Universidade de Brasília
Instituto de Ciências Exatas
Departamento de Estatística

Modelos de Regressão com Resposta Ordinal para Avaliação de Textura do Arroz

Geiziane Silva de Oliveira

Trabalho de Conclusão de Curso apresentado ao Departamento de Estatística da Universidade de Brasília, como parte dos requisitos para a obtenção do título de Bacharel em Estatística.

Brasília
2015

Geiziane Silva de Oliveira

Modelos de Regressão com Resposta Ordinal para Avaliação de Textura do Arroz

Trabalho de Conclusão de Curso apresentado ao Departamento de Estatística da Universidade de Brasília, como parte dos requisitos para a obtenção do título de Bacharel em Estatística.

Orientador: Prof. Dr. **George von Borries**

Brasília
2015

,

Agradecimentos

A Deus, agradeço pela força e proteção concebida para a conclusão deste trabalho.

A Embrapa Arroz e Feijão (CNPAF), agradeço pela disponibilização dos dados do projeto QualiArroz. Especialmente à pesquisadora Dr. Priscila Zaczuk Bassinello pela disponibilização dos dados e informações prestadas que permitiram a realização deste trabalho.

À amiga Érica Rios, por ter me dado essa oportunidade maravilhosa de continuar o seu trabalho de conclusão de graduação. Por ter me apoiado bastante nesse estudo. Por toda amizade e carinho.

Ao professor Dr. George von Borries, por toda dedicação na orientação desse trabalho, e por todo conhecimento que sempre me proporcionou durante a minha vida acadêmica. Além, disso por toda atenção, conselhos e amizade que contribuíram muito para a minha formação acadêmica e pessoal.

Às professoras Dra. Juliana Betini Fachini e Ma. Maria Tereza Leão Costa por terem acreditado em mim e por todo incentivo, conselhos e os ensinamentos que foram essenciais para o meu desenvolvimento acadêmico e pessoal.

Ao professor Me. Luís Gustavo do Amaral Vinha por todos os ensinamento e conselhos durante a orientação do projeto Banco de Questões para Estatística Básica.

Ao professor Dr. Bernardo Borda de Andrade por ter aceito participar da banca examinadora desse trabalho. E pelas aulas com "efeitos especiais e sonoros"que me proporcionou muitos aprendizados.

À professora Dr. Joanlise Marco de Leon pela participação na banca examinadora desse trabalho. Por todas contribuições para melhorar esse trabalho. E pelos conselhos tão sábios.

Ao Professor Me. Demerson André Polli por ter sido um dos incentivadores na escolha do curso de estatística, da qual me orgulho. Pela amizade e ensinamentos durante toda a minha vida acadêmica.

Ao professor Eduardo Nakano, pelas contribuições estatística dadas na apresentação deste trabalho.

Aos meus pais, por todos os valores que me ensinaram e por toda compreensão.

À minha irmã Maria José Oliveira da Silva, por todo apoio em diversos aspectos

durante á minha vida, e principalmente na realização desse curso. Por toda atenção e conselhos.

Ao meu cunhado Aparecido Lima, por toda atenção e pelas caronas para Unb durante toda a realização desse curso.

À minha amiga Jéssica Oliveira e sua mãe Rute Oliveira, por todo apoio e acolhimento, que foram essenciais para conclusão deste trabalho. E pela, amizade, conselhos e cuidado.

A todos os meus irmãos e familiares pelo apoio, especialmente a Neide, Geovane, Gilberto e Gilmar.

Às minhas amigas, Denise Rayanne e Cíntia Soares, pela amizade e compreensão.

Aos meus amigos fofos Erique Pereira e Felipe Quintino, pelas contribuições estatísticas e amizade.

Aos funcionários do Departamento de Estatística, especialmente a Tathyanna Cordeiro e Edenilson.

Por fim, ao Bonde da Estatística, Agda Galletti, Andressa Lima, Márcia Maia, Mariana Fehr, Mateus Carbone, Rodrigo Ferrari. Pelos dias de estudo e muito aprendizado conjunto, por todos os domingos de Unb, inclusive nas datas comemorativas, por toda a vida de estudos no aquário e no Espaço Chiarini. Pelas inúmeras madrugadas de estudos, caronas e hospedagens. Por todo carinho e amizade.

