori* é importante porque o único "estimador Bayesiano"é *a posteriori*, ou seja, toda inferência é feita observando os valores de tal distribuição. Assim se o interesse for calcular a média e variância da distribuição *a posteriori*, basta gerar amostras aleatórias da distribuição *a posteriori* $f(\theta|y)$ e calcular a média amostral de $G(\theta)$. Para uma abordagem mais precisa [Gamerman and Lopes(2006), p. 95].

2.2.4 Métodos Monte Carlo via Cadeia de Markov

Atualmente, usar Estatística Bayesiana sem usar métodos de Monte Carlo via Cadeias de Markov (MCMC) é muito difícil. Isso ocorre devido à complexidade dos modelos que os dados reais exigem. Na abordagem Bayesiana, quando não se sabe a família conjugada de distribuições *a priori* é necessária a utilização do método MCMC. O método de Monte Carlo via Cadeias de Markov é baseado em simulações iterativas, que por sua vez são ancoradas nas Cadeias de Markov, tendo como objetivo obter uma amostra da distribuição conjunta dos parâmetros de interesse, ou seja, gerar uma amostra da distribuição *a posteriori* e calcular estimativas amostrais desta distribuição ou até mesmo intervalos empíricos de credibilidade.

Os algoritmos desenvolvidos nestes trabalho são determinísticos baseados em iterações de cadeias de Markov, mas com passos que dependem de números pseudo-aleatório. A escolha do algoritmo eficiente está relacionada com o tempo em que a cadeia demora para esquecer os valores anteriores, buscando atingir a estacionariedade. Dois pontos importantes em MCMC é o *burn-in* e o *thin*. O *burn-in* é o espaço que a cadeia de Markov necessita para chegar à distribuição estacionária, ele é considerado o período de aquecimento da cadeia, esse período sofre influência do estado inicial e por isso deve ser descartado. O *thin* é o lag que, se houver autocorrelação positiva, é necessário para que a observações sucessivas sejam independentes, ele serve para diminuir a dependência entre as observações subsequentes ao longo da cadeia, diminuindo a autocorrelação e armazenando valores a cada k iterações.

Segundo [Ntzoufras(2009)], para gerar uma amostra de $f(\theta|y)$ é necessário satisfazer duas propriedades: Primeiro que a cadeia de Markov satisfaça $(f(\theta^{(t+1)}|\theta^{(t)}))$; e segundo que a distribuição de equilíbrio da cadeia de Markov seja a distribuição *a posteriori* de interesse $f(\theta|y)$.

A generalização do processo MCMC:

1. Selecione um valor inicial $\theta^{(0)}$ do vetor $\underset{\sim}{\theta}^{(t)}$
2. Gere $\{T : t = 1, ..., T\}$ iterações até a distribuição de equilíbrio ser atingida
3. Monitore a convergência do algoritmo usando diagnósticos de convergência. Se no diagnóstico a convergência falhar, é preciso gerar mais observações.

4. Realize o *burn-in*, corte as primeiras observações, $\underset{\sim}{\theta}^{(t)} = \{\theta^{B+1}, \theta^{B+2}, ..., \theta^{T}\}$, onde B é o tamanho do *burn-in*

5. Realize o *thin* e conside $\underset{\sim}{\theta}^{(t)} = \{\theta^{B+1k}, \theta^{B+2k}, ..., \theta^{T}\}$ como a amostra para análise posterior

6. Trace a distribuição posterior (Geralmente, o foco está sobre as distribuições marginais univariadas)

7. Finalmente, obtenha estatísticas da distribuição posterior (média, mediana, desvio padrão, quantis, correlações)

2.3 Algoritmos populares dos MCMC

Entre os Métodos de Monte Carlo via Cadeia de Markov, os dois mais utilizados são o algoritmo de Metropolis-Hasting [Metropolis et al. (1953)] e [Hastings(1970)] e o Amostrador de Gibbs [Geman and Geman(1984)] detalhados nesta seção.

2.3.1 Amostrador de Gibbs

O amostrador de Gibbs foi apresentado inicialmente por [Geman and Geman(1984)], logo depois [Gelfand and Smith(1990)] foram os primeiros autores a mostrar que o Amostrador de Gibbs poderia ser usado para uma série de outras distribuições posteriores.

A vantagem do Amostrador de Gibbs é tornar um problema multivariado numa sequência de problemas univariados, para o qual existe uma grande variedade de ferramentas computacionais.

Se a distribuição conjunta existir, é possível determiná-la apenas conhecendo suas distribuições condicionais utilizando o Amostrador de Gibbs. A função *a posteriori* condicional é dada por $f(\theta_j|\theta_{\setminus j}, y)$, onde $\theta_{\setminus j} = (\theta_1, ..., \theta_{j-1}, \theta_{j+1}, ..., \theta_d)^T$.

No passo (2) do caso geral, na seção 2.2.4, são gerados $\{T : t = 1, ..., T\}$ iterações até a distribuição de equilíbrio ser atingida. Para isso é necessário resolver o seguinte algoritmo:

1. Defina os valores iniciais $\theta^{(0)} = (\theta_1^{(0)}, ..., \theta_d^{(0)})'$

2. Inicie o contador de iteração da cadeia com $t = 1$, onde $t = 1, ..., T$

3. Obtenha um novo valor θ_j de $\theta_j \sim f(\theta_j|\theta_{\setminus j}, y)$ através de valores sucessivos da geração de

$$\begin{array}{rcc}
\theta_1^{(t)} & \sim & f(\theta_1|\theta_2^{(t-1)}, \theta_3^{(t-1)}, ..., \theta_p^{(t-1)}, y) \\
\theta_2^{(t)} & \sim & f(\theta_2|\theta_1^{(t)}, \theta_3^{(t-1)}, ..., \theta_p^{(t-1)}, y) \\
\vdots & \vdots & \vdots \\
\theta_j^{(t)} & \sim & f(\theta_j|\theta_1^{(t)}, \theta_2^{(t)}, ..., \theta_{j-1}^{(t)}, \theta_{j+1}^{(t-1)}, ..., \theta_p^{(t-1)}, y) \\
\vdots & \vdots & \vdots \\
\theta_p^{(t)} & \sim & f(\theta_p|\theta_1^{(t)}, \theta_2^{(t)}, ..., \theta_{p-1}^{(t)}, y)
\end{array}$$

4. Troque o contador de t para $t + 1$ e retorne ao passo 3 até $t = T$

Geração de valores de $f(\theta_j|\theta_{\setminus j}, y) = (\theta_j|\theta_1^{(t)}, \theta_2^{(t)}, ..., \theta_{j-1}^{(t)}, \theta_{j+1}^{(t-1)}, ..., \theta_p^{(t-1)}, y)$ é relativamente fácil, uma vez que uma distribuição univariada pode ser escrita como $f(\theta_j|\theta_{\setminus j}, y) \propto f(\theta|y)$, onde todas as variáveis, exceto θ_j, são mantidas constantes nos seus dados valores. [Ntzoufras(2009), p. 72]. O conjunto de t valores amostrais armazenados em $\underset{\sim}{\theta}^{(t)}$ representa a distribuição conjunta *a posteriori*. Com os valores amostrais de $\theta^{(t)}$ pode-se obter estimativas como média, mediana e intervalo de credibilidade. Uma descrição mais detalhada do Amostrador de Gibbs é encontrada em [Casella and George(1992)].

2.3.2 Metropolis-Hastings

Os algoritmos de Metropolis-Hastings são de grande importância para a Inferência Bayesiana, pois, resumidamente, garantem a convergência da cadeia para a distribuição de equilíbrio, que é a distribuição *a posteriori*. Esse poderoso MCMC é de extrema importância na resolução de problemas multidimensionais.

Seja θ o parâmetro de interesse da distribuição $f(x)$, $f(\theta|y)$ a distribuição *a posteriori* e $\underset{\sim}{\theta}^{(t)}$ é um vetor que armazena os valores gerados da interação do algoritmo.

No passo (2) do caso geral, seção 2.2.6, são gerados $\{T : t = 1, ..., T\}$ iterações até a distribuição de equilíbrio ser atingida. Para isso é necessário resolver o seguinte algoritmo:

1. Escolha um valor inicial do vetor $\underset{\sim}{\theta}^{(t)} = \theta^{(0)}$

2. Gere um novo valor θ' da distribuição $q(\theta'|\theta)$

3. Calcule a probabilidade de aceitação α

$$\alpha = min\left(1, \frac{f(\theta'|y)q(\theta|\theta')}{f(\theta|y)q(\theta'|\theta)}\right) \tag{2.21}$$

4. Gere $u \sim U(0,1)$ de u_i com tamanho igual a T

5. Compare α com os números aleatórios μ. Se $u \leqslant \alpha$ então $\theta^{(t)} = \theta'$, caso contrário, $\theta^{(t)}$ receberá θ

6. Faça a comparação até T iterações.

A vantagem desse algoritmo é poder usar uma outra densidade mais simples $q(\theta)$ para gerar uma amostra de $f(\theta)$, pois a probabilidade (2.21) não se altera.

O algoritmo de Metropolis-Hastings converge para a sua distribuição de equilíbrio independentemente de qualquer distribuição proposta q. Entretanto, na prática, a escolha de uma distribuição q é de suma importância, uma escolha ruim pode atrasar consideravelmente a convergência para a distribuição de equilíbrio. O Amostrador de Gibbs, visto na seção anterior, é um caso especial do Metropolis-Hastings quando a densidade proposta é $q(\theta'|\theta^{(t)})$.

2.4 Diagnósticos de convergência

Existem métodos formais e informais de diagnosticar a convergência de uma cadeia. Os métodos formais são contemplados com técnicas gráficas e uma combinação de critérios univariados e multivariados de avaliações das cadeias de Markov. Existem diversos critérios de verificar a convergência pelo método formal, os critérios abordados nessa seção são os de [Gelman and Rubin(1992)], [Heidelberger and Welch(1983)], [Geweke(1992)] e [Raftery and Lewis(1992)].

Os métodos informais foram introduzidos por [Gelfand and Smith(1990)] e consistem em técnicas gráficas para verificar a convergência: gráficos de média ergótica, série temporal, função de autocorrelação (FAC) e função de autocovariância (FACV). Os métodos informais são tratados na subseção 2.4.5.

2.4.1 Critério de Gelman-Rubin

O diagnóstico de [Gelman and Rubin(1992)] é baseado na análise de múltiplas cadeias. Após a convergência para a distribuição de interesse, as amostras obtidas em cada cadeia não apresentam diferenças significativas. A análise consiste na comparação dos desvios de cada cadeia e entre as cadeias. Desvios grandes entre as duas cadeias indicam a não convergência.

Seja $\left\{\underset{\sim}{\theta}^{(t)}\right\}$ as amostras de uma única cadeia e M a quantidade de cadeias simuladas, resultando em uma sequência de amostras $\left\{\theta_m^{(t)}\right\}_{t=1,...,M}$, $t = 1, ..., n$ e $m = 1, ..., M$. Então a variância entre as cadeias é dada por

$$B = \frac{n}{M-1}\sum_{m=1}^{M}(\bar{\theta}_{\cdot m} - \bar{\theta}_{\cdot\cdot})^2 \tag{2.22}$$

Onde $\bar{\theta}_{\cdot m}$ é a média amostral dos elementos da cadeia m e $\bar{\theta}_{\cdot\cdot}$ é a média amostral de todos os valores amostrados.

E a variância amostral das cadeias é dada por

$$W = \frac{1}{M}\sum_{m=1}^{M} s_m^2 \tag{2.23}$$

Onde s_m^2 é a variância das amostras de cada cadeia, ou seja

$$s_m^2 = \frac{1}{n-1}\sum_{t=1}^{n}(\theta_m^t - \bar{\theta}_{\cdot m})^2$$

A variância marginal da distribuição posterior, $var(\theta|y)$, é a média ponderada de W e B e sua estimativa é dada por

$$\hat{V} = \frac{n-1}{n}W + \frac{M+1}{nM}B \tag{2.24}$$

Se todas as M cadeias tiverem convergido, a estimativa $\hat{V}$ será muito próximo da variância amostral das cadeias W. Portanto, a razão $\dfrac{\hat{V}}{W}$ estará próximo de 1. Porém, se isso não ocorre, $\hat{V}$ ficará superestimado. Para detectar esse problema é calculado a raiz quadrada dessa razão, conhecida como *potential scale reduction factor* (PSRF). Então, um PSFR grande indica que a variância entre as cadeias é essencialmente maior que a variação dentro das cadeias, sendo necessário aumentar o número de simulações. E, se, o PSRF estiver próximo de 1, pode-se concluir que as M cadeias convergiram

e, consequentemente, foi encontrada a distribuição de interesse. O calculo do PSRF é dado por

$$\hat{R}_c = \sqrt{\frac{\hat{d}+3}{\hat{d}+1}\frac{\hat{V}}{W}}, \quad \text{onde} \quad \hat{d} = \frac{2\hat{V}^2}{\widehat{Var}(\hat{V})} \tag{2.25}$$

$\hat{R}$ sempre será maior que 1 e tende para 1 a medida que $M \longrightarrow \infty$. Os autores sugerem que valores de $\hat{R}$ a baixo de $1,2$ indicam convergência.

2.4.2 Critério de Geweke

O critério de [Geweke(1992)] consiste em dividir a cadeia de Markov em duas partes e testar se os valores da primeira parte da cadeia são iguais em média aos valores da segunda parte. Se a distribuição é estacionária, espera-se que a média da primeira parte seja igual a segunda. Geralmente, a primeira parte corresponde aos primeiros 10% das iterações após o período de *burn in* e a segunda, parte os últimos 50% das iterações da cadeia. A estatística do teste possui distribuição assintótica normal.

Seja a cadeia de Markov $\{\theta^t\}$, duas subsequenciais são extraídas $\{\theta_1^t : t = 1, ..., n_1\}$ e $\{\theta_2^t : t = n_a, ..., n\}$, onde $1 < n_1 < n_a < n$ e $n_2 = n - n_a + 1$. Então, as médias das subsequencias são

$$\bar{\theta}_1 = \sum_{t=1}^{n_1} \frac{\theta^t}{n_1} \text{ e } \bar{\theta}_2 = \sum_{t=n_a}^{n} \frac{\theta^t}{n_2}$$

E com respectivas variâncias estimadas, S_1^2 e S_2^2. A estatística do teste é dada por

$$Z_{n=}\frac{\bar{\theta}_1 - \bar{\theta}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \tag{2.26}$$

Admitindo-se que $\frac{n_1}{N}$ e $\frac{n_2}{N}$ são fixos, a distribuição da estatística converge para uma $N(0,1)$ quando $N \longrightarrow \infty$. Se o *valor p* obtido for maior que o nível de significância pré-fixado, então não existe evidências contra a convergência dos parâmetros.

2.4.3 Critério de Heidelberger-Welch

O diagnóstico de [Heidelberger and Welch(1983)] avalia a estacionariedade da cadeia e se o tamanho da amostra é adequado para estimar a média com precisão. Se a cadeia não chegar a estacionariedade, descarta-se 10% das iterações iniciais e o teste

é repetido. Se não chegar na estacionariedade novamente, descarta-se mais 10% das iterações iniciais. Esse processo é repetido, no máximo, 5 vezes, a partir da sexta rejeição é considerado falha de estacionariedade e indica que é necessário aumentar o número de iterações. Se o teste chegar a estacionariedade, a parte descartada é considerada o tamanho do *burn in*.

Para verificar se a média foi calculada com precisão, calcula-se o intervalo de confiança utilizando o desvio padrão assintótico. Faz-se o seguinte teste da razão

$$\frac{\text{(limite superior - limite inferior)}}{2}$$

Se o resultado for menor do que $0,1$ é considerado adequado para um nível de confiança de 95%, ou seja, a média foi calculada com acurácia. Caso contrário, concluí-se que não há dados suficientes para estimar com precisão a média da cadeia.

2.4.4 Critério de Raftery-Lewis

O critério proposto por [Raftery and Lewis(1992)] determina o período de *burn-in* e a distância mínima (k) de uma interação à outra para diminuir a autocorrelação amostral *thin*. Segundo os autores, o fator de convergência é responsável pelo valor multiplicativo ao número de iterações necessárias para chegar a convergência, se esse valor for menor do que 5 a convergência é obtida.

2.4.5 Outros critérios

Quando o algoritmo converge, a distribuição de interesse é atingida, isto é, a distribuição *a posteriori* alcançou, aproximadamente, sua distribuição estacionária. A escolha do algoritmo eficiente está relacionada com o tempo em que a cadeia demora para esquecer os valores anteriores, ou seja, as interações de aquecimento da cadeia. A seguir serão apresentados alguns métodos de monitoramento de convergência.

Gráfica de série temporal

Uma forma empírica de monitorar a cadeia é desenhar gráficos de série temporal. A figura a seguir mostra um exemplo de gráfico de série temporal.

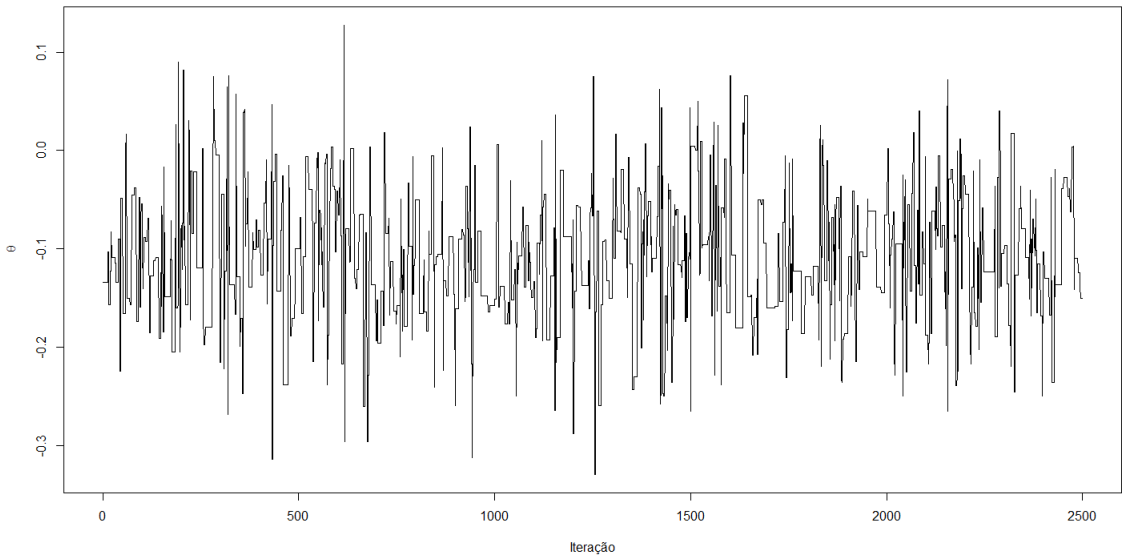

Figura 2.3: Gráfico de série temporal

Pode-se notar que há indícios de que a cadeia converge, os valores parecem oscilar em torno de uma média, sem aspectos de tendência.

Gráfico de média ergódica

Outra gráfico importante é o de média ergódica. A média ergódica (M_{eg}) é a média até o instante atual da cadeia. Se a média ergódica estabilizar após algumas iterações, então há indicativos de convergência do algoritmo.

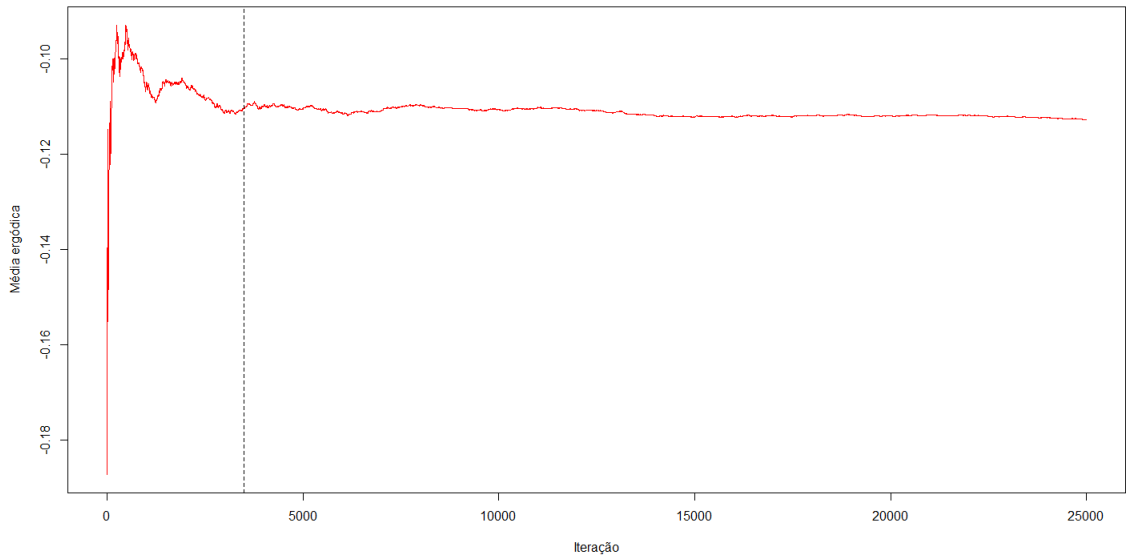

Figura 2.4: Gráfico de média ergódica

A linha pontilhada mostra o período de aquecimento da cadeia. A figura a baixo mostra a média ergódica antes e depois do período de aquecimento do algoritmo.

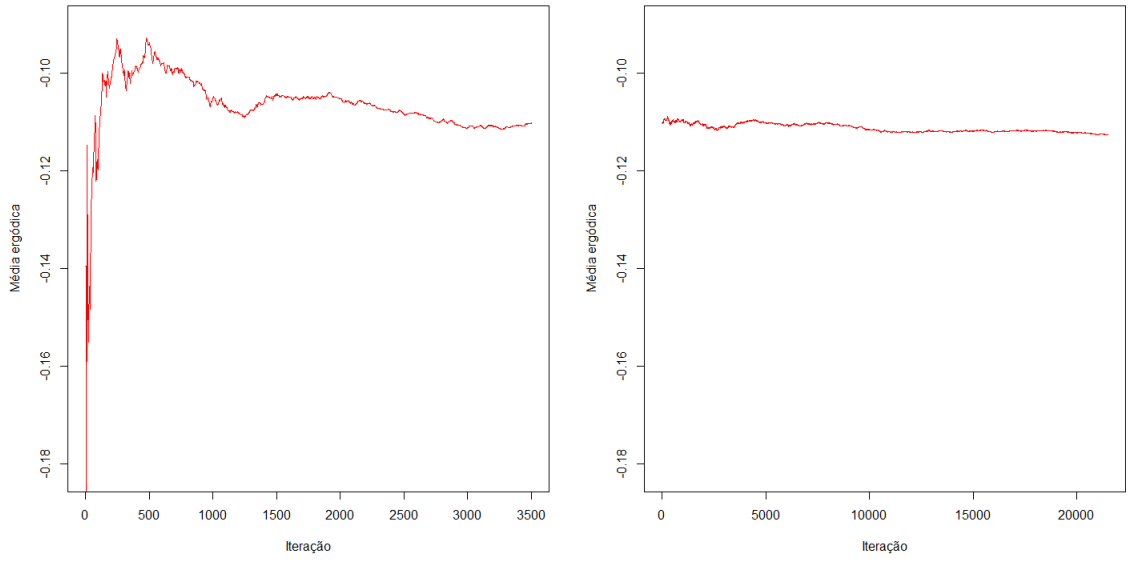

Figura 2.5: Gráfico de média ergódica antes e depois do período de aquecimento

Função de autocorrelação e autocovariância

Na definição de cadeia de Markov, os valores gerados são correlacionados ao valor anterior ao longo das iterações. Por esta razão, é preciso monitorar as autocorrelações dos valores gerados. Quanto maior a correlação, menor será o ganho de informação dos valores armazenados da cadeia, o que acarreta em uma amostra não representativa e um desperdício de espaço em disco, assim sendo, para a cadeia convergir para a distribuição estacionária o número de iterações tem que ser muito grande. A função de autocorrelação (FAC) é utilizada para verificar a velocidade de convergência do algoritmo e a função de autocovariância (FACV) para verificar a estacionariedade. Para maiores detalhes de como calcular a FAC e a FACV veja [Ehlers(2003), p. 14].