Resumo

Neste trabalho estudamos a técnica de regressão logística Bayesiana para a avaliação de medidas de textura de arroz. Esse estudo é parte do projeto QualiArroz da Embrapa Arroz e Feijão (CNPAF). Sendo o objetivo principal estudar modelos para averiguar a relação existente entre a avaliação sensorial de textura e as medidas instrumentais de viscosidade. A avaliação da textura do arroz é um aspecto muito importante porque incide na qualidade dos grãos com reflexo no preço, e está associada com outras variáveis, tal como tamanho do grão, coloração, clima e tipo do terreno. A principal ferramenta para avaliar a textura do arroz é a análise sensorial, porém é muito cara e trabalhosa e há poucos profissionais para realizar esse trabalho. Portanto, buscou-se estudar a substituição da avaliação sensorial por medidas instrumentais de viscosidade, pois são mais fáceis e baratas de se obter. Por conseguinte, foram aplicados modelos para predição da avaliação sensorial por meio de medidas de viscosidade utilizando as técnicas de Análise de Componentes Principais e Regressão Logística Bayesiana e Clássica (binária). A qualidade da previsão dos modelos foi avaliada pela estimativa do erro de classificação, no caso do modelos clássicos foi estimada por meio de validação cruzada. Os modelos Bayesianos foram aplicados para aperfeiçoar o processo de classificação devido ao problema de poucas observações. As análises mostraram que os resultados para as técnicas de regressão logística Bayesiana e Clássica foram semelhantes. E, portanto, sugere que pode ser feita a substituição da avaliação sensorial por medidas de viscosidade, porém, ainda é necessário o aperfeiçoamento das categorias informadas na avaliação sensorial para ter-se um resultado mais preciso.

Palavras-chave: Textura de Arroz, Avaliação Sensorial, Medidas de Viscosidade, Regressão Logística Bayesiana, Regressão Logística Clássica, Componentes Principais.

Sumário

Lista de Figuras

Lista de Tabelas

Introdução e Justificativa

Este trabalho tem como objetivo abordar a técnica de regressão logística Bayesiana para a avaliação de medidas de textura de arroz. Usualmente, a avaliação da textura do arroz é realizada por avaliação sensorial, que verifica características dos grãos por meios dos sentidos (visão, audição, tato e olfato). Outras formas de avaliação envolvem as análises instrumentais de viscosidade e de textura. Essas técnicas mensuram características importantes sobre o cozimento do arroz.

As análises realizadas neste estudo foram feitas com a base de dados sobre a textura e cultivo do arroz dos anos de 2013 e 2014 fornecida pela Embrapa CNAPAF-GO. A base de dados (que é composta por 18 variáveis, sendo que nove são quantitativas e nove qualitativas) contém variáveis como medidas de textura sensorial, medidas de textura instrumental e medidas instrumentais de viscosidade (perfil viscoamilográfico). Algumas variáveis são relacionadas ao cultivo do arroz e identificação das medidas analisadas.

A avaliação da textura do arroz é um aspecto muito importante porque incide sobre a qualidade dos grãos com reflexo no preço, e está associada a outras variáveis, tal como tamanho do grão, coloração, clima e tipo de terreno.

Segundo a Norma ISO [8] a textura pode ser definida como o "conjunto de propriedades mecânicas, geométricas e de superfície de um produto, detectáveis pelos receptores mecânicos e tácteis e, eventualmente pelos receptores visuais e auditivos". Rios [16] destaca que para o caso específico do arroz, a textura é consequência da estrutura interna do grão e é determinada através do tato.