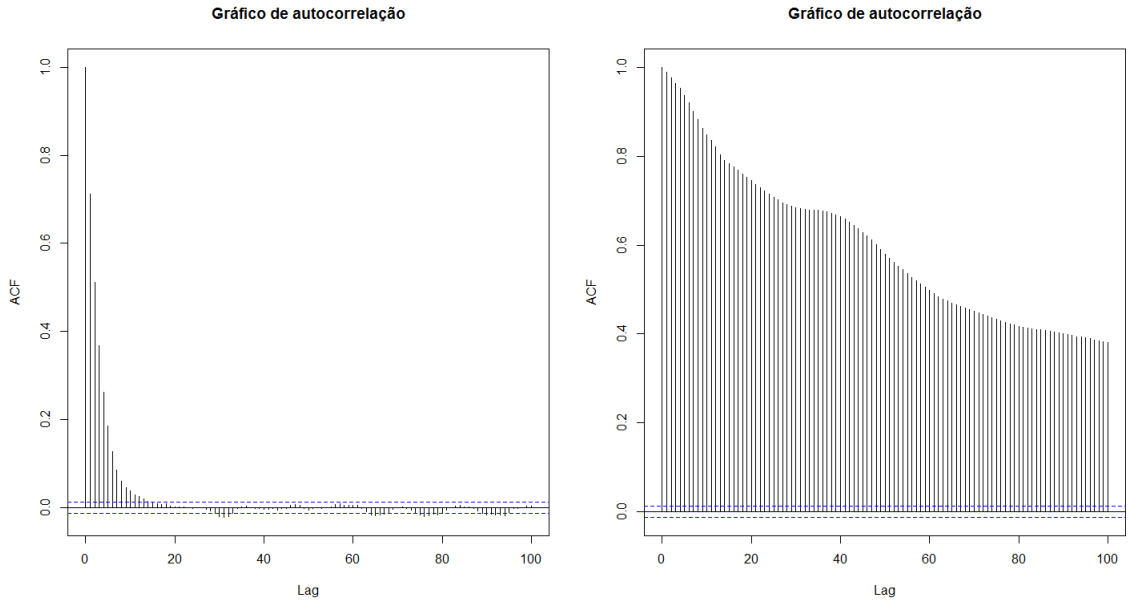

Figura 2.6: Funções de autocorrelação do algoritmo, na figura a esquerda os dados estão poucos correlacionados e na figura da direita os dados estão muito correlacionados. Quanto menos correlacionados estão os dados mais rápido o algoritmo tende a estacionariedade.

Na figura acima são apresentados gráficos de autocorrelação da cadeia para os primeiros 100 lags. No primeiro caso, percebe-se que a função de autocorrelação chega a estacionariedade rapidamente, já no segundo caso, os valores estão muito correlacionados, todavia, com tendência de estacionariedade, necessitando de um aumento na

quantidade de iterações.

Uma outra abordagem, conhecida como *thin*, consiste em guardar os valores simulados a cada k iterações após o período de aquecimento. Então, a amostra para a análise posterior consiste em $\{\theta^{B+k}, \theta^{B+2k}, ..., \theta^{B+Tk}\}$, onde o tamanho do vetor θ será igual a T. Em problemas de maior dimensão essa abordagem é usada para economizar espaço de armazenamento ou velocidade computacional.

Um exemplo está na figura a seguir, foram feitas 25 mil simulações. Após o período de aquecimento utilizou-se o *thin* de $k = 20$. O corte foi feito para os primeiros 5 mil valores, então o vetor da amostra θ é igual a $\{\theta^{5020}, \theta^{5040}, ..., \theta^{25000}\}$ e tem o tamanho de 1000 iterações. No primeiro gráfico a função de autocorrelação com todas as iterações e no segundo gráfico, a FAC da amostra de 1000.

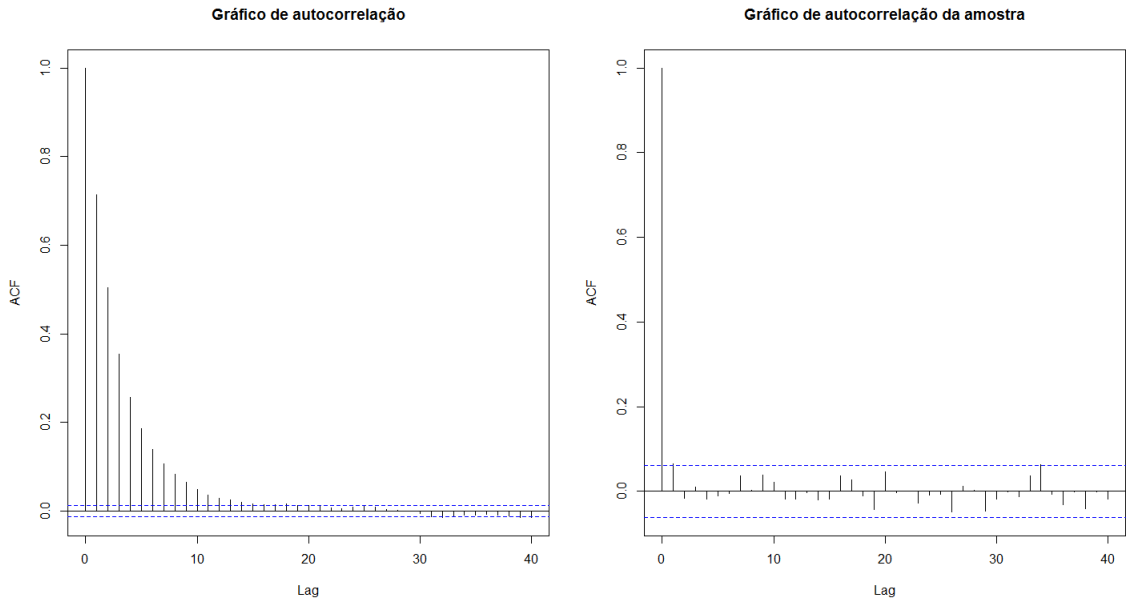

Figura 2.7: Funções de autocorrelação do algoritmo, a esquerda tem-se a FAC antes e a direita, a FAC da amostra

Percebe-se que no gráfico da esquerda os valores são correlacionados e decrescem rapidamente, já os valores da direita são estacionários, o que representa melhor a distribuição de interesse. Esse critério é importante quando não é possível eliminar a correlação entre os elementos da amostra obtida.

2.5 Introdução aos Modelos Lineares Generalizados

O estudo da modelagem iniciou-se antes na década de 70, porém ela só foi unificada por [Nelder and Wedderburn(1972)] em seu artigo de nome *Generalized linear models*. Os autores definiram que os Modelos Lineares Generalizados (MLG) consistem basicamente em três componentes: a variável resposta Y, as variáveis explicativas $X_1, X_2, ..., X_p$ e os mecanismos de ligação entre o conjuntos dessas duas variáveis.

A variável resposta $Y_i,\ i = 1, ..., n$ é a variável aleatória independente e tem distribuição pertencente à família exponencial em sua forma canônica, com média u_i e parâmetro de dispersão constante $\phi > 0$. Por isso ela é conhecida como a **parte aleatória**.

As variáveis explicativas $X_i^t = (X_1, X_2, ..., X_{qi})$, $i = 1, ..., n$ são chamadas de **parte sistemática**, pois associam-se ao modelo junto ao vetor de preditores lineares dado por

$$\eta_i = x_i^T \beta \tag{2.27}$$

O terceiro componente é a **função de ligação**, que associa a parte aleatória com a parte sistemática, e, portanto,

$$g(\mu_i) = x_i^T \beta \Rightarrow \mu_i = g^{-1}(x_i^T \beta) \tag{2.28}$$

Se a função de probabilidade da variável aleatória Y_i puder ser escrita por

$$f(y_i; \theta_i, \phi) = exp\left\{ \frac{1}{a(\phi)}[y_i\theta_i - b(\theta_i)] + c(y_i, \phi) \right\} \tag{2.29}$$

Com média

$$E(Y_i) = \mu_i = b'(\theta_i) \tag{2.30}$$

E variância

$$Var(Y_i) = a_i(\phi)b''(\theta_i) = a_i(\phi)V(\mu_i) = a_i(\phi)V_i \tag{2.31}$$

Então Y_i tem distribuição pertencente à família exponencial na forma canônica, em que $a(.)$, $b(.)$ e $c(.)$ são funções específicas e θ_i é um parâmetro canônico. Geralmente é utilizado $a_i(\phi) = \frac{\phi}{w_i}$ em que ϕ é o parâmetro de dispersão ou escala, w_i um peso *a priori* e $V_i = \frac{d\mu_i}{d\theta_i}$.

Seja $y^T = (y_1, y_2, ..., y_n)$ observações da variável aleatória Y_i pertencente à família exponencial, então a função de verossimilhança é dada por

$$\begin{aligned} L(\theta, \phi; y) &= \prod_{i=1}^{n} f(y_i; \theta_i, \phi) \\ &= exp\left\{ \sum_{i=1}^{n} \left[\frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y_i, \phi) \right] \right\} \\ &= exp\left\{ \frac{n\bar{y}\theta - nb(\theta)}{a(\phi)} + \sum_{i=1}^{n} c(y_i, \phi) \right\} \end{aligned} \tag{2.32}$$

Se a distribuição a priori for do tipo

$$f(\theta|\theta_0, \tau_0, \phi) \propto exp\left(\frac{\theta\theta_0 - \tau_0 b(\theta)}{a(\phi)} \right) \tag{2.33}$$

Se ambas as distribuições são da família exponencial, então a distribuição *a posteriori* terá *a priori* conjugada, sendo necessária apenas a atualização dos parâmetros $\tilde{\theta} = n\bar{y} + \theta_0$ e $\tilde{\tau} = n + \tau_0$ (7.5). E com a distribuição *a posteriori* dada por

$$f(\theta|y, \phi) \propto exp\left(\frac{(n\bar{y} + \theta_0)\theta - (n + \tau_0)b(\theta)}{a(\phi)} + \sum_{i=1}^{n} c(y_i, \phi) \right) \tag{2.34}$$

Assumindo que o parâmetro de dispersão ϕ é conhecido e fixo. Nas próximas subseções, serão revisados algumas distribuições que fazem parte da família exponencial.

2.5.1 Distribuição Normal

Seja $Y \sim N(\mu, \sigma^2)$, com média μ, variância σ^2 e com função densidade de probabilidade dada por

$$\begin{aligned} f(y|\mu, \sigma_2) &= \frac{1}{\sigma\sqrt{2\pi}} exp\left\{ -\frac{1}{2}\frac{(y-\mu)^2}{\sigma^2} \right\} \\ &= exp\left\{ \frac{1}{\sigma^2}\left(y\mu - \frac{\mu^2}{2} \right) - \frac{1}{2}\left(\frac{y^2}{\sigma^2} + ln(2\pi\sigma^2) \right) \right\} \end{aligned} \tag{2.35}$$

Sendo $\sigma > 0$ e $\mu \in \mathbb{R}$. Utilizando a notação da família exponencial tem-se

$$\begin{aligned} \theta = \mu, &\qquad b(\theta) = \frac{\mu^2}{2} \\ a(\phi) = \sigma^2, &\qquad b'(\theta) = \mu \\ c(y, \phi) = -\tfrac{1}{2}\left(\tfrac{y^2}{\sigma^2} + ln(2\pi\sigma^2) \right) &\qquad b''(\theta) = 1 \end{aligned}$$

Portanto, a distribuição normal pertence a família exponencial com média dadas por (2.30) e variância (2.31), ou seja,

$$E(Y_i) = b'(\theta_i) = \mu$$
$$Var(Y_i) = a_i(\phi)b''(\theta_i) = \sigma^2$$

Em modelos de regressão normais, a variável Y é considerada uma variável aleatória contínua pertencente ao conjunto dos reais que segue a distribuição normal. Portanto, tem-se a seguinte equação

$$Y|X_1, ..., X_p \sim N(\mu(\beta, X_1, ..., X_p), \sigma^2) \tag{2.36}$$

Onde

$$\begin{aligned} \mu(\beta, X_1, ..., X_p) &= \beta_0 + \beta_1 X_1 + ... + \beta_p X_p \\ &= \beta_0 + \sum_{j=1}^{p} \beta_j X_{ij} \end{aligned}$$

O modelo também pode ser escrito por

$$Y_i = \beta_0 + \sum_{j=1}^{p} \beta_j X_{ij} + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma^2)\ i = 1, 2, ..., n \tag{2.37}$$
$$Y_i \sim N(\mu_i, \sigma^2) \tag{2.38}$$
$$\mu_i = \beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip}, i = 1, ..., n$$

A maneira mais simples é supor que os parâmetros tem distribuição *a priori* independe e dadas por

$$\begin{aligned} f(\beta, \sigma^2) &= f(\sigma^2) \prod_{j=0}^{p} f(\beta_j) \\ \beta_j &\sim N(\mu_{b_j}, c_j^2), j = 0, ..., p \\ \sigma^2 &\sim GI(a, b) \end{aligned}$$

2.5.2 Distribuição Poisson

Seja $Y \sim Poisson(\lambda)$, com parâmetro $\lambda > 0$ e função densidade de probabilidade dada por

$$f(y|\lambda) = \frac{e^{-\lambda}\lambda^y}{y!}, \text{ para } y = 0, 1, 2, ...$$

Desenvolvendo tem-se

$$f(y|\lambda) = exp\left[ylog\lambda - \lambda - log(y!)\right] \tag{2.39}$$

Portanto, para a notação da família exponencial tem-se

$$\begin{array}{cc} \theta = log(\lambda), & \lambda = e^{\theta}, \\ \phi = 1 & b(\theta) = \lambda = e^{\theta} \\ a(\phi) = 1 & b'(\theta) = e^{\theta} \\ c(y,\phi) = -log(y!) & b''(\theta) = e^{\theta} \end{array}$$

Portanto, a distribuição de poisson pertence a família exponencial com média e variâncias dadas por (2.30) e (2.31), respectivamente, ou seja,

$$\begin{aligned} E(Y_i) &= b'(\theta_i) = e^{\theta} \\ Var(Y_i) &= a_i(\phi)b''(\theta_i) = e^{\theta} \end{aligned}$$

O modelo de poisson é dado por

$$\begin{aligned} Y_i &\sim Poisson(\lambda_i) \\ log\lambda_i &= \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} \end{aligned} \tag{2.40}$$

2.5.3 Distribuição Gama

Seja $Y \sim gama\left(v, \frac{v}{\mu}\right)$, função densidade de probabilidade dada por

$$f(y|\mu, v) = \frac{1}{\Gamma(v)} \left(\frac{v}{u}\right)^{v} y^{v-1} exp\left(\frac{-yv}{\mu}\right),\ y > 0$$

Desenvolvendo tem-se

$$\begin{aligned} f(y|\mu, v) &= exp\left\{log\left(\frac{v}{u}\right)^{v} - log\Gamma(v) + (v-1)log(y) - \frac{yv}{\mu}\right\} \\ &= exp\left\{v\left[y\left(-\frac{1}{\mu}\right) - log(\mu)\right] + vlog(vy) - log\Gamma(v) - log(y)\right\} \end{aligned}$$

Portanto, para a notação da família exponencial tem-se

$$\theta = -\frac{1}{\mu}, \qquad \mu = -\frac{1}{\theta},$$
$$\phi = \frac{1}{v} \qquad b(\theta) = log(\mu) = -log(-\theta)$$
$$a(\phi) = \frac{1}{v} \qquad b'(\theta) = -\frac{1}{\theta}$$
$$c(y, \phi) = vlog(vy) - log\Gamma(v) - log(y) \qquad b''(\theta) = \frac{1}{\theta^2} = \mu^2$$

Portanto, a distribuição gama pertence a família exponencial com média e variâncias dadas por (2.30) e (2.31), respectivamente, ou seja,

$$E(Y_i) = b'(\theta_i) = -\frac{1}{\theta} = \mu$$
$$Var(Y_i) = a_i(\phi)b''(\theta_i) = \frac{1}{v}\mu^2 = \frac{\mu^2}{v}$$

3 Material e métodos

Os dados utilizados neste trabalho são relativos ao resveratrol, baseados no trabalho de dissertação de mestrado de [Carvalho(2013)]. Os experimentos foram conduzidos na Embrapa Recursos Genéticos e Biotecnologia, Brasília – DF. Nesta seção são feitos esclarecimentos sobre o resveratrol e sua importância, bem como a realização do experimento e seus objetivos

3.1 Resveratrol e sua importância

Uma das defesas naturais da planta mediante estresses bióticos e abióticos consiste na produção de fitoalexinas, que são caracterizadas como compostos de baixo peso molecular, muitos da classe dos estilbenos, com atividade antifúngica e antibacteriana (Ingham et al., 1976; Paxton et al., 1991; Sobolev et al., 1995 apud [Carvalho(2013)]). Um dos principais estilbenos produzido em vários tecidos da planta do amendoim é o resveratrol (Sobolev et al., 1995 apud [Carvalho(2013)]).

O resveratrol é uma fitoalexina envolvida na resposta da planta a estresses bióticos e abióticos (Chung et al., 2003 apud [Carvalho(2013)]). A análise de algumas cultivares de A. hypogaea demonstrou relação entre maiores teores de resistência e maiores teores de resveratrol (Sobolev et al., 2007 apud [Carvalho(2013)]). Portanto, o aumento dos teores de resveratrol no amendoim pode resultar em aumento de resistência, além do aumento de seu valor nutracêutico. Estudos anteriores mostram que o resveratrol

- Inibe o desenvolvimento de tumores;
- Protege contra doenças cardíacas;
- Protege contra danos cerebrais (isquemias); e
- Pode ser usado no tratamento contra diabetes.

3.2 Objetivos

Os objetivos são analisar e avaliar a variabilidade das espécies para a produção de resveratrol e a expressão de *resveratrol sintase* em função da produção de resveratrol após tratamento com ultravioleta (UV) sob a abordagem Clássica e a Bayesiana.

3.3 Experimento

Foram realizados dois tipos de experimentos, o de concentração de resveratrol (micrograma/grama de folha) e o de expressão relativa do *resveratrol sintase* (quatitativo de gene). Os experimentos consistem em amostras de quatro especies, sendo estas:

- A. hypogaea (cultivar runner, genoma AB);
- A. ipaënsis (acesso KGPScS 30076, genoma B);
- A. duranensis (acesso VNvEv 14167, genoma A); e
- Anfidiploide sintético (originado pelo cruzamento artificial de A. ipaënsis e de A. duranensis , genoma AB).

As sementes foram obtidas do Banco Ativo de Germoplasma de Arachis da Embrapa Recursos Genéticos e Biotecnologia, Brasília – DF, e foram cultivadas durante o período de dezembro de 2011 a março de 2012. Para os experimentos utilizou-se tecido foliar, pois este é produzido em abundância pela planta e é o tecido mais suceptível as doenças fúngicas, fator limitante para o aumento do cultivo no Brasil (Freitas et al., 2005 apud [Carvalho(2013)]).

O tempo de indução, tempo de coleta pós-indução e o tratamento utilizado para indução da produção de resveratrol foram escolhidos baseando-se em dados da literatura que comparam várias formas de indução e demonstram que a forma de indução mais eficiente na produção de resveratrol é por meio de tratamento com UV.

As folhas foram coletadas em casa de vegetação durante os meses de janeiro e fevereiro de 2012, com intervalo de vinte dias entre as repetições biológicas do experimento. Depois de coletadas, foram colocadas em sacos plásticos e umedecidas e no laboratório foram dispostas em duas bandejas que foram revestidas com uma camada

de algodão umedecida com 500 ml de agua, sob a qual foi colocada uma camada de folhas de papel germitest.

Uma bandeja foi colocada 50 cm abaixo da luz ultravioleta (UV-C) (Fluxo Laminar: Trox Modelo FLV série: 235-81, com Lâmpada Philips TUV 30W/ 630 TB LONGLIFE), por duas horas e trinta minutos, enquanto a outra (grupo controle) ficou em uma sala ao lado, livre da radiação. Após o período indicado, as duas bandejas foram reunidas em uma mesma sala, onde permaneceram por quinze horas à temperatura ambiente e protegidas da luz. As folhas tratadas com UV de cada espécie foram divididas em seis tubos falcon de 50 ml, sendo três destinados para a análise de conteúdo de resveratrol, contendo 1 g de folha cada, e três para análise da expressão do gene da *resveratrol sintase*, esses últimos contendo 300 mg de folhas. A mesma divisão foi feita para o grupo controle.

Todas as espécies foram induzidas simultaneamente e o experimento foi repetido três vezes (repetições biológicas). A cada nova indução, a disposição das folhas de cada espécie na bandeja foi diferente, certificando que o local na bandeja não influenciaria na intensidade da indução por UV e, consequentemente, nos resultados. Todas as amostras foram congeladas em nitrogênio líquido e depois guardadas em congelador -80℃ até o momento das análises laboratoriais.

4 Resultados

Os resultados são apresentados nesta seção. Para as variáveis produção de resveratrol (seção 4.1) e expressão de *resveratrol sintase* (seção 4.2), os ajustes serão feitos tanto pela metodologia Bayesiana tanto pela a Clássica, além de uma breve análise descritiva dos dados. Os dados coletados no experimento estão no anexo (7) e a programação no apêndice (6.1)

4.1 Produção de resveratrol

A seguir é apresentado a análise descritiva da produção de resveratrol. Percebe-se, pela tabela 4.1, que a espécie A. duranensis tem a maior média, com 371,97, e a maior mediana, com 394,20 micrograma por grama de folha. Porém, a espécie Anfidiplóide apresentou a menor media e mediana, com valores de 193,66 e 204,52, respectivamente.

Tabela 4.1: Resumo da variável resveratrol

Espécies	Média	Desvio Padrão	1^{o} Quartil	Mediana	3^{o} Quartil
A. hypogaea	241,99	70,20	218,96	225,34	244,07
A. duranensis	371,97	64,60	308,37	394,20	408,24
Anfidiplóide	193,66	28,95	177,60	204,52	217,53
A. ipaensis	226,04	12,60	223,00	228,78	230,72
Total	262,47	88,60	214,52	227,06	288,12

A Figura a seguir apresenta o boxplot da variável resveratrol por espécie.

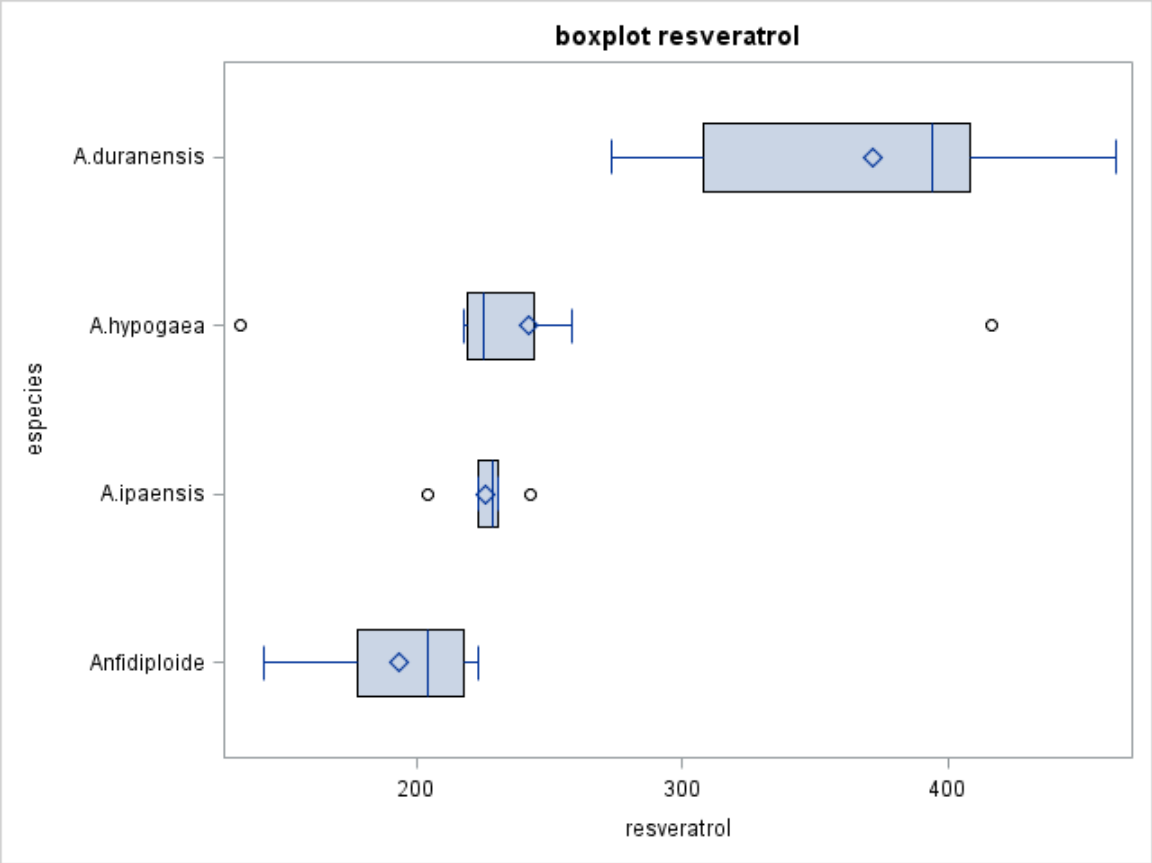

Figura 4.1: Boxplot por tratamento

Pode-se notar, pela figura 4.1, que a espécie A. Duranensis tem média e mediana bem maiores que as demais espécies, assim como o desvio interquartílico. A espécie A. Ipaensis ficou prejudicada na análise por ter 4 dados perdidos entre os 9 coletados.

4.1.1 Análise Clássica

Baseado na maneira que o experimento foi realizado, ver 3.3, em que as folhas foram coletadas com intervalo de vinte dias entre os experimentos, foi escolhido o modelo misto, com efeito aleatório para experimentos, para testar se existe variabilidade entre os experimentos. O modelo é apresentado a seguir:

$$y_{ijk} = \mu + \alpha_i + \gamma_j + e_{ijk} \tag{4.1}$$

Onde:

y_{ijk} é o valor observado do resveratrol

μ é uma constante inerente a todas as observações

α_i é o efeito do i-ésimo tratamento, considerado fixo

γ_j é o efeito do j-ésimo experimento, $\gamma_j \sim N(0, \sigma^2_{exp})$

e_{ijk} é o erro aleatório associado à abservação y_{ijk} e $e_{ijk} \sim N(0, \sigma^2)$

Os dados coletados da produção de resveratrol são apresentados na tabela 7.1. Foram utilizados 32 observações do total de 36, dessas, 4 valores não foram usados por apresentarem problemas na coleta. As quatro espécies são os tratamentos, cada um com 3 repetições biológicas e cada repetição com uma tréplica.

A figura a seguir mostra a análise residual do experimento.

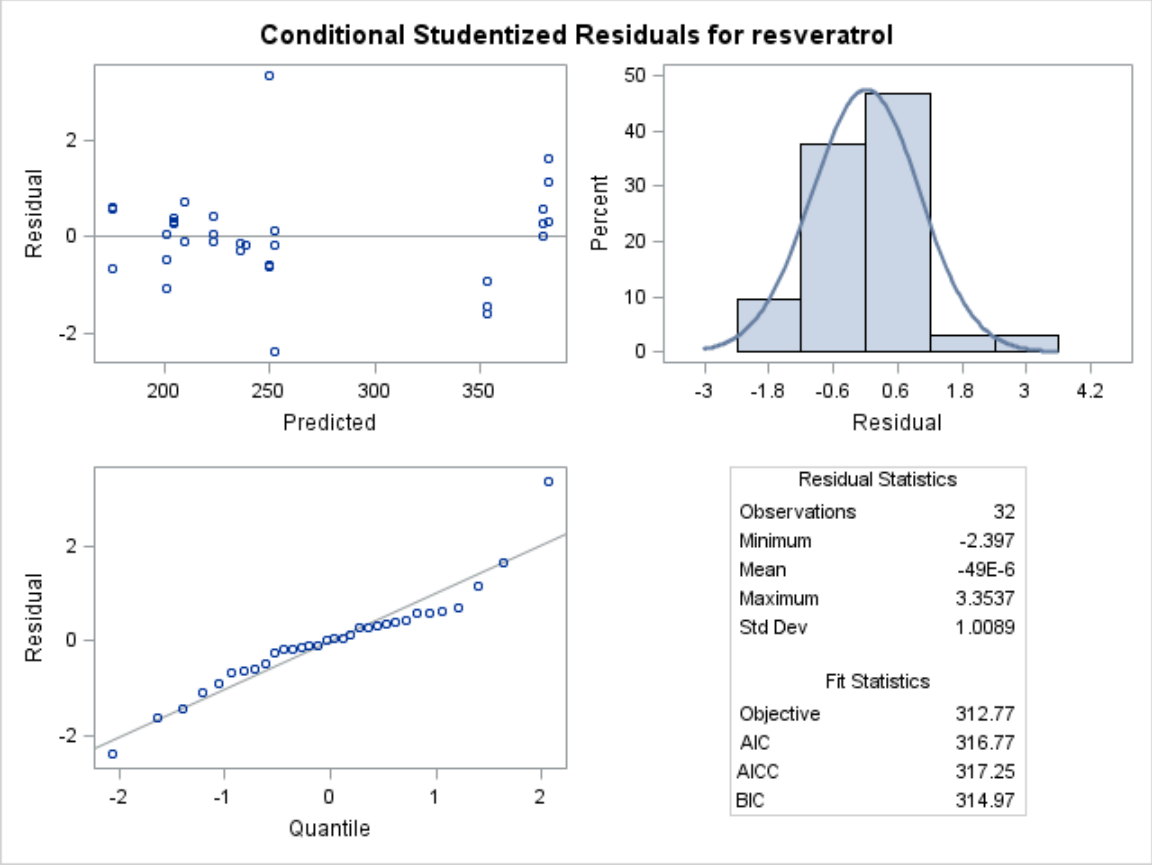

Figura 4.2: Resíduo Studentizado condicional

Observa-se, pela figura 4.2, que os pontos se mantêm próximos a reta no gráfico de normalidade, sugerindo que os resíduos seguem uma distribuição normal. Pode-se

notar também, que existem 2 outliers que ulltrapassam o intervalo $-2 \leq -r_i \leq 2$ no primeiro gráfico acima.

O teste-z de Wald foi utilizado para avaliar a significância dos efeitos aleatórios do modelo. O teste é baseado nos parâmetros de covariância e nas suas estimativas e nos erros padrões dessas estimativas, para maiores detalhes veja [West et al. (2007)]. Para o parâmetro σ^2 o teste de Wald resultou em valor de Z de $3,61$ e *valor p* igual a $0,0002$, o que indica que a hipótese nula, $\sigma^2 = 0$, pode ser rejeitada e σ^2 é significativo. Por outro lado, σ^2_{exp} resultou em Z de 0,61 e *valor p* de 0,2723, sendo assim, não existe evidências de que σ^2_{exp} seja significativo. Isso mostra que a variabilidade entre os experimentos é irrelevante.

Analisando a ANOVA (análise de variância), o teste F para os tratamentos teve valor de 18,39 e o *valor* $p < 0,001$. Existe evidências de que as espécies são diferentes em relação a produção de resveratrol, como já suspeitado na estatística descritiva dos dados.

Para a comparação múltipla das médias foi escolhido o teste de Tukey-Kramer, para maiores detalhes sobre o teste veja [Abdi and Williams(2010)]. O teste é apresentado na tabela 4.2.

Tabela 4.2: Diferença de médias das espécies - resveratrol com efeito aleatório

Espécies	Espécies	Estimativa	Erro padrão	G.l.	valor T	P-valor
A. duranensis	A. hypogaea	129,98	25,393	26	5,12	0,0001
A. duranensis	A. ipaensis	143,82	30,225	26,2	4,76	0,0003
A. duranensis	Anfidiploide	178,31	25,393	26	7,02	<0,0001
A. hypogaea	A. ipaensis	13,838	30,225	26,2	0,46	0,9675
A. hypogaea	Anfidiploide	48,332	25,393	26	1,90	0,2512
A. ipaensis	Anfidiploide	34,494	30,225	26,2	1,14	0,6680

Percebe-se, pelo tabela 4.2, que apenas a espécie A. duranensis se difere das outras, para um nível de significância de 5%.

As estimativas para os parâmetros são apresentados na tabela a seguir

Tabela 4.3: Estimativas dos parâmetros - reveratrol efeito aleatório

Efeito	Espécies	Estimativa	Erro padrão	G.l.	valor T	P-valor
Espécies	A. duranensis	178,31	25,3930	26	7,02	<0,0001
Espécies	hypogaea	48,3318	25,3930	26	1,9	0,0681
Espécies	A. ipaensis	34,4938	30,2253	26,2	1,14	0,2614
Espécies	Anfidiploide	0	–	–	–	–
Intercepto	–	193,66	21,5154	7,47	9	<0,0001
Intercepto	Experimento	421,53	695,60	–	0,61	0,2723
Resíduo	–	2901,61	804,46	–	3,61	0,0002

Após o ajuste do modelo, é indicado refazer a análise sem o efeito aleatório dos blocos, visto que a hipótese nula, $\sigma^2_{exp} = 0$, não foi rejeitada para um nível de significância de 5%.

Modelo sem efeito aleatório

O modelo proposto para a análise sem efeito aleatório é dado por

$$y_{ijk} = \mu + \alpha_i + e_{ijk} \tag{4.2}$$

Onde:

y_{ijk} é o valor observado do resveratrol

μ é uma constante inerente a todas as observações

α_i é o efeito do i-ésimo tratamento, considerado fixo

e_{ijk} é o erro aleatório associado à abservação y_{ijk} e $e_{ijk} \sim N(0, \sigma^2)$

A figura a seguir mostra a análise residual do experimento.

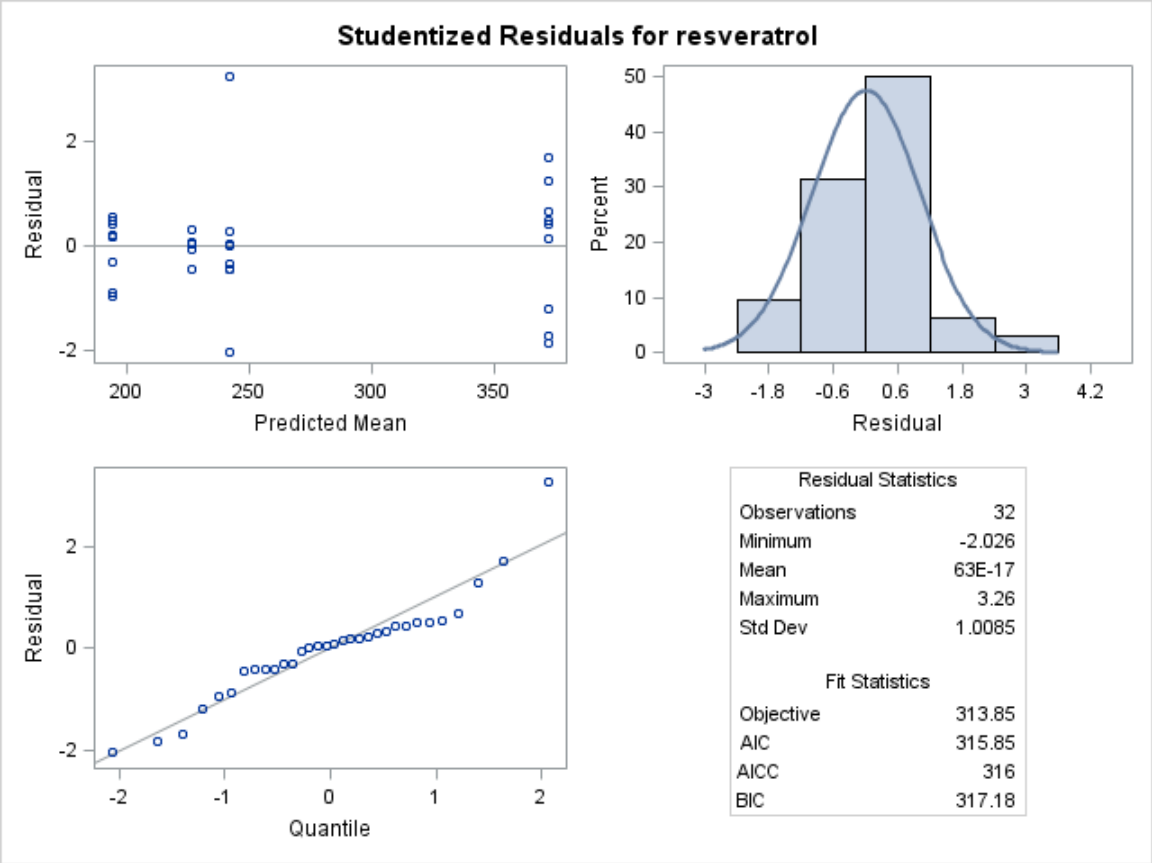

Figura 4.3: Resíduo Studentizado condicional

Percebe-se, pela figura 4.3, que assim como no modelo com efeito aleatório, os pontos se mantêm próximos da reta, o que sugere que os resíduos seguem, aproximadamente, uma distribuição normal e que existe um valor que ultrapassa o intervalo $-2 \leq -r_i \leq 2$ no último gráfico da figura.

Tabela 4.4: Critérios de seleção de modelos - resveratrol sem efeito aleatório

Modelo	AIC	BIC
Com efeito aleatório	316,77	314,97
Sem efeito aleatório	315,85	317,18

Pelos critérios de seleção de modelos AIC (critério de informação de Akaike) e BIC (critério de informação Bayesiano), na tabela 4.4, pode-se notar que para AIC o modelo mais parcimonioso é o modelo sem efeito aleatório e para BIC é o modelo com efeito aleatório.

O teste-z de Wald para σ^2 resultou em valor de Z igual a $3,74$ e *valor p* <0,0001, o que indica que a hipótese nula, $\sigma^2 = 0$, pode ser rejeitada e σ^2 é significativo.

A ANOVA teve o valor F de 16,64 para os tratamentos, e o *valor p* de $< 0,001$. Rejeita-se a hipótese nula de que as espécies são iguais. Portanto, existem evidências de que pelo menos uma espécie é diferente de outra, como já suspeitado na estatística descritiva dos dados.

A tabela a seguir, apresenta o teste de Tukey-Kramer pareado.

Tabela 4.5: Comparação de médias das espécies

Espécies	Espécies	Estimativa	Erro padrão	G.l.	valor T	P-valor
A. duranensis	A. hypogaea	129,98	26,76	28	4,86	0,0002
A. duranensis	A. ipaensis	145,93	31,67	28	4,61	0,0004
A. duranensis	Anfidiploide	178,31	26,76	28	6,66	<0,0001
A. hypogaea	A. ipaensis	15,95	31,67	28	0,50	0,9575
A. hypogaea	Anfidiploide	48,33	26,76	28	1,81	0,2918
A. ipaensis	Anfidiploide	32,38	31,67	28	1,02	0,7378

Percebe-se, pelo teste de Tukey-Kramer, que apenas a espécie A. duranensis se difere das outras espécies para um nível de significância de 5%, assim como constatado no modelo com efeito aleatório.

As estimativas dos parâmetros são apresentados na tabela a seguir

Tabela 4.6: Estimativas dos parâmetros - resveratrol sem efeito aleatório

Efeito	Espécies	Estimativa	Erro padrão	G.l.	valor T	P-valor
Espécies	A. duranensis	178,31	26,7638	28	6,66	<0,0001
Espécies	hypogaea	48,3318	26,7638	28	1,81	0,0817
Espécies	A. ipaensis	32,3808	31,6674	28	1,02	0,3153
Espécies	Anfidiploide	0	–	–	–	–
Intercepto	–	193,66	18,9249	28	10,23	<0,0001
Resíduo	–	3223,36	861,48	–	3,74	<0,0001

4.1.2 Análise Bayesiana

Dados os resultados obtidos na metodologia Clássica, utilizou-se, primeiramente, o modelo com efeito aleatório para desenvolver o ajuste Bayesiano. O modelo normal foi ajustado por meio do algoritmo Metropolis-Hasting, com 1.000.000 de iterações, com *burn-in* de 1.000.000 e o *thin* de 25 unidades.

$$\begin{aligned} Y &\sim N(\mu, \sigma^2) \\ \mu &= \beta_0 + \beta_i x_i + \gamma_j \end{aligned} \tag{4.3}$$

Onde:

- x_i $i = 1, 2, 3$ as espécies: A. duranensis, A, hypogaea e A. ipaensis
- μ é o preditor linear
- β_i $i = 1, 2, 3$ e $\beta_i \sim N(0, 10^6)$
- σ^2_{exp} $\sim GI(0,01; 0,01)$
- γ_j é o efeito aleatório do j-ésimo experimento $j = 1, 2, 3$. $\gamma_j \sim N(0, \sigma^2_{exp})$
- σ^2 $\sim GI(0,001; 0,001)$

A convergência da cadeia foi monitorada para cada um dos parâmetros através dos métodos informais (2.4.5): gráfico de traço, função de autocorrelação (FAC) e gráfico da densidade *posteriori*. Além dos critérios propostos por Geweke 2.4.2, Raftery-Lewis 2.4.4 e Heidelberger-Welch 2.4.3.

A tabela 4.7 a seguir apresenta os resultados para os critérios de convergência das cadeias. Pelo critério de Geweke, como o *valor p* foi sempre maior que o nível de significância escolhido, de 0,05, então não existem evidências contra a convergência dos parâmetros. Pelo critério de Raftery-Lewis, os valores devem ser menores que cinco para garantir a convergência. Como na cadeia, σ^2_{exp} foi maior do que cinco, a cadeia não atingiu a convergência.

Tabela 4.7: Critério de Geweke (*valor p*) e Raftery-Lewis (fator de dependência - FD) - resveratrol com efeito aleatório

Parâmetro	Geweke (*valor p*)	Raftery-Lewis (FD)
β_0	0,6809	4,7952
β_1	0,2956	1,0360
β_2	0,9581	1,0192
β_3	0,1775	1,0598
σ^2	0,2792	1,0037
σ^2_{exp}	0,3311	15,7718

Para verificar a convergência dos parâmetros foram utilizados também o critério de Heidelberger-Welch e o teste de Half-Width, apresentados na tabela 4.8. Percebe-se que para o teste de estacionariedade de Heidelberger-Welch, todas as cadeias convergiram, entretanto, para o teste de Half-Width na cadeia, σ^2_{exp} não convergiu.

Tabela 4.8: Critério de Heidelberger-Welch e Half-Width - resveratrol com efeito aleatório

Parâmetro	Teste de estacionariedade			Teste de Half-width		
	Cramer-von	p	Resultado	Half-width	Média	Resultado
β_0	0,0524	0,8617	Passou	0,4836	193,9	Passou
β_1	0,2824	0,1516	Passou	0,3211	178	Passou
β_2	0,1561	0,3715	Passou	0,2520	48,1903	Passou
β_3	0,4423	0,0560	Passou	0,3219	33,2418	Passou
σ^2	0,0591	0,8194	Passou	8,6769	3335,2	Passou
σ^2_{exp}	0,1437	0,4093	Passou	139,3	988,8	Falhou

Os resultados das tabelas 4.7 e 4.8 podem ser confirmados pelos gráficos apresentados a seguir para cada parâmetro.

Percebe-se, pelos gráficos dos parâmetros, ver anexo: parâmetro β_0 figura 7.1; β_1 figura 7.2; β_2 figura 7.3; e β_3 figura 7.4, que os gráficos de traço permanecem constantes em torno de um valor fixo médio. O gráfico de autocorrelação decresce, o que indica estacionariedade da cadeia. E, ainda, é possível perceber que há simetria

na distribuição dos parâmetros, o que indica que a distribuição normal foi uma boa escolha como distribuição *a priori* para os parâmetros.

No que tange o parâmetro da variância residual, σ^2, figura 7.5, pode-se notar que o gráfico de traço permanece constante em torno de um valor fixo médio. O gráfico de autocorrelação decresce, o que indica estacionariedade da cadeia. E, ainda, é possível perceber que há assimetria na distribuição *a posteriori* do parâmetro, o que confirma que a distribuição gama inversa é adequada. Todavia, para o parâmetro σ^2_{exp}, figura 7.6, observa-se que o gráfico de traço fica aproximadamente em zero e a distribuição *a posteriori* fica distorcida para aproximadamente zero. Isso indica que, ou esse parâmetro não está bem ajustado, ou ele não é necessário.

A tabela a seguir apresenta a autocorrelação das distribuições *a posteriori* dos parâmetros. Constata-se que os parâmetros β_0, σ^2 e σ^2_{exp} demoram mais que os outros para convergir, entretanto, aparentemente, todos os parâmetros convergem.

Tabela 4.9: Histórico de autocorrelação conforme parâmetro - resveratrol com efeito aleatório

Parâmetro	Lag 1	Lag 5	Lag 10	Lag 50
β_0	0,3125	0,1163	0,0541	-0,0033
β_1	0,0694	0,0005	0,0029	0,0153
β_2	0,0325	-0,0031	0,0041	0,0006
β_3	0,0941	-0,0020	-0,0009	0,0031
σ^2	0,0267	0,0138	0,0107	0,0092
σ^2_{exp}	0,0267	0,0344	0,0185	-0,0007

É apresentado na tabela a seguir, o resumo dos parâmetro acerca das estimativas, desvios e quartis.