A pesquisa realizada por Rios [16] procura estimar a classificação sensorial do arroz através de medidas laboratoriais (Perfil Viscoamilográfico) que são mais fáceis e baratas de se obter. Nesse estudo procuramos aperfeiçoar este processo de classificação utilizando o modelo logístico Bayesiano. Os dados foram disponibilizados pela Embrapa Arroz e Feijão (CNPAF), por meio do projeto QualiArroz, e relacionam medidas instrumentais com respostas sensoriais numa escala de Likert de 7 níveis. Alguns dos níveis não foram observados ou possuem poucas observações, o que sugere a modificação da escala original e a aplicação de modelos que permitam fazer inferências mais precisas com pequenas amostras.

O objetivo principal do estudo é a avaliação de modelos para verificar a relação exis-

tente entre medidas de textura sensorial, medidas de textura instrumental e medidas instrumentais de viscosidade. A substituição da avaliação sensorial por medidas instrumentais de viscosidade se justifica na Embrapa, uma vez que a análise sensorial do arroz é demorada e cara porque envolve o treinamento de pessoas, e a medida instrumental de textura é trabalhosa e envolve equipamento de alto custo. Ressalta-se ainda que o maior problema enfrentado pelos pesquisadores da CNPAF com relação a análise da textura sensorial é que o número de avaliadores sensoriais é bastante pequeno para a demanda existente desse tipo de análise.

As medidas de textura instrumental servem de padrão para a avaliação da análise sensorial. (já mostrada por Rios [16]) Uma vez validadas as medidas sensoriais, procedemos no estudo para substituição da análise sensorial por medidas instrumentais de viscosidade.

Primeiramente, será realizada a análise descritiva dos dados para conhecer o comportamento das variáveis em estudo. Em seguida, serão ajustados alguns modelos de regressão logística, visto que a regressão linear não é válida, pois as suposições de normalidade dos erros não são atendidas.

Modelos de regressão logística são usados quando se tem interesse em verificar a influência que um conjunto de variáveis explicativas desempenha em relação a ocorrência ou não de um evento (Hosmer e Lemeshow [6]).

As covariáveis em estudo são as medidas de viscosidades, que estão altamente correlacionadas. Assim, foi utilizada a técnica de componentes principais para reduzir a dimensão do estudo e evitar problemas de multicolinearidade nos dados. Logo, os modelos foram ajustados considerando-se apenas duas covariáveis, que são os escores das componentes principais das medidas de viscosidade.

Os modelos Bayesianos foram ajustados com base em *prioris* não informativas e informativas, baseadas nos resultados dos modelos clássicos ajustados para o ano de 2013, segundo o tipo de arroz. As distribuições *a posteriori* foram obtidas via o metódo MCMC por meio do algoritmo de Metropolis-Hastings implementados na `PROC MCMC` do software SAS. As análise realizadas nesse estudo foram realizadas por meio do sotware SAS. O software SAS foi utilizado através da parceria acadêmica entre SAS Institute Brasil e Departamento de Estatística da Universidade de Brasília-UnB.

Ressalta-se ainda que serão comparados os resultados dos modelos clássicos com os Bayesianos, mais especificamente, com a regressão logística binária já implementada no estudo de Rios [16].

Capítulo 1

Revisão Bibliográfica

1.1 Qualidade e padronização do Arroz no Brasil

O Arroz é considerado um dos alimentos com maior valor comercial em países em desenvolvimento. É também o alimento com maior balanceamento nutricional, sendo ainda parte essencial da dieta básica dos brasileiros, onde o consumo médio é por volta de 45kg ao ano. Segundo Ferreira et. al [4], o cultivo do arroz no Brasil sempre esteve ligado com a relação assimétrica entre quantidade e qualidade, que a preocupação com a qualidade era referente ao interesse privado e a quantidade ao interesse dos agentes públicos. Ferreira et. al [4] ressaltam ainda que, no início do século XIX começou o processo de preocupação com a qualidade e padronização do arroz no Brasil, período marcado diretamente pelas leis de fiscalização de produtos agropecuários. Os tipos de cultivares no Brasil também apresentaram aspectos sobre qualidade e padronização, sendo destacados os cultivares de arroz irrigado (terras baixas) e sequeiro (terras altas). O arroz irrigado desenvolveu-se mais na região Sul, com maior representatividade no Rio Grande do Sul. O arroz de terras altas depende apenas das condições endo climáticas e em 1974 representava 80% da produção de arroz no Brasil. Este arroz é desenvolvido principalmente na parte central do Brasil. Estas cultivares passaram por diversas fases e produziam diversos tipos de arroz como os de grãos finos, médios, curtos e solto. O preço foi bastante influenciando pela qualidade medida pelo gosto do consumidor. Os produtores de cultivares de terras irrigadas sofreram consequências no preço devido a não adequação das exigências dos consumidores. Em 1974 surgiu o Centro Nacional de Pesquisas sobre o Arroz e Feijão (CNPAF) da Embrapa que dedicou esforços ao estudo sobre o melhoramento do arroz em cultivares de terras altas no Brasil. Hoje, a Embrapa Arroz e Feijão se dedica também ao estudo do arroz irrigado.