Tabela 4.10: Estimativas das cadeias *a posteriori* dos parâmetros - resveratrol com efeito aleatório

Parâmetro	Média	Desvio	1º Quartil	Mediana	3º Quartil
(β_0) Intercepto	193,9	24,1508	179,4	193,7	208,0
(β_1) A. duranensis	178,0	27,2102	160,0	178,1	196,0
(β_2) A. hypogaea	48,1903	27,3305	30,1383	48,1938	66,1832
(β_3) A. ipaensis	33,2418	32,2090	11,9111	33,1604	54,4842
(σ^2) Resíduo	3335,2	989,30	2640,7	3165,0	3844,6
(σ^2_{exp}) Experimento	988,8	9849,9	0,5048	20,6439	338,5

Os intervalos de credibilidade e de HPD (Highest Posterior Density) são apresentados na tabela 4.11

Tabela 4.11: Intervalos *a posteriori* dos parâmetros - resveratrol com efeito aleatório

Parâmetro	Intervalo de credibilidade		Intervalo HPD	
(β_0) Intercepto	147,6	241,4	147,4	241,2
(β_1) A. duranensis	124,2	231,7	122,7	230,2
(β_2) A. hypogaea	-5,7819	102,5	-5,7789	102,6
(β_3) A. ipaensis	-30,0731	96,6562	-30,2632	96,3308
(σ^2) Resíduo	1907,0	5720,7	1657,5	5267,9
(σ^2_{exp}) Experimento	0,0122	6346,2	0,00147	3233,8

Modelo sem efeito aleatório

Para melhor análise de convergência, foi testado o modelo sem efeito aleatório. O modelo normal foi ajustado por meio do algoritmo Metropolis-Hasting, com 1.000.000 de iterações, com *burn-in* de 1.000.000 e o *thin* de 25 unidades.

$$\begin{aligned} Y &\sim N(\mu, \sigma^2) \\ \mu &= \beta_0 + \beta_i x_i \end{aligned} \tag{4.4}$$

Assim como no modelo Bayesiano anterior, equação 4.3, só que agora sem o efeito aleatório γ_j, onde:

x_i $i = 1, 2, 3$ as espécies: A. duranensis, A, hypogaea e A. ipaensis

μ é o preditor linear

β_i $i = 1, 2, 3$ e $\beta_i \sim N(0, 10^6)$

σ^2 $\sim GI(0,001; 0,001)$

A convergência da cadeia foi monitorada para cada um dos parâmetros através dos métodos informais (2.4.5): gráfico de traço, função de autocorrelação (FAC) e gráfico da densidade *posteriori*. Além dos critérios propostos por Geweke 2.4.2, Raftery-Lewis 2.4.4 e Heidelberger-Welch 2.4.3.

A tabela 4.12 a seguir apresenta os resultados para os critérios de convergência das cadeias. Pelo critério de Geweke, como o *valor p* não foi menor que o nível de significância escolhido, de 0,05, então não existem evidências contra a convergência para os parâmetros. Pelo critério de Raftery-Lewis, como o fator de dependência não foi maior que 5 para nenhum parâmetro, conclui-se que a cadeia atingiu a convergência.

Tabela 4.12: Critério de Geweke (*valor p*) e Raftery-Lewis (fator de dependência - FD) - resveratrol sem efeito aleatório

Parâmetro	Geweke (*valor p*)	Raftery-Lewis (FD)
β_0	0,8048	1,0235
β_1	0,2877	1,0339
β_2	0,3504	1,0235
β_3	0,0686	1,0403
σ^2	0,1049	1,0149

Para verificar a convergência dos parâmetros foram utilizados também o critério de Heidelberger-Welch e teste de Half-Width, apresentado na tabela 4.13. Esses resultados confirmam que as cadeias de todos os parâmetros são estacionárias e que a convergência foi alcançada.

Tabela 4.13: Critério de Heidelberger-Welch e Half-Width - resveratrol sem efeito aleatório

Parâmetro	Teste de estacionariedade			Teste de Half-width		
	Cramer-von	p	Resultado	Half-width	Média	Resultado
β_0	0,0784	0,7014	Passou	0,1821	193,7	Passou
β_1	0,0707	0,7471	Passou	0,2524	178,3	Passou
β_2	0,1787	0,3128	Passou	0,2502	48,1597	Passou
β_3	0,1484	0,3945	Passou	0,4069	32,5662	Passou
σ^2	0,2366	0,2065	Passou	9,5588	3470,5	Passou

Os resultados apresentados nas tabelas 4.12 e 4.13 podem ser confirmados pelos gráficos apresentados a seguir para cada parâmetro.

Observa-se, pelos gráficos dos parâmetros, ver anexo; parâmetro β_0 figura 7.7; β_1 figura 7.8; β_2 figura 7.9; e β_3 figura 7.10, que os gráficos de traço permanecem constantes em torno de um valor fixo médio. O gráfico de autocorrelação decresce, o que indica estacionariedade da cadeia. E, ainda, é possível perceber que há simetria na distribuição do parâmetro, o que indica que a distribuição normal foi uma boa escolha *a priori* para os parâmetros.

No que refere-se ao parâmetro da variância residual, σ^2, figura 7.11, pode-se notar que o gráfico de traço permanece constante em torno de um valor fixo médio. O gráfico de autocorrelação decresce, o que indica estacionariedade da cadeia. E, ainda, é possível perceber que há assimetria na distribuição *a posteriori* do parâmetro, o que confirma que a distribuição gama inversa é adequada.

A tabela a seguir apresenta a autocorrelação das distribuições *a posteriori* dos parâmetros. Constata-se, que a partir do quinto lag todas as cadeias já estavam com valores menores que 0,01, o que indica convergência dos parâmetros, afirmando os gráficos de autocorrelação.

Tabela 4.14: Histórico de autocorrelação conforme parâmetro - resveratrol sem efeito aleatório

Parâmetro	Lag 1	Lag 5	Lag 10	Lag 50
β_0	0,0366	-0,0019	0,0071	0,0085
β_1	0,0305	0,0004	0,0052	0,0163
β_2	0,0595	0,0038	0,0012	0,0108
β_3	0,0178	-0,0017	-0,0028	0,0116
σ^2	0,0178	0,0030	-0,0001	-0,0063

São apresentados na tabela a seguir, o resumo dos parâmetros acerca das estimativas, desvios e quartis.

Tabela 4.15: Estimativas das cadeias *a posteriori* dos parâmetros - resveratrol sem efeito aleatório

Parâmetro	Média	Desvio	1º Quartil	Mediana	3º Quartil
(β_0) Intercepto	193,7	19,5375	180,7	193,7	206,5
(β_1) A. duranensis	178,3	27,8143	159,9	178,4	196,8
(β_2) A. hypogaea	48,1597	27,7927	29,7492	48,2941	66,4257
(β_3) A. ipaensis	32,5662	32,7001	11,1118	32,7088	54,1899
(σ^2) Resíduo	3470,5	1012,1	2761,1	3296	3977,8

Os intervalos de credibilidade e de HPD (Highest Posterior Density) são apresentados na tabela 4.16

Tabela 4.16: Intervalos *a posteriori* dos parâmetros - resveratrol sem efeito aleatório

Parâmetro	Intervalo de credibilidade		Intervalo HPD	
(β_0) Intercepto	155,6	232,5	155,2	232
(β_1) A. duranensis	123,1	233	122,7	232,4
(β_2) A. hypogaea	-6,3257	102,9	-7,5922	101,5
(β_3) A. ipaensis	-32,1375	96,9988	-33,7706	95,0967
(σ^2) Resíduo	2022,5	5948,8	1831	5483,5

4.2 Gene *resveratrol sintase*

A seguir é apresentada a análise descritiva da produção da expressão do gene *resveratrol sintase*. Percebe-se, pela tabela 4.17, que a espécie A. ipaensis tem as maiores média e medianas, com valores de 58,6 e 58,95, nesta ordem. Porém, a espécie A. hypogaea apresentou as menores média e mediana quantificadas em 8,3 e 7,65, respectivamente.

Tabela 4.17: Resumo da variável expressão de *resveratrol sintase*

Espécies	Média	Desvio Padrão	1º Quartil	Mediana	3º Quartil
A. hypogaea	8,30	2,56	7,21	7,68	7,84
A. duranensis	29,86	0,84	29,60	30,20	30,43
Anfidiplóide	14,48	0,81	13,87	14,87	15,00
A. ipaensis	58,60	0,99	58,75	58,95	59,11
Total	28,14	19,92	13,10	21,96	50,21

A Figura a seguir apresenta o boxplot da variável gene por espécie.

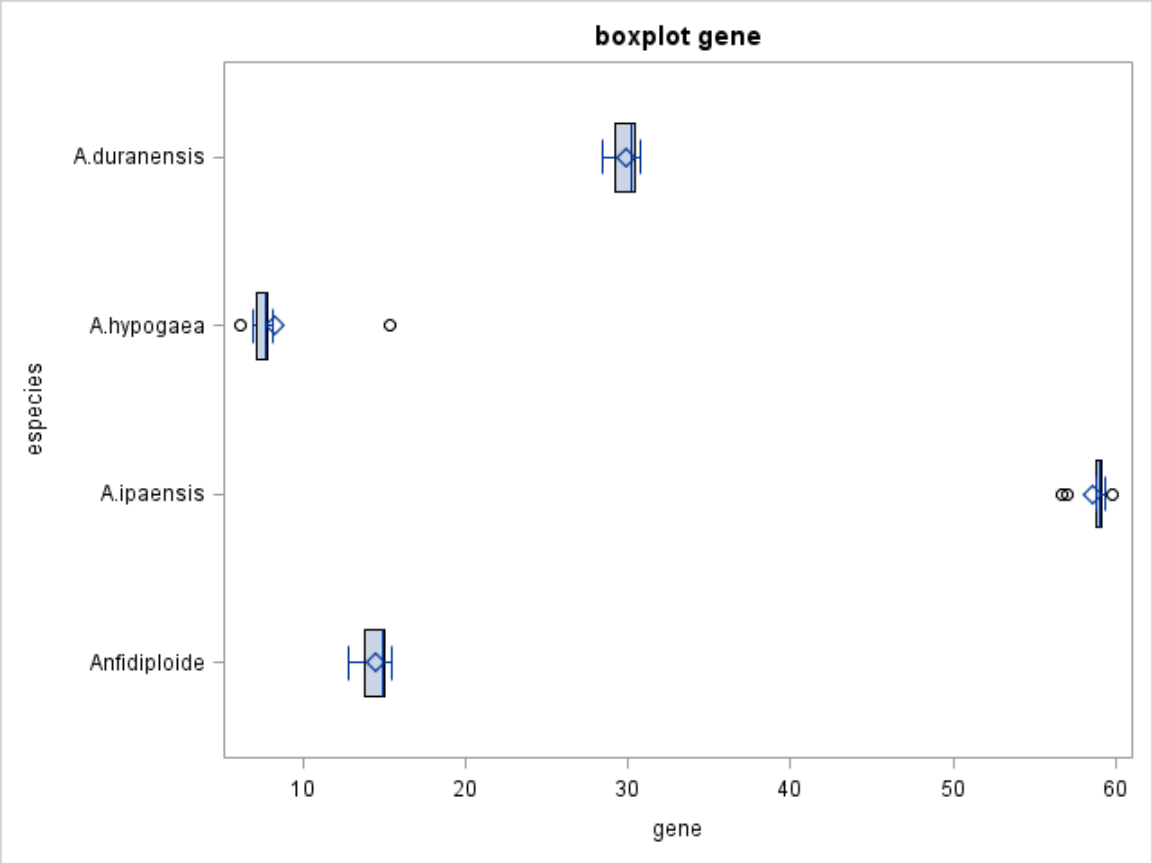

Figura 4.4: Boxplot por tratamento

Pode-se notar, pela figura 4.4, a baixa variabilidade intra espécies. A espécie A. ipaensis tem média e mediana bem maiores que as demais espécies.

4.2.1 Análise Clássica

Baseado na maneira que o experimento foi realizado, ver seção 3.3, em que as folhas foram coletadas com intervalo de vinte dias entre os experimentos, foi escolhido o modelo misto, com efeito aleatório para experimentos para testar se existe variabilidade entre os experimentos. O modelo é apresentado a seguir:

$$y_{ijk} = \mu + \alpha_i + \gamma_j + e_{ijk} \tag{4.5}$$

Onde:

y_{ijk} é o valor observado do gene *resveratrol sintase*

μ é uma constante inerente a todas as observações

α_i é o efeito do i-ésimo tratamento, considerado fixo

γ_j é o efeito do j-ésimo experimento, $\gamma_j \sim N(0, \sigma^2_{exp})$

e_{ijk} é o erro aleatório associado à observação y_{ijk} e $e_{ijk} \sim N(0, \sigma^2)$

Os dados coletados do gene *resveratrol sintase* são apresentados na tabela 7.3. Foram utilizados 34 observações do total de 36, dessas, 2 valores não foram usados por apresentarem problemas na coleta. Os tratamentos são as espécies, cada uma com 3 repetições biológicas e cada repetição com uma tréplica.

A figura a seguir mostra a análise residual do experimento.

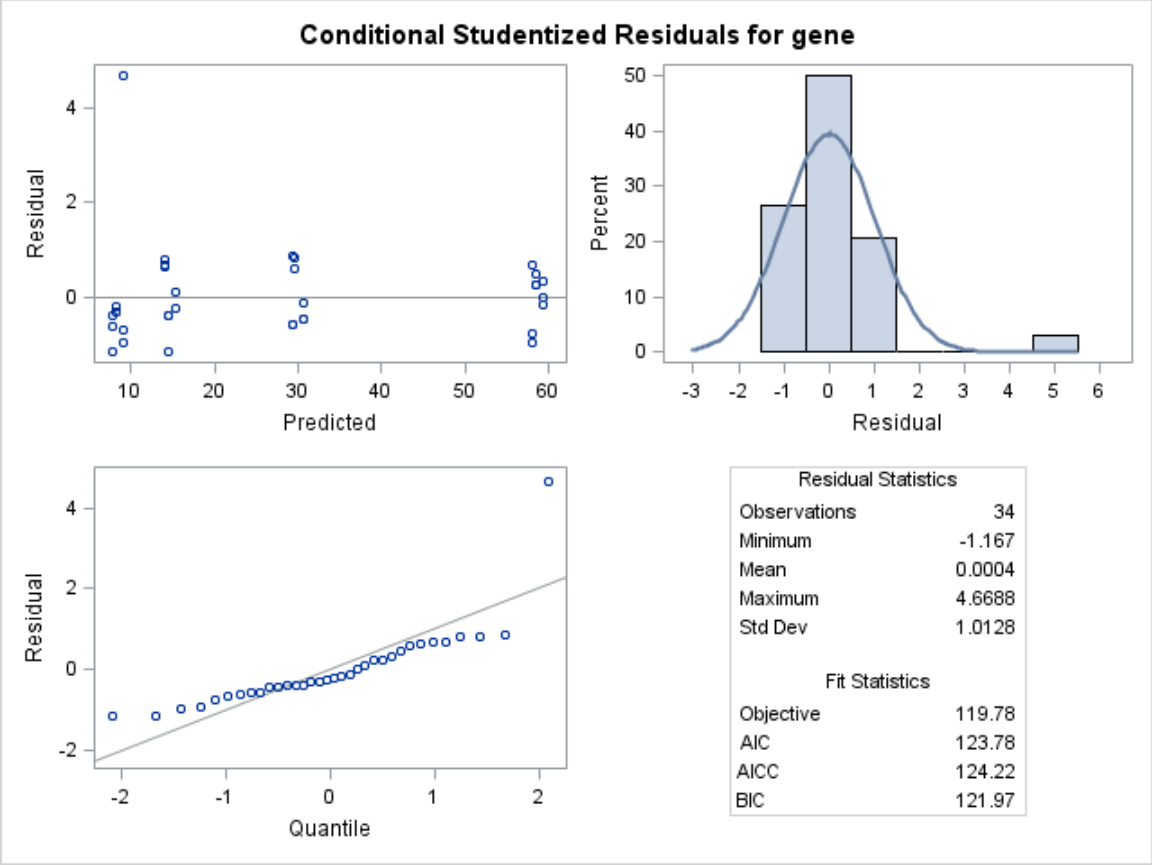

Figura 4.5: Resíduo Studentizado condicional

Observa-se, pela figura 4.5, que os pontos se mantêm próximos a reta no gráfico de normalidade, sugerindo que os resíduos seguem uma distribuição normal. Pode-se

notar também, que existe um outlier que ulltrapassam o intervalo $-2 \leq -r_i \leq 2$ no primeiro gráfico da figura.

O teste-z de Wald para σ^2 resultou em valor de Z de $3,74$ e *valor p* <0,0001, o que indica que a hipótese nula $\sigma^2 = 0$, pode ser rejeitada e σ^2 é significativo. Todavia, σ^2_{exp} resultou em Z de 0,76 e *valor p* de 0,2245, sendo assim, não existe evidências de que σ^2_{exp} seja significativo. Isso mostra que a variabilidade entre os experimentos é irrelevante.

Analisando a ANOVA (análise de variância), o teste F para os tratamentos teve valor de 2054,15 e *valor* $p < 0,001$. Portanto, existem evidências de que as espécies são diferentes em relação a expressão de gene, como já suspeitado na estatística descritiva dos dados.

A tabela a seguir, apresenta o teste de Tukey-Kramer pareado.

Tabela 4.18: Diferença de médias das espécies - gene com efeito aleatório

Espécies	Espécies	Estimativa	Erro padrão	G.l.	valor T	P-valor
A. duranensis	A. hypogaea	21,5383	0,7177	28,1	30,01	<0,0001
A. duranensis	A. ipaensis	-28,7617	0,7177	28,1	-40,08	<0,0001
A. duranensis	Anfidiploide	15,2728	0,7420	28,2	20,58	<0,0001
A. hypogaea	A. ipaensis	-50,3000	0,6945	28	-72,43	<0,0001
A. hypogaea	Anfidiploide	-6,2655	0,7177	28,1	-8,73	<0,0001
A. ipaensis	Anfidiploide	44,0345	0,7177	28,1	61,36	<0,0001

Percebe-se, pelo tabela 4.18, que todas as espécies se diferem uma das outras, para um nível de significância de 5%.

As estimativas para os parâmetros são apresentadas na tabela 4.26

Tabela 4.19: Estimativa dos parâmetros - gene com efeito aleatório

Efeito	Espécies	Estimativa	Erro padrão	G.l.	valor T	P-valor
Espécies	A. duranensis	15,2728	0,7420	28,2	20,58	<0,0001
Espécies	hypogaea	-6,2655	0,7177	28,1	-8,73	<0,0001
Espécies	A. ipaensis	44,0345	0,7177	28,1	61,36	<0,0001
Espécies	Anfidiploide	0	–	–	–	–
Intercepto	–	14,5694	21,5154	6,07	21,12	<0,0001
Intercepto	Experimento	0,6062	0,8005	–	0,76	0,2245
Resíduo	–	2,1703	0,5800	–	3,74	<0,0001

Após o ajuste do modelo, é indicado refazer a análise sem o efeito aleatório dos blocos, visto que a hipótese nula, $\sigma^2_{exp} = 0$, não foi rejeitada para um nível de significância de 5%.

Modelo sem efeito aleatório

O modelo proposto para a análise sem efeito aleatório é dado por

$$y_{ijk} = \mu + \alpha_i + e_{ijk} \tag{4.6}$$

Onde:

y_{ijk} é o valor observado do gene *resveratrol sintase*

μ é uma constante inerente a todas as observações

α_i é o efeito do i-ésimo tratamento, considerado fixo

e_{ijk} é o erro aleatório associado à abservação y_{ijk} e $e_{ijk} \sim N(0, \sigma^2)$

A figura a seguir mostra a análise residual do experimento.

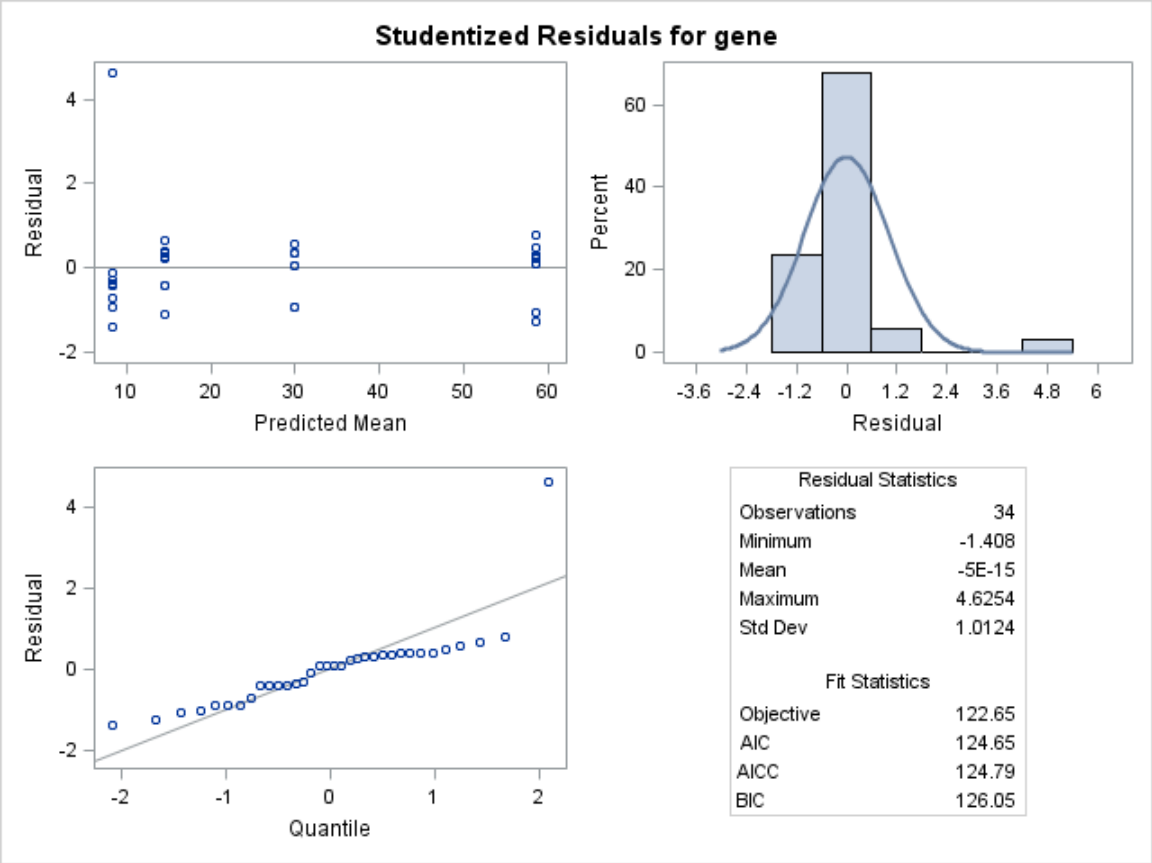

Figura 4.6: Resíduo Studentizado condicional

Percebe-se, pela figura 4.6, que assim como no modelo com efeito aleatório, os pontos se mantêm próximos da reta, o que sugere que os resíduos seguem, aproximadamente, uma distribuição normal e que existe um valor que ultrapassa o intervalo $-2 \leq -r_i \leq 2$ no último gráfico da figura.