1.2 Análise da Textura Sensorial do Arroz no projeto QualiArroz.

O projeto QualiArroz foi realizado pela Embrapa CNPAF nos anos de 2012 a 2014 com o objetivo de aplicar metodologias refinadas na caracterização de diferentes genótipos brasileiros de arroz de terras altas e irrigado quanto às propriedades sensoriais e de amido associadas à qualidade culinária do grão. Entre os objetivos específicos destaca-se a identificação de metodologia de avaliação sensorial de arroz que seja viável na prática laboratorial e treinamento de painel sensorial.

Minim [13] define análise sensorial como uma ferramenta usada para evocar, medir, analisar, interpretar reações às características de um alimento por meios dos sentidos (visão, olfato, tato e audição).

A análise sensorial desenvolveu bastante no século XIX com o processo de industrialização, onde cresceu a preocupação em avaliar a textura de alimentos. A avaliação sensorial em muitos países passou por quatros fases. No Brasil chegou em 1967, no Instituto Agronômico de Campinas, sendo no início testada apenas para café (Rios [16]).

No Projeto QualiArroz, os avaliadores foram treinados para realizarem a avaliação sensorial por tato. A área de condução dos testes consiste de seis cabines individuais conforme mostrado na Figura 1.1, ou seja, uma para cada avaliador. Todas as cabines são brancas e cada uma delas possui três luzes, sendo de colorações branca, vermelha e azul. Para conduzir o processo de amostragem, as cabines possuem uma escotilha que permite entregar e retirar as amostras sem contato com os avaliadores. A comunicação entre o técnico que prepara as amostras para avaliação é dada por meio de uma luz que fica do lado de fora da cabine e o interruptor dentro da cabine.

Figura 1.1: Local de condução da análise sensorial na Embrapa Arroz e Feijão.

As amostras de arroz são levadas aos avaliadores de forma sequencial e uma após a outra são submetidas a análise sob a luz de coloração branca. Depois da avaliação das

amostras, os avaliadores preenchem um questionário com dois atributos, onde um trata da avaliação sensorial da dureza e outro da pegajosidade, sendo dispostos em numa escala de 7 níveis. As categorias referente a dureza são: Extremamente Macio, Macio, Ligeiramente Macio, Macio com Centro Firme, Levemente Firme, Muito Firme e Extremamente Firme. Para a pegajosidade as sete categorias são: Extremamente Solto, Muito Solto, Solto, Ligeiramente Solto, Pegajoso, Muito Pegajoso, Extremamente Pegajoso.

A análise sensorial é muito influenciada por fatores psicológicos e fisiológicos, pois o instrumento de avaliação é o homem. Assim, podem ocorrer muitos erros causados por fatores da personalidade e atitudes do avaliador ou pela maneira que o organismo se adapta a um determinado estímulo. Ressalta-se que a análise sensorial é um processo demorado e caro e o maior problema enfrentado na CNPAF é a indisponibilidade de técnicos para fazer as análise. A CNPAF conta com um número pequeno de avaliadores, esses não restringe apenas ao estudo do arroz.

1.3 Análise Instrumental da textura de arroz no projeto QualiArroz

Como descrito anteriormente, a análise sensorial é influenciada por diversos fatores, além de ter alto custo por envolver treinamento de pessoas. Devido à necessidade constante de atualização e substituição dos avaliadores, existem equipamentos que procuram determinar a textura de alimentos.