Tabela 4.20: Critérios de seleção de modelos - gene *resveratrol sintase* sem efeito aleatório

Modelo	AIC	BIC
Com efeito aleatório	123,78	121,97
Sem efeito aleatório	124,65	126,05

Pelos critérios de seleção de modelos AIC (critério de informação de Akaike) e BIC (critério de informação Bayesiano), tabela 4.20, pode-se notar que para AIC o modelo mais parcimonioso é o modelo com efeito aleatório e para BIC é o modelo sem efeito aleatório.

O teste-z de Wald, para σ^2 resultou em valor de Z igual a $3,87$ e *valor p* <0,0001, o que indica que a hipótese nula, $\sigma^2 = 0$, pode ser rejeitada e σ^2 é significativo.

Na ANOVA o valor do teste F obtido foi de 1702,70 para os tratamentos, e o *valor p* de <0,0001. Sendo assim, rejeita-se a hipótese nula de que as espécies são iguais. Existem evidências de que pelo menos uma espécie é diferente de outra, como já suspeitado na estatística descritiva dos dados.

A tabela a seguir, apresenta o teste de Tukey-Kramer pareado.

Tabela 4.21: Comparação de médias das espécies

Espécies	Espécies	Estimativa	Erro padrão	G.l.	valor T	P-valor
A. duranensis	A. hypogaea	21,56	0,7873	30	27,38	<0,0001
A. duranensis	A. ipaensis	-28,74	0,7873	30	-36,5	<0,0001
A. duranensis	Anfidiploide	15,39	0,8102	30	18,99	<0,0001
A. hypogaea	A. ipaensis	-50,3	0,7638	30	-65,85	<0,0001
A. hypogaea	Anfidiploide	-6,17	0,7873	30	-7,84	<0,0001
A. ipaensis	Anfidiploide	44,13	0,7873	30	56,05	<0,0001

Percebe-se, pelo teste de Tukey-Kramer, que todas as espécies se diferem uma das outras, considerando um nível de significância de 5%, assim como constatado no modelo com efeito aleatório.

As estimativas dos parâmetros são apresentados na tabela 4.22 a seguir

Tabela 4.22: Estimativa dos parâmetros - gene *resveratrol sintase* sem efeito aleatório

Efeito	Espécies	Estimativa	Erro padrão	G.l.	valor T	P-valor
Espécies	A. duranensis	15,389	0,8102	30	18,99	<0,0001
Espécies	hypogaea	-6,1718	0,7873	30	-7,84	<0,0001
Espécies	A. ipaensis	44,1282	0,7873	30	56,05	<0,0001
Espécies	Anfidiploide	0	–	–	–	–
Intercepto	–	14,4757	0,5729	30	25,27	<0,0001
Resíduo	–	2,6255	0,6779	–	3,87	<0,0001

4.2.2 Análise Bayesiana

Dados os resultados obtidos na metodologia Clássica, utilizou-se, primeiramente, o modelo com efeito aleatório para desenvolver o ajuste Bayesiano. O modelo normal foi ajustado por meio do algoritmo Metropolis-Hasting, com 1.000.000 de iterações, com *burn-in* de 1.000.000 e o *thin* de 25 unidades.

$$\begin{aligned} Y &\sim N(\mu, \sigma^2) \\ \mu &= \beta_0 + \beta_i x_i + \gamma_j \end{aligned} \tag{4.7}$$

Onde:

x_i	$i = 1, 2, 3$ as espécies: A. duranensis, A, hypogaea e A. ipaensis
μ	é o preditor linear
β_i	$i = 1, 2, 3$ e $\beta_i \sim N(0, 10^6)$
σ^2_{exp}	$\sim GI(0,01; 0,01)$
γ_j	é o efeito aleatório do j-ésimo experimento $j = 1, 2, 3$. $\gamma_j \sim N(0, \sigma^2_{exp})$
σ^2	$\sim GI(0,001; 0,001)$

A convergência da cadeia foi monitorada para cada um dos parâmetros através dos métodos informais (2.4.5): gráfico de traço, função de autocorrelação (FAC) e gráfico da densidade *posteriori*. Além dos critérios propostos por Geweke 2.4.2, Raftery-Lewis 2.4.4 e Heidelberger-Welch 2.4.3.

A seguir, na tabela 4.23, apresenta os resultados para os critérios de convergência das cadeias. Pelo critério de Geweke, como o *valor p* foi sempre maior que o nível de significância escolhido, de 0,05, então não existem evidências contra a convergência dos parâmetros. Pelo critério de Raftery-Lewis, os valores devem ser menores que cinco para garantir a convergência. Como a cadeia β_0 foi maior do que cinco, a cadeia não atingiu a convergência.

Tabela 4.23: Critério de Geweke (*valor p*) e Raftery-Lewis (fator de dependência - FD) - gene *resveratrol sintase* com efeito aleatório

Parâmetro	Geweke (*valor p*)	Raftery-Lewis (FD)
β_0	0,6772	7,5809
β_1	0,8948	1,1343
β_2	0,7587	1,0619
β_3	0,1675	1,0774
σ^2	0,8880	1,0149
σ^2_{exp}	0,4665	1,1556

Para verificar a convergência dos parâmetros foram utilizados também o critério de Heidelberger-Welch e o teste de Half-Width, apresentados na tabela 4.24. Percebe-se que para o teste de estacionariedade de Heidelberger-Welch todas as cadeias convergiram, entretanto, para o teste de Half-Width a cadeia σ^2_{exp} não convergiu.

Tabela 4.24: Critério de Heidelberger-Welch e Half-Width - gene *resveratrol sintase* com efeito aleatório

Parâmetro	Teste de estacionariedade			Teste de Half-width		
	Cramer-von	p	Resultado	Half-width	Média	Resultado
β_0	0,2038	0,2602	Passou	0,0745	14,643	Passou
β_1	0,1293	0,4591	Passou	0,0099	15,285	Passou
β_2	0,0430	0,9173	Passou	0,0085	-6,257	Passou
β_3	0,4052	0,0700	Passou	0,0083	44,046	Passou
σ^2	0,1300	0,4566	Passou	0,0069	2,402	Passou
σ^2_{exp}	0,2260	0,2223	Passou	1,5409	3,827	Falhou

Os resultados das tabelas 4.23 e 4.24 podem ser confirmados pelos gráficos apresentados a seguir para cada parâmetro.

Percebe-se, pelas gráficos dos parâmetros, ver anexo: parâmetro β_0 figura 7.12; β_1 figura 7.13; β_2 figura 7.14; e β_3 figura 7.15, que os gráficos de traço permanecem constantes em torno de um valor fixo médio. O gráfico de autocorrelação decresce, o que indica estacionariedade da cadeia, exceto para β_0 que não convergiu após 50 lags,

o que indica falta de estacionariedade da cadeia. E, ainda, é possível perceber que há simetria na distribuição do parâmetro, o que indica que a distribuição normal foi uma boa escolha *a priori* para os parâmetros.

Se tratando do parâmetro da variância residual, σ^2, figura 7.16, pode-se notar que o gráfico de traço permanece constante em torno de um valor fixo médio O gráfico de autocorrelação decresce, o que indica estacionariedade da cadeia. E, ainda, é possível perceber que há assimetria na distribuição *a posteriori* do parâmetro, o que confirma que a distribuição gama inversa é adequada. Todavia, para o parâmetro σ^2_{exp}, presentes figura 7.17, observa-se que o gráfico de traço fica aproximadamente em zero e a distribuição *a posteriori* fica distorcida para aproximadamente zero. Isso indica que, ou esse parâmetro não está bem ajustado, ou ele não é necessário.

A tabela a seguir apresenta a autocorrelação das distribuições *a posteriori* dos parâmetros. Constata-se que os parâmetros β_0 e σ^2_{exp} não convergiram após 50 lags.

Tabela 4.25: Histórico de autocorrelação conforme parâmetro - gene *resveratrol sintase* com efeito aleatório

Parâmetro	Lag 1	Lag 5	Lag 10	Lag 50
β_0	0,7148	0,4802	0,3743	0,1075
β_1	0,1674	0,0031	0,0014	-0,0006
β_2	0,1120	0,0025	-0,0019	0,0083
β_3	0,1148	0,0003	-0,0039	0,0043
σ^2	0,0255	0,0021	0,0108	-0,0093
σ^2_{exp}	0,1719	0,1732	0,0959	0,0545

É apresentado na tabela a seguir, o resumo dos parâmetros acerca das estimativas, desvios e quartis.

Tabela 4.26: Estimativas das cadeias *a posteriori* dos parâmetros - gene *resveratrol sintase* com efeito aleatório

Parâmetro	Média	Desvio	1º Quartil	Mediana	3º Quartil
(β_0) Intercepto	14,643	1,1400	14,0607	14,5716	15,0902
(β_1) A. duranensis	14,285	0,7880	14,7706	15,2810	15,8025
(β_2) A. hypogaea	-6,257	0,7629	-6,7550	-6,2624	-5,8025
(β_3) A. ipaensis	44,046	0,7576	43,5470	44,0528	44,5423
(σ^2) Resíduo	2,402	0,6994	1,9067	2,2879	2,7641
(σ^2_{exp}) Experimento	3,827	37,2330	0,1817	0,5611	1,6121

Os intervalos de credibilidade e de HPD (Highest Posterior Density) são dado na tabela a seguir

Tabela 4.27: Intervalos *a posteriori* dos parâmetros - gene *resveratrol sintase* com efeito aleatório

Parâmetro	Intervalo de credibilidade		Intervalo HPD	
(β_0) Intercepto	12,8119	16,8209	12,6996	16,6153
(β_1) A. duranensis	13,7330	16,8404	13,7007	16,8019
(β_2) A. hypogaea	-7,7539	-4,7336	-7,7634	-4,7484
(β_3) A. ipaensis	42,5352	45,5327	42,5553	45,5466
(σ^2) Resíduo	1,3911	4,0843	1,2350	3,7827
(σ^2_{exp}) Experimento	0,0145	20,3920	0,00123	9,9627

Modelo sem efeito aleatório

Para aprimoramento da análise de convergência, foi testado o modelo sem efeito aleatório. O modelo normal foi ajustado por meio do algoritmo Metropolis-Hasting, com 1.000.000 de iterações, com *burn-in* de 1.000.000 e o *thin* de 25 unidades.

$$\begin{aligned} Y &\sim N(\mu, \sigma^2) \\ \mu &= \beta_0 + \beta_i x_i \end{aligned} \tag{4.8}$$

Assim como no modelo Bayesiano anterior, equação 4.7, só que agora sem o efeito aleatório γ_j, onde:

x_i $i = 1, 2, 3$ as espécies: A. duranensis, A, hypogaea e A. ipaensis

μ é o preditor linear

β_i $i = 1, 2, 3$ e $\beta_i \sim N(0, 10^6)$

σ^2 $\sim GI(0,001; 0,001)$

A convergência da cadeia foi monitorada para cada um dos parâmetros através dos métodos informais (2.4.5): gráfico de traço, função de autocorrelação (FAC), gráfico da densidade *posteriori*. Além dos critérios propostos por Geweke 2.4.2, Raftery-Lewis 2.4.4 e Heidelberger-Welch 2.4.3.

A seguir, na tabela 4.28, apresenta os resultados para os critérios de convergência das cadeias. Pelo critério de Geweke, destaca-se o valor de σ^2 que ficou menor do que 0,05, o que indica atenção na convergência do parâmetro. Pelo critério de Raftery-Lewis, como o fator de dependência não foi maior que 5 para nenhum parâmetro, conclui-se que a cadeia atingiu a convergência.

Tabela 4.28: Critério de Geweke (*valor p*) e Raftery-Lewis (fator de dependência - FD) - gene *resveratrol sintase* sem efeito aleatório

Parâmetro	Geweke (*valor p*)	Raftery-Lewis (FD)
β_0	0,6570	1,0598
β_1	0,8684	1,0192
β_2	0,6795	1,0574
β_3	0,8669	1,0128
σ^2	0,0165	1,0275

Para verificar a convergência dos parâmetros foram utilizados também o critério de Heidelberger-Welch e teste de Half-Width, apresentado na tabela 4.29. Esses resultados confirmam que as cadeias de todos os parâmetros são estacionárias e que a convergência foi alcançada.

Tabela 4.29: Critério de Heidelberger-Welch e Half-Width - gene *resveratrol sintase* sem efeito aleatório

Parâmetro	Teste de estacionariedade			Teste de Half-width		
	Cramer-von	p	Resultado	Half-width	Média	Resultado
β_0	0,1425	0,4132	Passou	0,00542	14,4763	Passou
β_1	0,1567	0,3697	Passou	0,00706	15,3906	Passou
β_2	0,0620	0,8016	Passou	0,00797	-6,1705	Passou
β_3	0,2238	0,2257	Passou	0,00746	44,1229	Passou
σ^2	0,4335	0,0590	Passou	0,00901	3470,5	Passou

Os resultados das tabelas 4.28 e 4.29 podem ser confirmados pelos gráficos apresentados a seguir para cada parâmetro.

Percebe-se, pelas gráficos dos parâmetros, ver anexo: parâmetro β_0 figura 7.18; β_1 figura 7.19; β_2 figura 7.20; e β_3 figura 7.21, que os gráficos de traço permanecem constantes em torno de um valor fixo médio. O gráfico de autocorrelação decresce, o que indica estacionariedade da cadeia. E, ainda, é possível perceber que há simetria na distribuição do parâmetro, o que indica que a distribuição normal foi uma boa escolha como distribuição *a priori* para os parâmetros.

No que tange o parâmetro da variância residual, σ^2, figura 7.22, pode-se notar que o gráfico de traço permanece constante em torno de um valor fixo médio. O gráfico de autocorrelação decresce, o que indica estacionariedade da cadeia. E, ainda, é possível perceber que há assimetria na distribuição *a posteriori* do parâmetro, o que confirma que a distribuição gama inversa é adequada.

A tabela a seguir apresenta a autocorrelação das distribuições *a posteriori* dos parâmetros. Constata-se que a partir do quinto lag todas as cadeias já estavam com valores de autocorrelação menores que 0,01, o que indica convergência dos parâmetros, afirmando os gráficos de autocorrelação.

Tabela 4.30: Histórico de autocorrelação conforme parâmetro - gene *resveratrol sintase* sem efeito aleatório

Parâmetro	Lag 1	Lag 5	Lag 10	Lag 50
β_0	0,0443	-0,0046	-0,0020	0,0006
β_1	0,0432	-0,0079	-0,0103	0,0008
β_2	0,0477	0,0073	-0,0098	0,0084
β_3	0,0301	-0,0039	0,0029	0,0012
σ^2	0,0152	0,0023	0,0080	0,0013

São apresentados na tabela a seguir, o resumo dos parâmetros acerca das estimativas, desvios e quartis.

Tabela 4.31: Estimativas das cadeias *a posteriori* dos parâmetros - gene *resveratrol sintase* sem efeito aleatório

Parâmetro	Média	Desvio	1º Quartil	Mediana	3º Quartil
(β_0) Intercepto	14,4763	0,5925	14,0851	14,4732	14,8675
(β_1) A. duranensis	15,3906	0,8359	14,8419	15,3895	15,9361
(β_2) A. hypogaea	-6,1705	0,8140	-6,7091	-6,1701	-5,6318
(β_3) A. ipaensis	44,1229	0,8145	43,5877	44,1220	44,6643
(σ^2) Resíduo	2,8126	0,7861	2,2610	2,6784	3,2204

Os intervalos de credibilidade e de HPD (Highest Posterior Density) são dado na tabela 4.32

Tabela 4.32: Intervalos *a posteriori* dos parâmetros - gene *resveratrol sintase* sem efeito aleatório

Parâmetro	Intervalo de credibilidade		Intervalo HPD	
(β_0) Intercepto	13,3187	15,6423	13,3083	15,6302
(β_1) A. duranensis	13,7340	17,0385	13,7074	17,0104
(β_2) A. hypogaea	-7,7808	-4,5661	-7,7689	-4,5582
(β_3) A. ipaensis	42,5261	45,7378	42,5376	45,7471
(σ^2) Resíduo	1,6796	4,7064	1,5438	4,3951

4.3 Discussão

Os modelos com efeito aleatório não serão comparados, visto que as cadeias dos parâmetros tiveram problemas para atingir a convergência. As comparações foram feitas apenas entre os modelos fixos e serão apresentados na subseções a seguir.

4.3.1 Produção de resveratrol

Na tabela 4.33, é possível verificar que as estimativas para os parâmetros das abordagens Clássicas e Bayesianas estão próximas, com exceção da variância residual, em que a abordagem Bayesiana apresentou um valor de 3470,5, contra 3223,36 alcançado pela abordagem Clássica. Possivelmente seja melhor o uso da mediana para estimar σ^2, valor de 3296, que pode ser verificado na tabela 4.15, sabendo-se que a distribuição gama inversa é assimétrica.

Tabela 4.33: Comparação dos parâmetros - resveratrol sem efeito aleatório

Parâmetro	Clássica		Bayesiana	
	Estimativa	Erro padrão	EAP	Desvio padrão
(β_0) Intercepto	193,66	18,9249	193,7	19,5375
(β_1) A. duranensis	178,31	26,7638	178,3	27,8143
(β_2) A. hypogaea	48,3318	26,7638	48,1597	27,7927
(β_3) A. ipaensis	32,3808	31,6674	32,5662	32,7001
(σ^2) Resíduo	3223,36	861,48	3470,5	1012,1

A diferença entre parâmetros está apenas nas casas decimais, o que ja era esperado, visto que as informações utilizadas na metodologia Bayesiana para as distribuições *a priori* derivam dos resultados obtidos na metodologia Clássica. Em relação o erro padrão, a abordagem Bayesiana apresentou valores maiores do que a abordagem Clássica para todos os parâmetros.

Os intervalos de confiança, credibilidade e HPD são dados na tabela a seguir

Tabela 4.34: Comparação do Intervalos de confiança, credibilidade e HPD dos parâmetros - resveratrol sem efeito aleatório

Parâmetro	Intervalos 95%					
	Confiança		Credibilidade		HPD	
(β_0) Intercepto	154,90	232,43	155,6	232,5	155,2	232
(β_1) A. duranensis	123,49	233,13	123,1	233	122,7	232,4
(β_2) A. hypogaea	-6,4914	103,15	-6,3257	102,9	-7,5922	101,5
(β_3) A. ipaensis	-32,4869	97,2485	-32,1375	96,9988	-33,7706	95,0967
(σ^2) Resíduo	2029,97	5895,93	2022,5	5948,8	1831	5483,5

Constata-se, pela tabela 4.34, que o intervalo de credibilidade ficou muito próximo do intervalo de confiança, e o intervalo HPD teve a menor amplitude. Uma das vantagens da abordagem Bayesiana é que os intervalos de credibilidade tendem a ter menores amplitudes que os de confiança, obtidos na inferência Clássica.

4.3.2 Gene *resveratrol sintase*

Na tabela 4.35, é possível verificar que as estimativas para os parâmetros das abordagens Clássicas e Bayesianas são praticamente iguais. Se utilizado a mediana no lugar da média para estimar a variância residual, o valor ficará ainda mais próximo da estimativa da abordagem Clássica, isto é, de 2,8126 (média) para 2,6784 (mediana), contra 2,6255 estimativa abordagem Clássica.

Tabela 4.35: Comparação dos parâmetros - gene *resveratrol sintase* sem efeito aleatório

Parâmetro	Clássica		Bayesiana	
	Estimativa	Erro padrão	EAP	Desvio padrão
(β_0) Intercepto	14,4757	0,5729	14,4763	0,5925
(β_1) A. duranensis	15,3890	0,8102	15,3906	0,8359
(β_2) A. hypogaea	-6,1718	0,7873	-6,1705	0,8140
(β_3) A. ipaensis	44,1282	0,7873	44,1229	0,8145
(σ^2) Resíduo	2,6255	0,6779	2,8126	0,7861

Em relação o erro padrão, a abordagem Bayesiana apresentou valores maiores do que a abordagem Clássica para todos os parâmetros, assim como para a produção de resveratrol, visto anteriormente.

Os intervalos de confiança, credibilidade e HPD são apresentados na tabela 4.36

Tabela 4.36: Intervalos *a posteriori* dos parâmetros - gene *resveratrol sintase* sem efeito aleatório

Parâmetro	Intervalos 95%					
	Confiança		Credibilidade		HPD	
(β_0) Intercepto	13,3057	15,6457	13,3187	15,6423	13,3083	15,6302
(β_1) A. duranensis	13,7344	17,0436	13,7340	17,0385	13,7074	17,0104
(β_2) A. hypogaea	-7,7789	-4,5638	-7,7808	-4,5661	-7,7689	-4,5582
(β_3) A. ipaensis	42,5202	45,7362	42,5261	45,7378	42,5376	45,7471
(σ^2) Resíduo	1,6766	4,6909	1,6796	4,7064	1,5438	4,3951

Constata-se, pela tabela 4.36, que o intervalo de credibilidade ficou muito próximo do intervalo de confiança e o intervalo HPD teve a menor amplitude.

5 Conclusão

O presente trabalho, teve como objetivos avaliar a variabilidade das espécies para a produção de resveratrol e a expressão de gene *resveratrol sintase* em função da produção de resveratrol após tratamento com UV sobe a abordagem Clássica e a Bayesiana e realizar uma comparação entre os métodos.

Para a produção de resveratrol e expressão de gene, a abordagem Clássica rejeitou o modelo com efeito aleatório para os experimentos; já na abordagem Bayesiana, foram encontrados problemas para a convergência do parâmetro σ^2_{exp}. O modelo sem efeito aleatório foi o mais parcimonioso e apresentou a melhor qualidade de ajuste segundo os métodos utilizados.

Para a produção de resveratrol, afim de testar a diferença das médias, foi utilizado o teste de Tukey-Kramer. Apenas a espécie A. duranensis se difere das outras, tendo a maior concentração de resveratrol que as demais. E para a expressão de gene *resveratrol sintase*, todas as espécies se diferem, sendo A.ipaensis a que apresenta a maior quantidade da expressão de gene, seguida da A. duranensis, a Anfidiploide sintético e por fim a A. hypogaea com a menor quantidade.

As metodologias Clássica e Bayesiana foram igualmente viáveis e eficientes para a análise de dados, os valores estimados em ambos os métodos foram muito próximos, assim como os intervalos de confiança, credibilidade e HPD.

Referências Bibliográficas

[Abdi and Williams(2010)] Hervé Abdi and Lynne J Williams. Tukey's honestly significant difference (hsd) test. *Encyclopedia of Research Design. Thousand Oaks, CA: Sage*, 2010.

[Barbosa(2005)] Vanessa Barbosa. Inferência bayesiana no estudo genético quantitativo de características de carcaça, utilizando a tecnica de ultra-sonografia e suas relações com crescimento, em novilhos da raça nelore. 2005.

[Boligon and Alburquerque(2010)] Arione Augusti Boligon and LG de Alburquerque. Correlações genéticas entre escores visuais e características reprodutivas em bovinos nelore usando inferência bayesiana. *Pesquisa Agropecuária Brasileira*, 45 (12):1412–1418, 2010.

[Box and Tiao(1992)] G. E. P . Box and G. C. Tiao. *Bayesian inference in statistical analysis*. John Wiley & Sons, Inc., 1992.

[Box and Muller(1958)] George EP Box and Mervin E Muller. A note on the generation of random normal deviates. *The Annals of Mathematical Statistics*, 29(2): 610–611, 1958.

[Bussab and Moretin(2004)] W. O. Bussab and P. A. Moretin. *Estatística Básica*. Saraiva, 2004.

[Carvalho(2013)] Paula Andréa s. de Vasconcelos Carvalho. Concentração de resveratrol e expressão de resveratrol sintase em espécies de arachis. Mestrado, Universidade estadual paulista - UNESP, 2013.

[Casella and George(1992)] George Casella and Edward I George. Explaining the gibbs sampler. *The American Statistician*, 46(3), 1992.

[de Aquino(2008)] Luiz Henrique de Aquino. Inferência bayesiana na análise genética de populações diplóides: estimação do coeficiente de endogamia e da taxa de fecundação cruzada. *Ciência Rural*, 38(5):1258–1265, 2008.

[Ehlers(2003)] Ricardo Ehlers. Análise de séries temporais, 2003. *Departamento de Estatística, Universidade Federal do Paraná*, 2003.

[Gamerman and Lopes(2006)] D. Gamerman and H. F. Lopes. *Markov chain Monte Carlo : stochastic simulation for Bayesian inference*. Taylor & Francis, 2006.

[Gelfand and Smith(1990)] Alan E Gelfand and Adrian FM Smith. Sampling-based approaches to calculating marginal densities. *Journal of the American statistical association*, 85(410):398–409, 1990.

[Gelman and Rubin(1992)] Andrew Gelman and Donald B Rubin. Inference from iterative simulation using multiple sequences. *Statistical science*, pages 457–472, 1992.

[Geman and Geman(1984)] Stuart Geman and Donald Geman. Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, (6):721–741, 1984.

[Geweke(1992)] J. Geweke. Evaluating the accuracy of sampling-based approaches to the calculation of posteriori moments. *Bayesian Statistics 4*, 1992.

[Hastings(1970)] W Keith Hastings. Monte carlo sampling methods using markov chains and their applications. *Biometrika*, 57(1):97–109, 1970.

[Heidelberger and Welch(1983)] Philip Heidelberger and Peter D Welch. Simulation run length control in the presence of an initial transient. *Operations Research*, 31(6):1109–1144, 1983.

[Institute(2011)] Sas Institute. *SAS/STAT 9.3 user's guide*. SAS Institute, 2011.

[Karlis and Ntzoufras(2000)] D Karlis and I Ntzoufras. On modelling soccer data. *Student*, 3:229–244, 2000.

[Metropolis and Ulam(1949)] Nicholas Metropolis and Stanislaw Ulam. The monte carlo method. *Journal of the American statistical association*, 44(247):335–341, 1949.

[Nelder and Wedderburn(1972)] John A Nelder and Robert WM Wedderburn. Generalized linear models. *Journal of the Royal Statistical Society. Series A (General)*, pages 370–384, 1972.

[Ntzoufras(2009)] I. Ntzoufras. *Bayesian Modeling Using WinBUGS*. Wiley, 2009.

[Raftery and Lewis(1992)] Adrian E Raftery and Steven M Lewis. [practical markov chain monte carlo]: Comment: One long run with diagnostics: Implementation strategies for markov chain monte carlo. *Statistical Science*, 7(4):493–497, 1992.

[West et al.(2007)West, Welch, and Galecki] BT West, KB Welch, and AT Galecki. Linear mixed model. *Chapman Hall/CRC*, 2007.

6 Apêndice

6.1 Programação

```
data dados;
    input especies $ 12. experimento amostra resveratrol gene;
    datalines;
    A.hypogaea 1 1 416.4927183 15.37
    A.hypogaea 1 1 218.9641406 7.762
    A.hypogaea 1 1 219.8188994 8.129
    A.hypogaea 2 2 243.39 7.841
    A.hypogaea 2 2 258.26 7.684
    A.hypogaea 2 2 133.55 7.684
    A.hypogaea 3 3 244.0690671 6.904
    A.hypogaea 3 3 218.070385 7.208
    A.hypogaea 3 3 225.3358736 6.153
    Anfidiploide 1 4 204.5181386 14.985
    Anfidiploide 1 4 177.6047521 .
    Anfidiploide 1 4 147.4253977 15.448
    Anfidiploide 2 5 220.4484853 13.8580375
    Anfidiploide 2 5 217.5322102 13.8676625
    Anfidiploide 2 5 223.3647604 12.8418
    Anfidiploide 3 6 204.0673562 14.91
    Anfidiploide 3 6 205.4638827 14.834
    Anfidiploide 3 6 142.5397435 15.061
    A.duranensis 1 7 380.1714765 29.973
    A.duranensis 1 7 394.2 29.973
    A.duranensis 1 7 408.2399117 30.432
    A.duranensis 2 8 463.3947775 .
    A.duranensis 2 8 439.789419 30.432
    A.duranensis 2 8 398.9778261 30.741
```

```
    A.duranensis 3 9 281.37 28.4711875
    A.duranensis 3 9 273.25 28.4634375
    A.duranensis 3 9 308.3669074 30.432
    A.ipaensis 1 10 222.9988167 59.793
    A.ipaensis 1 10 228.7758124 59.111
    A.ipaensis 1 10 . 59.346
    A.ipaensis 2 11 . 58.748
    A.ipaensis 2 11 . 58.748
    A.ipaensis 2 11 230.7163654 59.047
    A.ipaensis 3 12 243.1694113 58.948
    A.ipaensis 3 12 204.5573392 56.991
    A.ipaensis 3 12 . 56.703
;
ods html;
ODS GRAPHICS ON;
***RESVERATROL***;
***EFEITO ALEATÓRIO***;
proc print data=dados;
run;
proc univariate data=dados;
    var resveratrol;
    class especies;
run;
proc sgplot data=dados;
    title "boxplot resveratrol";
    hbox resveratrol / category=especies;
run;
TITLE1 H=1.2 'MIXED COM EFEITO ALEATORIO - RESVERATROL';
proc mixed data=dados method=REML plot=all covtest cl;
    class especies experimento amostra;
    model resveratrol = especies / DDFM=KR solution cl;
    random intercept / subject=experimento type=VC;
    lsmeans especies / adjust=tukey;
run;
```

```
options cpucount=4 threads;
TITLE1 H=1.2 'MCMC COM EFEITO ALEATORIO - RESVERATROL';
PROC MCMC DATA=dados outpost=cassout nbi=1000000 nmc=100000 diag=all
    thin=25 seed=6534 mchistory=detailed;
    parms beta0-beta3 0 s2 1 s2g 1;
    prior beta:    normal(0,var=1e6);
    prior s2   igama(shape=0.001,scale=0.001);
    prior s2g   igama(shape=0.01,scale=0.01);
    random gama   normal(0, var=s2g) subject=amostra;
    mu= beta0 + beta1*(especies='A.duranensis')
    + beta2*(especies='A.hypogaea')
    + beta3*(especies='A.ipaensis')
    + gama;
    model resveratrol   normal(mu, var=s2);
run;
***EFEITO FIXO***;
TITLE1 H=1.2 'MIXED COM EFEITO FIXO - RESVERATROL';
proc mixed data=dados method=REML plot=all covtest cl;
    class especies experimento amostra;
    model resveratrol = especies / DDFM=KR solution cl;
    lsmeans especies / adjust=tukey;
run;
TITLE1 H=1.2 'MCMC COM EFEITO FIXO - RESVERATROL';
PROC MCMC DATA=dados outpost=cassout nbi=1000000 nmc=1000000 diag=all
    thin=25 seed=578 mchistory=detailed;
    parms beta0-beta3 0 sigma2 1;
    prior beta:    normal(0,var=1e6);
    prior sigma2   igama(shape=0.001,scale=0.001);
    mu= beta0 + beta1*(especies='A.duranensis')
    + beta2*(especies='A.hypogaea')
    + beta3*(especies='A.ipaensis');
    model resveratrol   normal(mu, var=sigma2);
run;
***GENE***;
```

```
***EFEITO ALEATÓRIO***;
proc print data=dados;
run;
proc univariate data=dados;
    var resveratrol;
    class especies;
run;
proc sgplot data=dados;
    title "boxplot gene";
    hbox gene / category=especies;
run;
TITLE1 H=1.2 'MIXED COM EFEITO ALEATORIO - GENE';
proc mixed data=dados method=REML plot=all covtest cl;
    title 'Modelo para gene';
    title2 'gene micrograma por grama de folha';
    class especies experimento amostra;
    model gene = especies / DDFM=KR solution cl;
    random intercept / subject=experimento type=VC;
    lsmeans especies / adjust=tukey;
run;
TITLE1 H=1.2 'MCMC COM EFEITO ALEATORIO - GENE';
PROC MCMC DATA=dados outpost=cassout nbi=1000000 nmc=1000000 diag=all
    thin=25 seed=432 mchistory=detailed;
    parms beta0-beta3 0 s2 1 s2g 1;
    prior beta:    normal(0,var=1e6);
    prior s2    igama(shape=0.001,scale=0.001);
    prior s2g    igama(shape=0.01,scale=0.01);
    random gama    normal(0, var=s2g) subject=amostra;
    mu= beta0 + beta1*(especies='A.duranensis')
    + beta2*(especies='A.hypogaea')
    + beta3*(especies='A.ipaensis')
    + gama;
    model gene    normal(mu, var=s2);
run;
```

```
***EFEITO FIXO***;
TITLE1 H=1.2 'MIXED COM EFEITO FIXO - GENE';
proc mixed data=dados method=REML plot=all covtest cl;
    class especies experimento amostra;
    model GENE = especies / DDFM=KR solution cl;
    lsmeans especies / adjust=tukey;
run;
TITLE1 H=1.2 'MCMC COM EFEITO FIXO - GENE';
PROC MCMC DATA=dados outpost=cassout nbi=1000000 nmc=1000000 diag=all
    thin=25 seed=4383 mchistory=detailed;
    parms beta0-beta3 0 sigma2 1;
    prior beta:    normal(0,var=1e6);
    prior sigma2   igama(shape=0.001,scale=0.001);
    mu= beta0 + beta1*(especies='A.duranensis')
    + beta2*(especies='A.hypogaea')
    + beta3*(especies='A.ipaensis');
    model gene   normal(mu, var=sigma2);
run;
ODS GRAPHICS OFF;
ods html CLOSE;
QUIT;
```

7 Anexo

7.1 Banco de dados

Tabela 7.1: Concentração de resveratrol por grama de folha

Espécie	Genoma	Experimento	Amostras	Resveratrol(mg/g)
		1	1	416,49
A. hypogaea	AABB	1	2	218,96
		1	3	219,82
		1	1	204,52
Anfidiplóide	AABB (Sint.)	1	2	177,60
		1	3	147,43
		1	1	380,17
A. duranensis	AA	1	2	394,20
		1	3	408,24
		1	1	223,00
A. ipaensis	BB	1	2	228,78
		1	3	–
		2	1	243,39
A. hypogaea	AABB	2	2	258,26
		2	3	133,55
		2	1	220,45
Anfidiplóide	AABB (Sint.)	2	2	217,53
		2	3	223,36

Tabela 7.2: Concentração de resveratrol por grama de folha (continuação)

Espécie	Genoma	Experimento	Amostras	Resveratrol(mg/g)
		2	1	463,39
A. duranensis	AA	2	2	439,79
		2	3	398,98
		2	1	–
A. ipaensis	BB	2	2	–
		2	3	230,72
		3	1	244,07
A. hypogaea	AABB	3	2	218,07
		3	3	225,34
		3	1	204,07
Anfidiplóide	AABB (Sint.)	3	2	205,46
		3	3	142,54
		3	1	281,37
A. duranensis	AA	3	2	273,25
		3	3	308,37
		3	1	243,17
A. ipaensis	BB	3	2	204,56
		3	3	–

Tabela 7.3: Concentração da expressão do gene *resveratrol sintase*

Espécie	Genoma	Experimento	Amostras	Gene
		1	1	15,37
A. hypogaea	AABB	1	2	7,76
		1	3	8,13
		1	1	14,99
Anfidiplóide	AABB (Sint.)	1	2	–
		1	3	15,45
		1	1	29,97
A. duranensis	AA	1	2	29,97
		1	3	30,43
		1	1	59,79
A. ipaensis	BB	1	2	59,11
		1	3	59,35
		2	1	7,84
A. hypogaea	AABB	2	2	7,68
		2	3	7,68
		2	1	13,86
Anfidiplóide	AABB (Sint.)	2	2	13,87
		2	3	12,84

Tabela 7.4: Concentração da expressão do gene *resveratrol sintase* (continuação)

Espécie	Genoma	Experimento	Amostras	Gene
		2	1	–
A. duranensis	AA	2	2	30,43
		2	3	30,74
		2	1	58,75
A. ipaensis	BB	2	2	58,75
		2	3	59,05
		3	1	6,90
A. hypogaea	AABB	3	2	7,21
		3	3	6,15
		3	1	14,91
Anfidiplóide	AABB (Sint.)	3	2	14,83
		3	3	15,06
		3	1	28,47
A. duranensis	AA	3	2	28,46
		3	3	30,43
		3	1	58,95
A. ipaensis	BB	3	2	56,99
		3	3	56,70

Tabela 7.5: Prioris conjugadas [Ntzoufras(2009), p. 15]

Distribuições	Verossimilhança	Distribuição *a priori*	Parâmetro *a posteriori*	
Poisson	$Y_i \sim Poisson(\lambda)$	$\lambda \sim Gama(a,b)$	$\tilde{a} = n\bar{y} + a$	$\tilde{b} = n + b$
Binomial	$Y_i \sim Binomial(p, N_i)$	$p \sim beta(a,b)$	$\tilde{a} = \sum_{i=1}^{n} y_i + a$	$\tilde{b} = \sum_{i=1}^{n} N_i + b$
Normal (σ^2 conhecida)	$Y_i \sim Normal(\mu, \sigma^2)$	$\mu\|\sigma^2 \sim Normal(\mu_0, \sigma_0^2)$	$\tilde{\mu} = w\bar{y} + (1-w)\mu_0$ $w = \sigma_0^2/(\sigma_0^2 + \sigma^2)$	$\tilde{\sigma^2} = w\sigma^2/n$
Normal (σ^2)	$Y_i \sim Normal(\mu, \sigma^2)$	$[\mu, \sigma^2] \sim NIG(\mu_0, c, a, b)$ $[Normal(\mu_0, c\sigma^2)$ x $GI(a,b)]$	$\tilde{\mu} = w\bar{y} + (1-w)\mu_0$ $\tilde{c} = w/n$ $w = nc/(1+nc)$	$\tilde{a} = n/2 + a$ $\tilde{b} = SS/2 + b$ $SS = \left((n-1)s^2 + (\bar{y} - \mu_0)^2 w/c\right)$
Gama	$Y_i \sim Gama(\nu, \theta)$	$\nu\|\theta \sim gama(a,b)$	$\tilde{a} = n\nu + a$	$\tilde{b} = n\bar{y} + b$
Exponencial	$Y_i \sim Exponencial(\theta)$ $= Gama(1, \theta)$	$\theta \sim Gama(a,b)$	$\tilde{a} = n + a$	$\tilde{b} = n\bar{y} + b$
Binomial negativa	$Y_i \sim BN(p, K_i)$	$p \sim beta(a,b)$	$\tilde{a} = \sum_{i=1}^{n} K_i + a$	$\tilde{b} = \sum_{i=1}^{n} y_i + b$
Multinomial	$Y_i \sim Multinomial(\alpha, N_i)$	$\alpha \sim Dirichlet(\alpha_0)$	$\tilde{a} = \sum_{i=1}^{n} y_i + \alpha_0$	
Família exponencial (τ *conhecido*)	$Y_i \sim FE(\vartheta, \phi, a(), b(), c())$	$exp\left[\frac{\vartheta\vartheta_0 - \tau_0 b(\vartheta)}{a(\phi)}\right]$	$\tilde{\vartheta} = n\bar{y} + \vartheta_0$	$\tilde{\tau} = n + \tau_0$

7.2 Gráficos

7.2.1 Produção de resveratrol

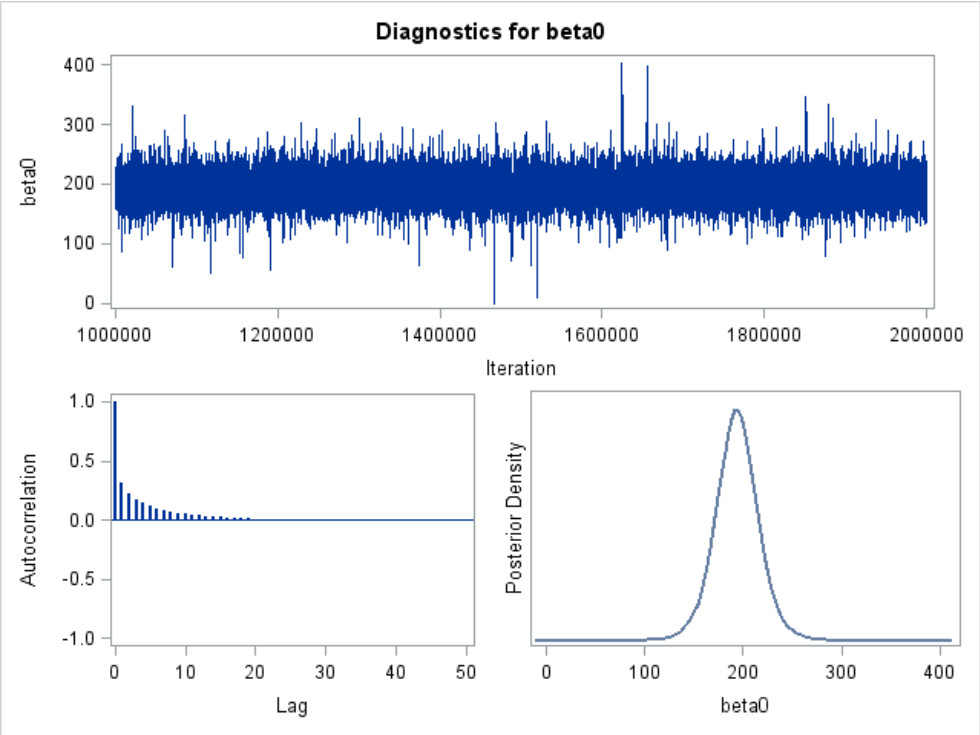

Figura 7.1: Diagnóstico gráfico de convergência do parâmetro β_0 - resveratrol com efeito aleatório

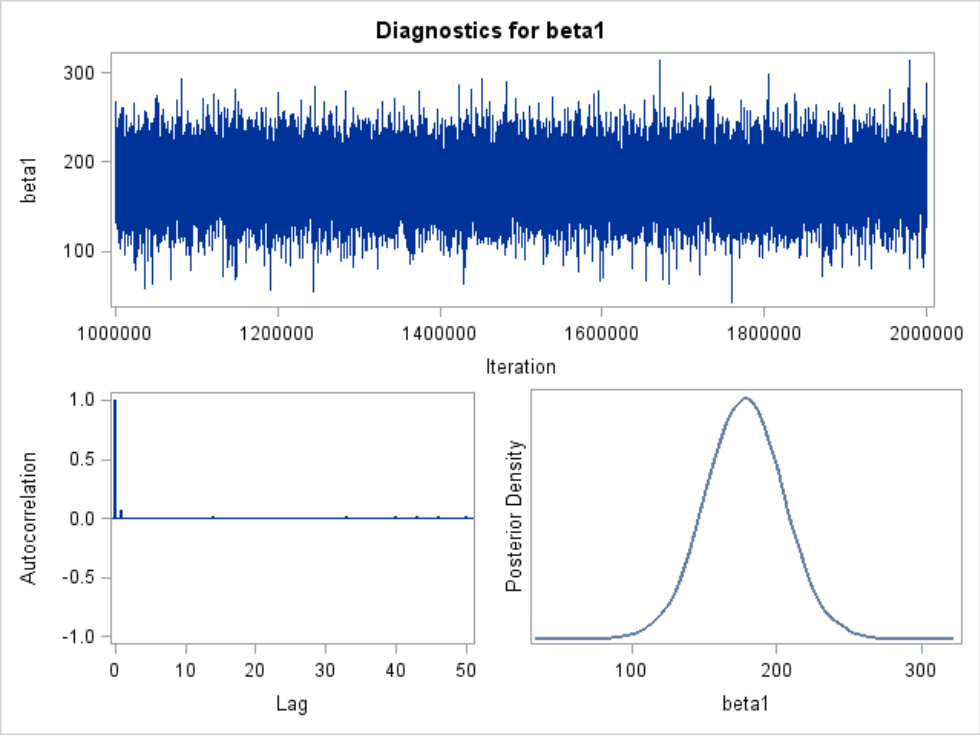

Figura 7.2: Diagnóstico gráfico de convergência do parâmetro β_1 - resveratrol com efeito aleatório

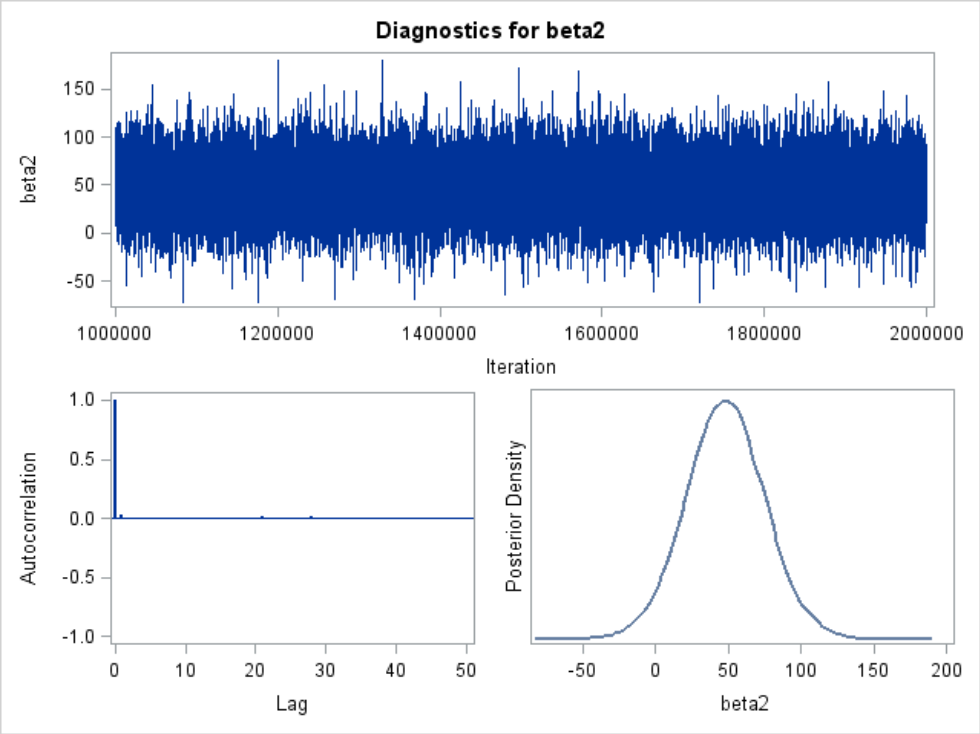

Figura 7.3: Diagnóstico gráfico de convergência do parâmetro β_2 - resveratrol com efeito aleatório