O texturômetro é um equipamento que analisa a textura através da resistência à deformação apresentada pelos grãos. Na Embrapa CNPAF, a análise instrumental de textura segue o método otimizado apresentado por Sesmat & Meullennet (2011). A textura do arroz é avaliada quanto aos parâmetros dureza e pegajosidade através amostras de arroz cozido no Texturômetro (TA.XT.plus, Stable Micro Systems, Godalming, Surrey, UK, Sesmat [20]).

O texturômentro registra, em um gráfico, à resistência a deformação apresentada pelos grãos do arroz. A compreensão do material é feita por uma sonda, com diâmetro igual ou superior ao diâmetro da amostra.

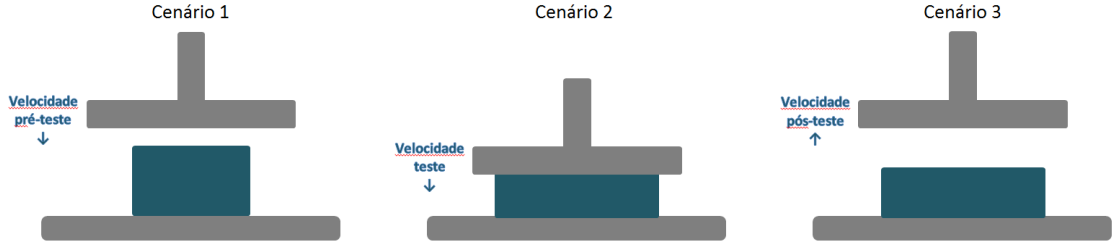

Figura 1.2: Análise de compressão

Segundo Rios [16] *"No cenário 1, a sonda que inicialmente se encontra a uma velocidade de pré-teste vai abaixando em direção a plataforma de análise do texturômetro onde se encontra a amostra de arroz. Isso ocorre até que a sonda atinja uma força chamada* trigger *que é a evidência de que a sonda entrou em contato com a amostra. Após o registro do trigger a sonda comprime a amostra até determinada altura quando muda da velocidade de compressão pré-teste para a velocidade de teste. De modo que, a dureza instrumental é a força máxima, medida em Newton, registrada durante a análise de compressão no cenário 2. Já a pegajosidade instrumental é dada pela energia de adesão medida após a compressão de uma amostra, durante a volta da sonda à sua posição inicial no cenário."*

1.4 Medidas instrumentais de viscosidade no Projeto QualiArroz

Algumas propriedades sobre o cozimento arroz são determinadas pelas características físico-químicas, tais como a composição do arroz que é de 95% de amido formado por moléculas de glicose. A amilose, é a molécula responsável pelas propriedades da textura do arroz e representa cerca de 20 a 30% do amido. E, visto a indisponibilidade de mensuração do amido no grão de arroz, as medidas instrumentais de viscosidade procuram mensurar a amilose (Bueno [2]).

De acordo com Ferreira et. al [4] o teor de amilose aparente nos grãos varia entre 0 e 35 %. Portanto, os cultivares com teor ($< 21\%$),(21 a 25%) e ($> 25\%$) são classificados como de teor baixo, intermediário e alto, respectivamente. Altos teores de amilose geralmente produzem grãos secos e soltos, que podem endurecer após o resfriamento. Baixos teores podem resultar em grãos macios, e pegajosos no cozimento. E teores intermediários, após o cozimento resultam em grãos secos e soltos. Este último é o tipo preferido pelos brasileiros.

O processamento da amostra dos grãos para a mensuração das medidas instrumentais de viscosidade é realizado imediatamente após a colheita, debulhamento e secagem natural dos grãos de arroz. Os grãos que ainda possuírem casca são processados na Suzuki MT 10 mil (Santa Cruz do Rio Pardo, São Paulo, Brasil. (Moreira et. al [14]).