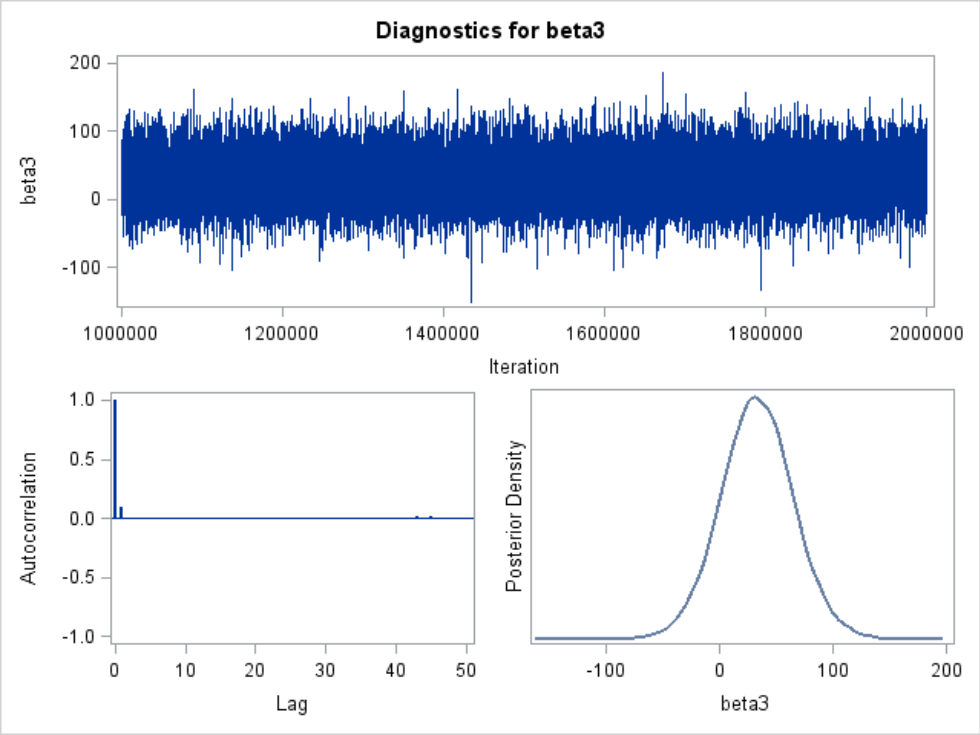

Figura 7.4: Diagnóstico gráfico de convergência do parâmetro β_3 - resveratrol com efeito aleatório

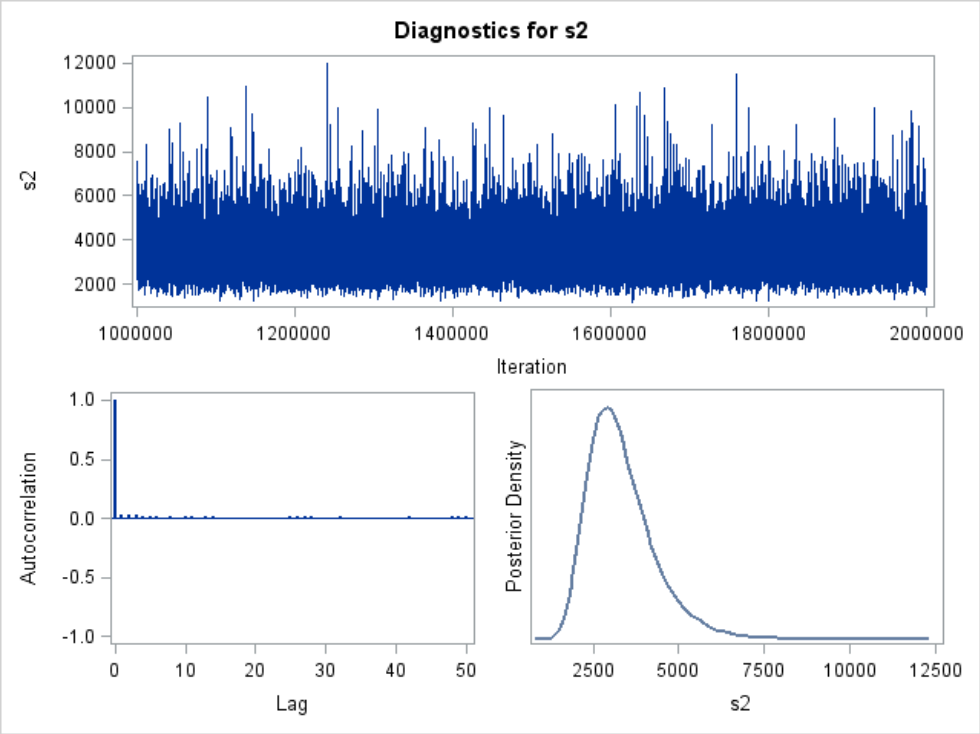

Figura 7.5: Diagnóstico gráfico de convergência do parâmetro σ^2 - resveratrol com efeito aleatório

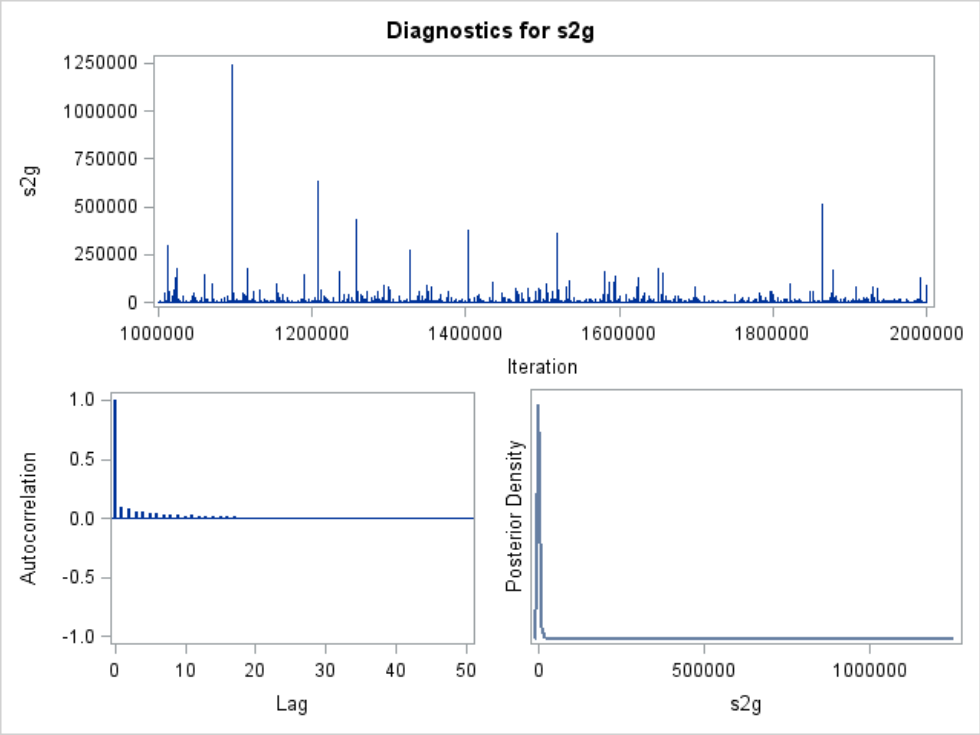

Figura 7.6: Diagnóstico gráfico de convergência do parâmetro σ^2_{exp} - resveratrol com efeito aleatório

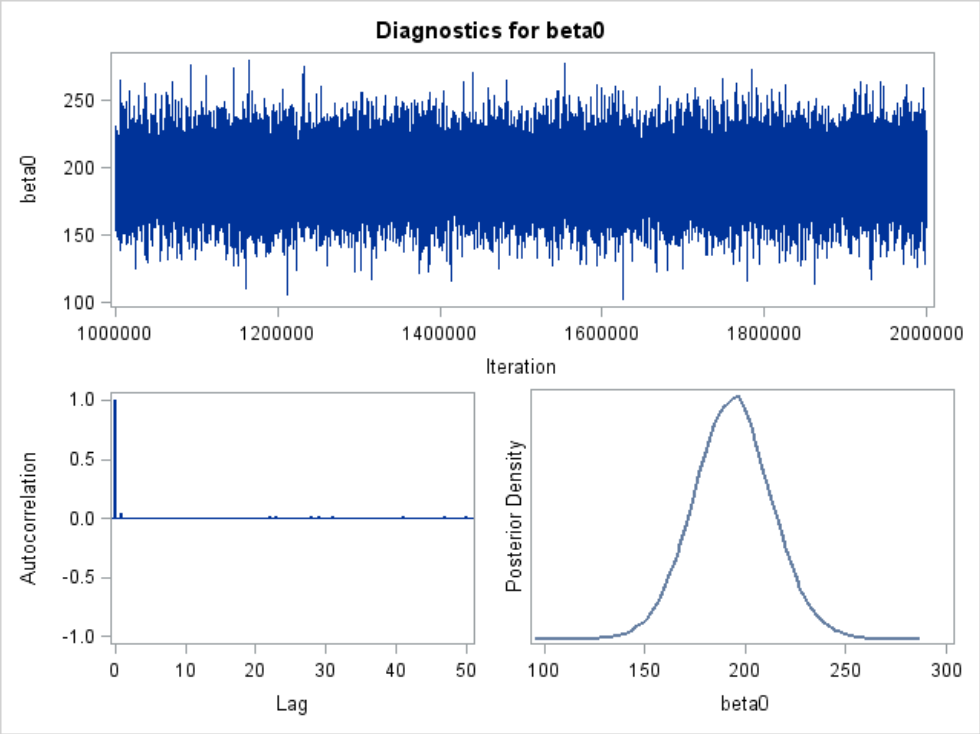

Figura 7.7: Diagnóstico gráfico de convergência do parâmetro β_0 - resveratrol sem efeito aleatório

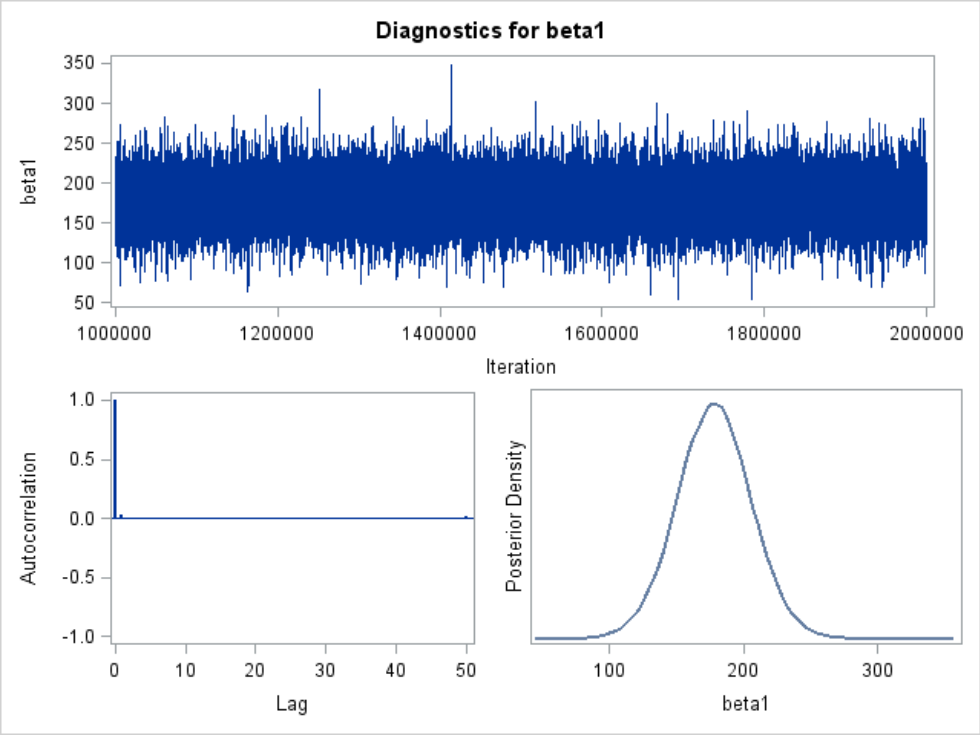

Figura 7.8: Diagnóstico gráfico de convergência do parâmetro β_1 - resveratrol sem efeito aleatório

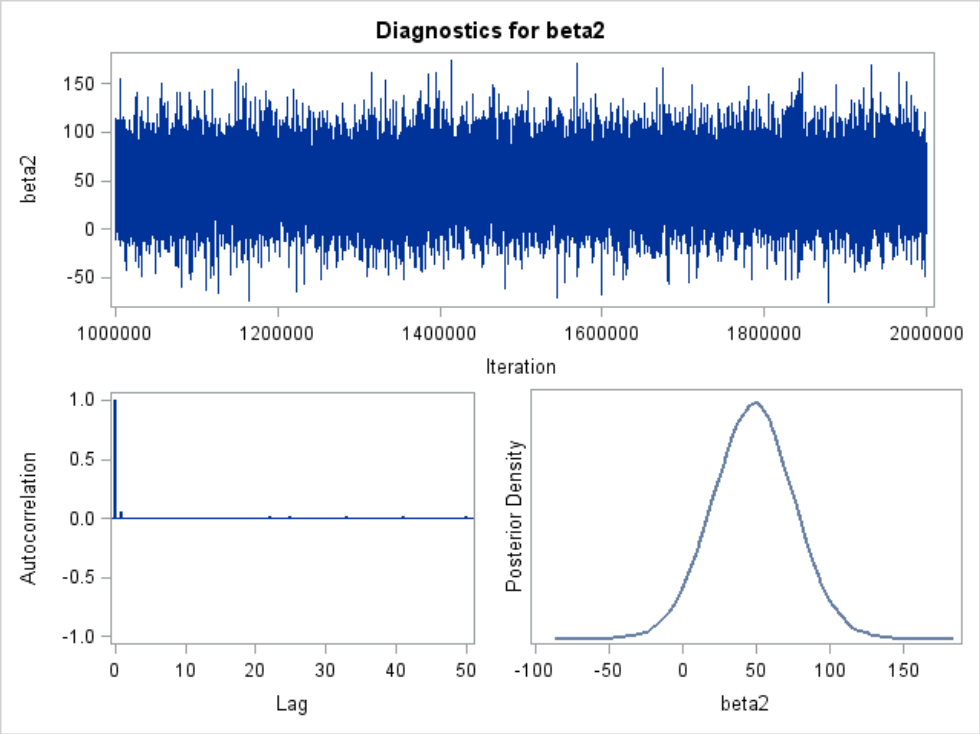

Figura 7.9: Diagnóstico gráfico de convergência do parâmetro β_2 - resveratrol sem efeito aleatório

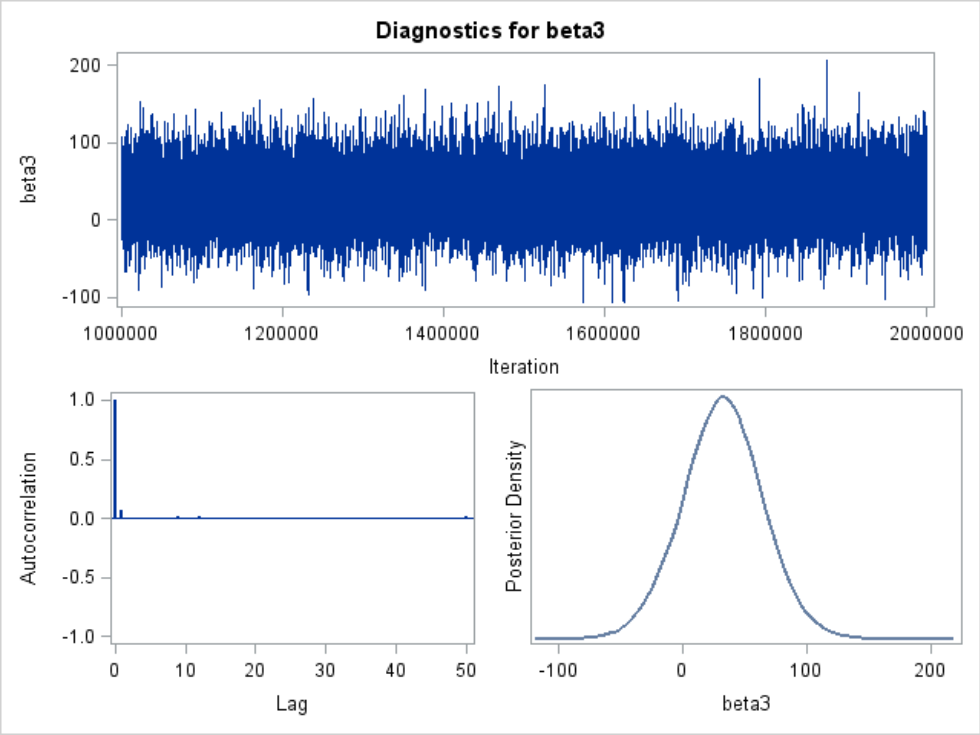

Figura 7.10: Diagnóstico gráfico de convergência do parâmetro β_3 - resveratrol sem efeito aleatório

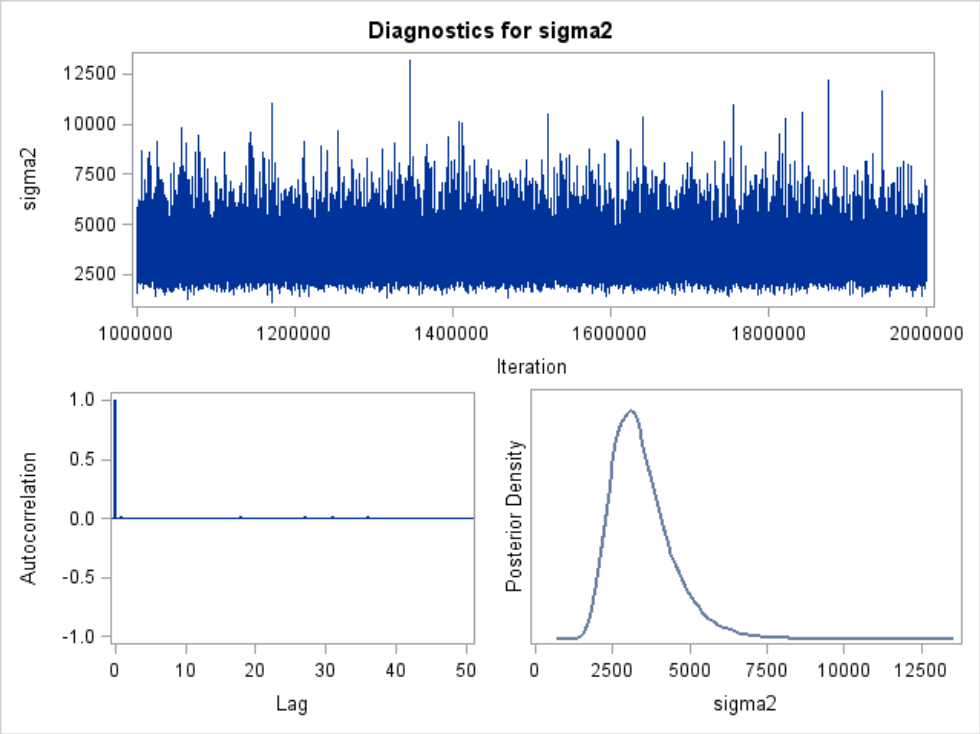

Figura 7.11: Diagnóstico gráfico de convergência do parâmetro σ^2 - resveratrol sem efeito aleatório

7.2.2 Gene *resveratrol sintase*

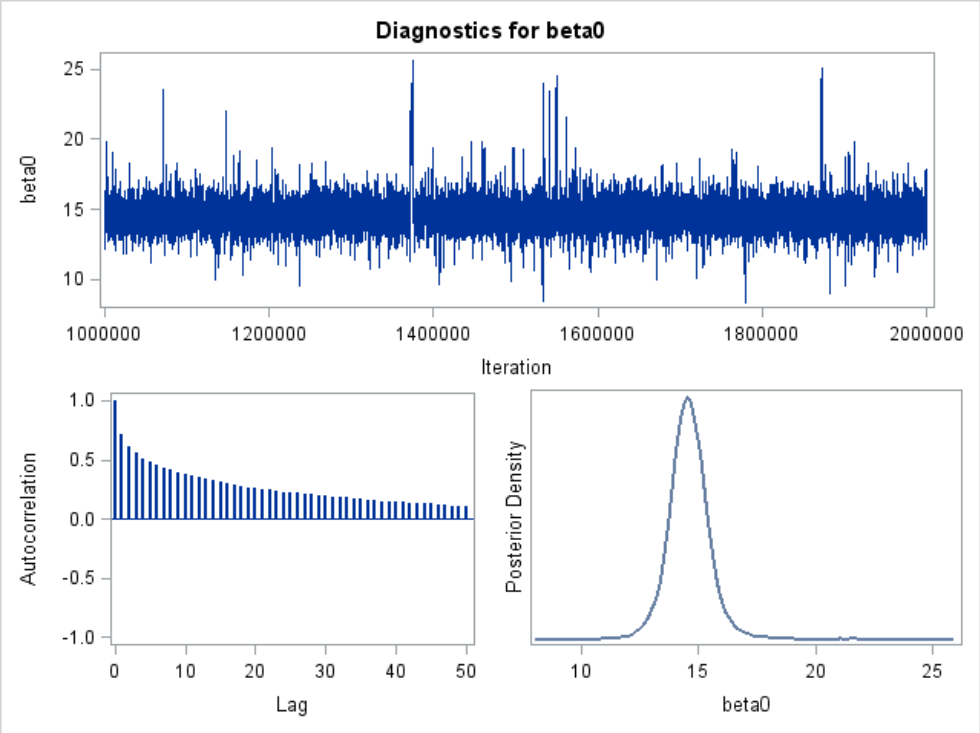

Figura 7.12: Diagnóstico gráfico de convergência do parâmetro β_0 - gene *resveratrol sintase* com efeito aleatório

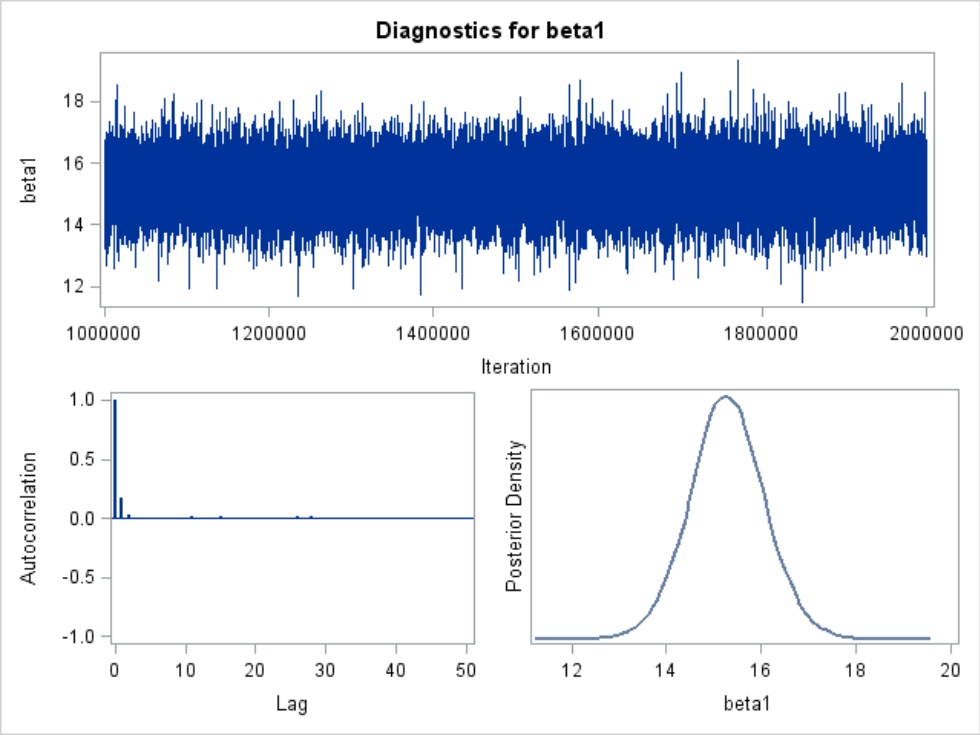

Figura 7.13: Diagnóstico gráfico de convergência do parâmetro β_1 - gene *resveratrol sintase* com efeito aleatório

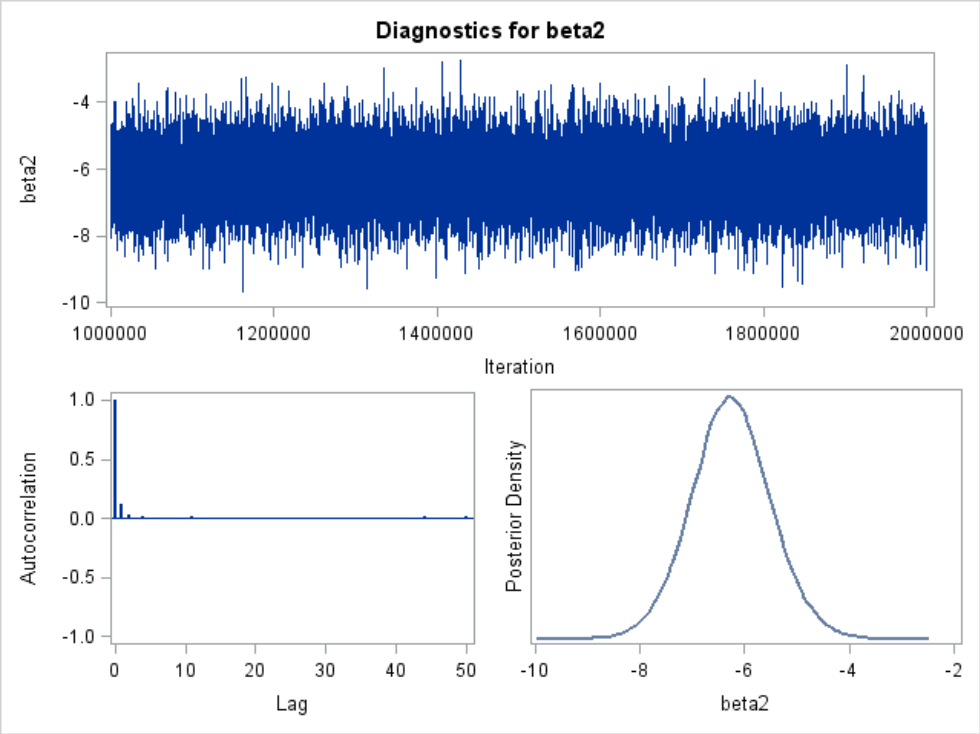

Figura 7.14: Diagnóstico gráfico de convergência do parâmetro β_2 - gene *resveratrol sintase* com efeito aleatório

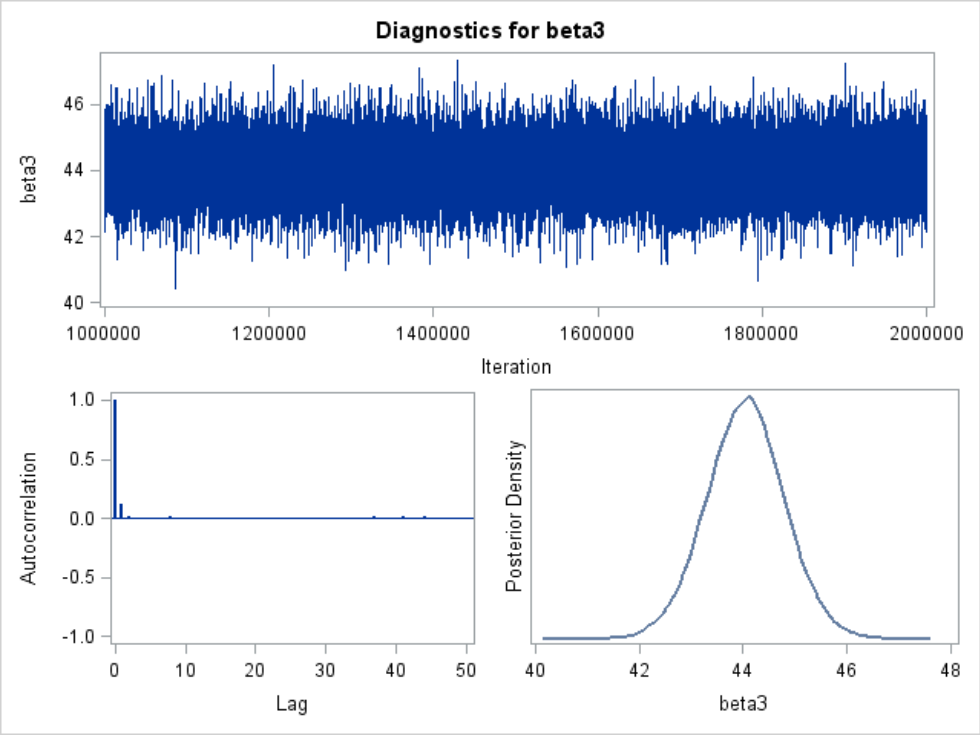

Figura 7.15: Diagnóstico gráfico de convergência do parâmetro β_3 - gene *resveratrol sintase* com efeito aleatório

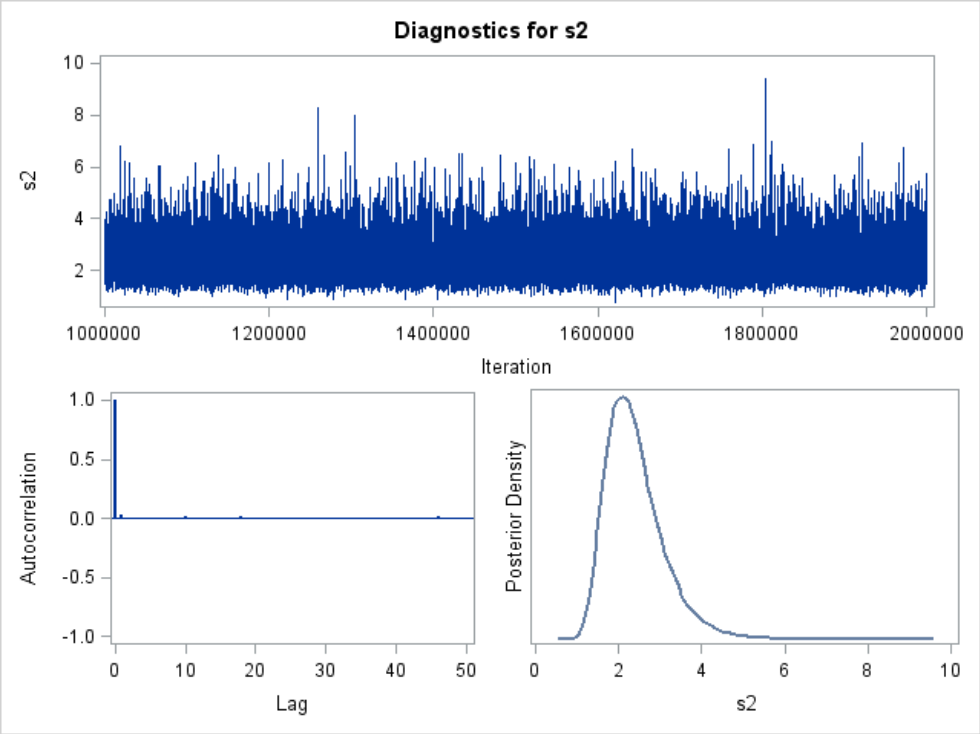

Figura 7.16: Diagnóstico gráfico de convergência do parâmetro σ^2 - gene *resveratrol sintase* com efeito aleatório

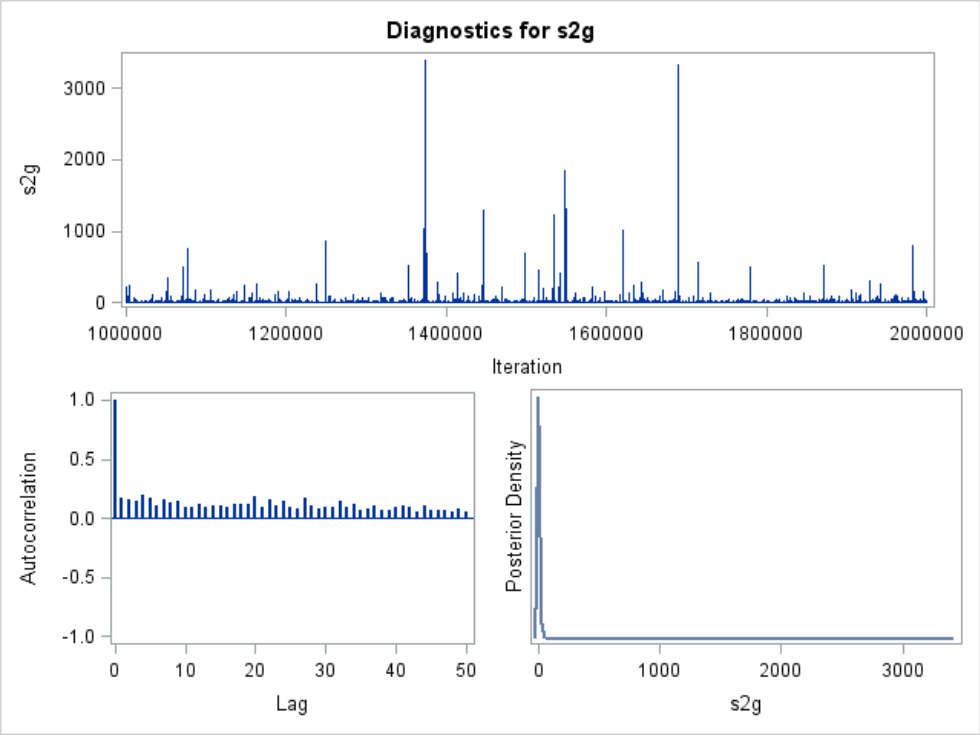

Figura 7.17: Diagnóstico gráfico de convergência do parâmetro σ^2_{exp} - gene *resveratrol sintase* com efeito aleatório

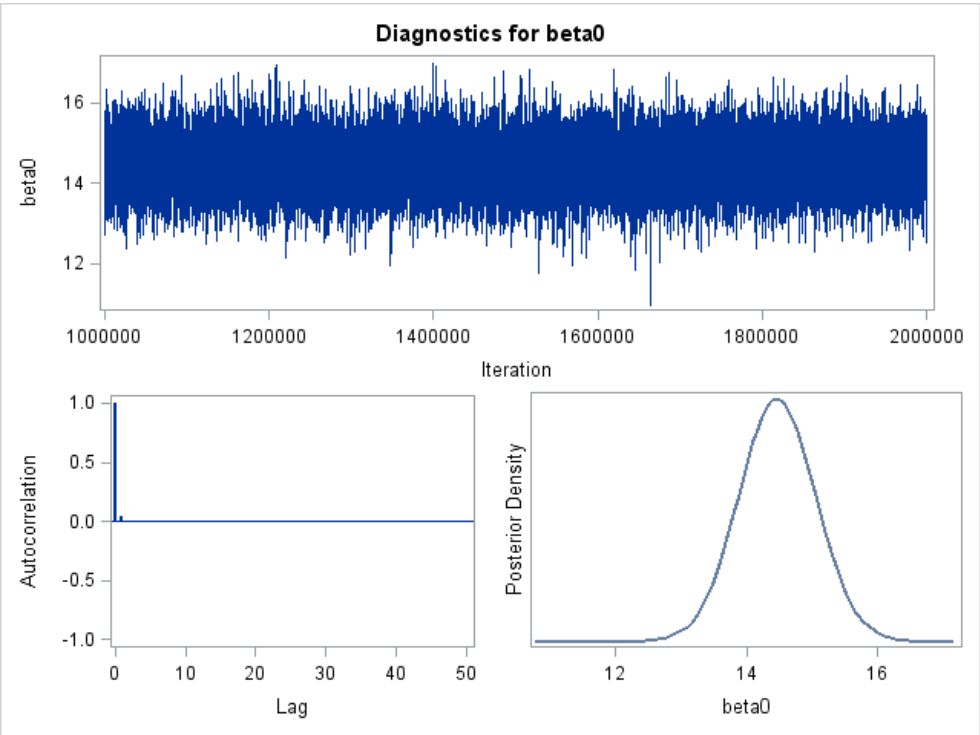

Figura 7.18: Diagnóstico gráfico de convergência do parâmetro β_0 - gene *resveratrol sintase* sem efeito aleatório

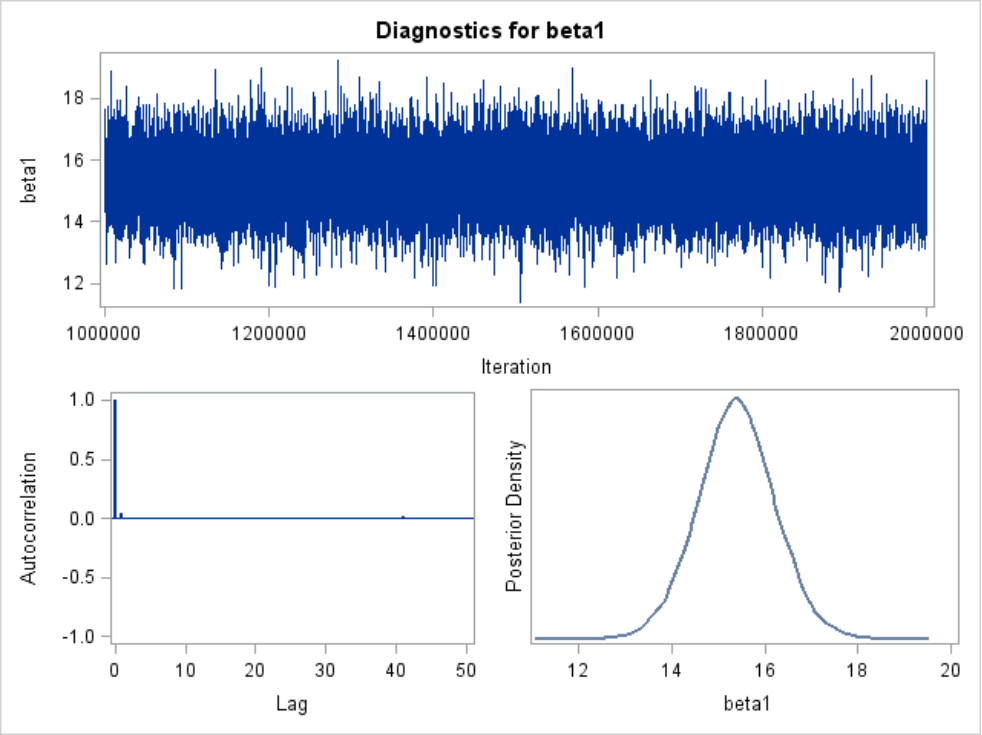

Figura 7.19: Diagnóstico gráfico de convergência do parâmetro β_1 - gene *resveratrol sintase* sem efeito aleatório

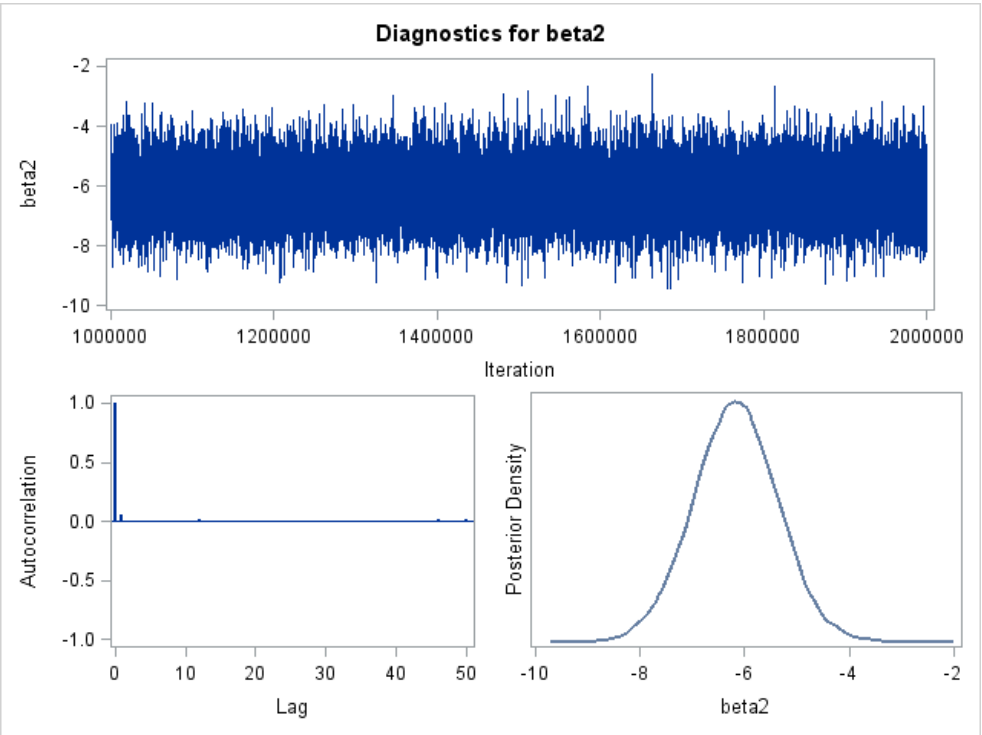

Figura 7.20: Diagnóstico gráfico de convergência do parâmetro β_2 - gene *resveratrol sintase* sem efeito aleatório

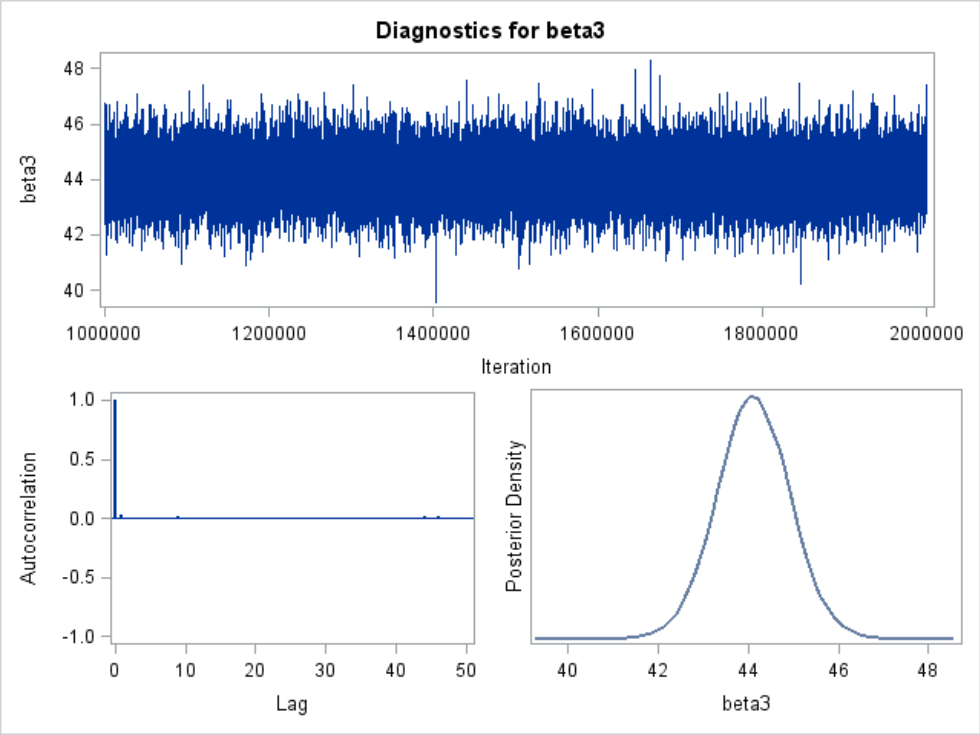

Figura 7.21: Diagnóstico gráfico de convergência do parâmetro β_3 - gene *resveratrol sintase* sem efeito aleatório

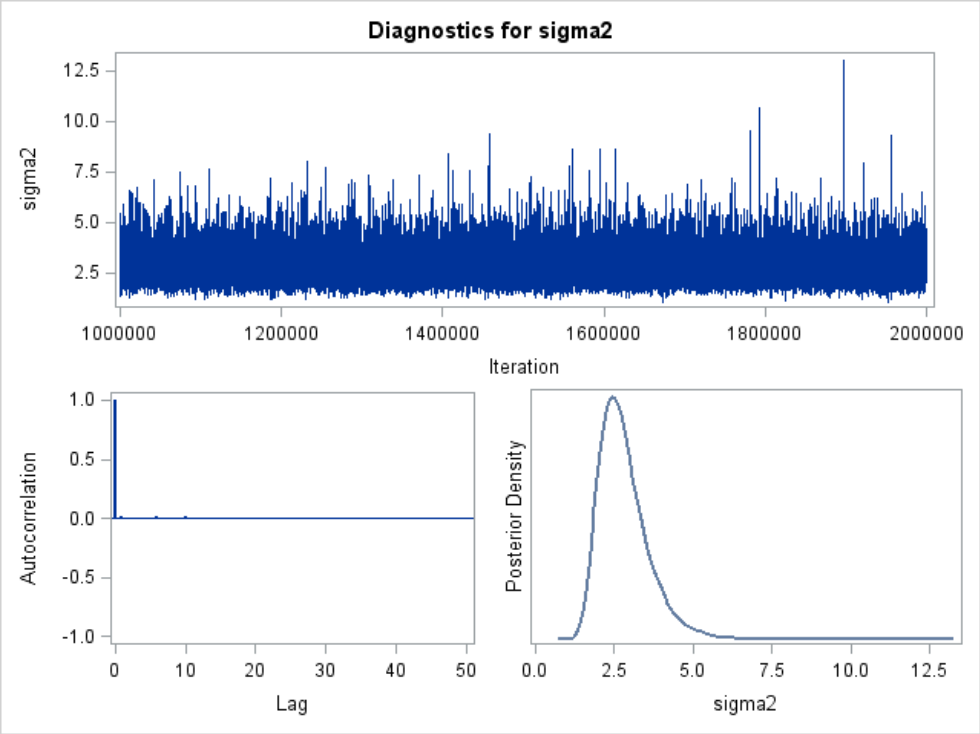

Figura 7.22: Diagnóstico gráfico de convergência do parâmetro σ^2 - gene *resveratrol sintase* sem efeito aleatório