O banco de dados contém 9 variáveis referentes as medidas de viscosidade. TAAFIA é uma medida de viscosidade que avalia o teor de amilose aparente nos grãos de arroz através do sistema FIA (Análise por injeção de Fluxo) da Foss Tecator (FIAStar 5000, Dinamarca). O teor de amilose aparente nos grãos é medido a partir de uma técnica calorimétrica que utiliza uma solução de iodo potássio como indicador. O procedimento usa aproximadamente 90 grãos de arroz e o teor de amilose é verificado à partir de uma curva de calibração.(Rios [16]).

A variável TAASEC representa a medida de viscosidade que avalia o teor de amilose absoluto dos grãos por meio do Sistema de Cromatografia Líquida de Alta Eficiência

(HPLC). Prominence (Shimadzu, Kyoto, Japão) acoplado com o detector de índice de Refração, conforme metodologia de FITZGERALD, McCOUCH e HALL, (2009) [3]. Para mais detalhes sobre a metologia ver Rios [16].

A temperatura de gelatinização (TG) mensura a reação dos grãos em uma solução alcalina. Esse índice avalia o tempo de cozimento, que é verificado por características do amido. Temperatura de gelatinização intermediária (69 a 73^oC) resulta em menor tempo e menor quantidade de água para o cozimento. A TG, que é medida indiretamente pelo teste de dispersão alcalina, de cada amostra é determinada pelo número de grãos de arroz, da amostra utilizada, multiplicado pelo seu correspondente nível de dispersão alcalina.

As características referentes à viscosidade dos grãos do arroz são determinadas pelo Rapid Visco Analyser (RVA) da marca Newport Scientific série 4 mostrado na Figura (1.3), de acordo com o procedimento definido por (Teba [22]).

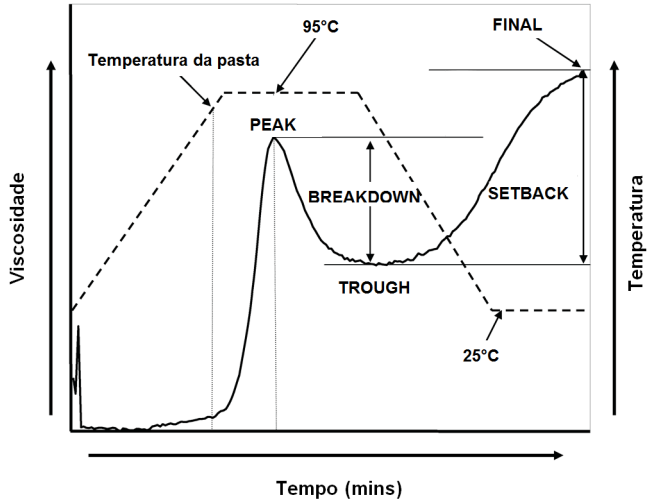

Figura 1.3: Medidas de Perfil Viscoamilográfico

As medidas de perfil viscoamilográfico são:

- *Viscosidade de pasta máxima* (**PEAK**): é o maior valor da viscosidade durante o ciclo de aquecimento, que é obtido no ponto máximo da curva,ou seja, quando a curva atinge 95°C apresentada na Figura (1.3).

- *Viscosidade de pasta mínima à quente* (**TROUGH**): é o menor valor da viscosidade durante os 5 minutos após atingir temperatura constante de 95°C.

- *Quebra de viscosidade* (**BREAKDOWN**): é a diferença entre a viscosidade de pasta máxima e a viscosidade de pasta mínima da curva viscoamilográfica.

- *Viscosidade final* (**FINAL**): é o valor final da viscosidade durante o ciclo de resfriamento, á temperatura de 25°C.

- *Tendência à retrogradação* (**SETBACK**): é a diferença entre a viscosidade final e a viscosidade de pasta mínima à quente.

1.5 Regressão Logística Clássica

Quando existe o interesse em descrever a influência que um conjunto de variáveis aleatórias tem sobre a ocorrência ou não de um determinado evento, a técnica mais utilizada é regressão logística (Hosmer & Lemeshow [6]).

Seja Y uma variável aleatória que indica a ocorrência ou não de um evento, tal que $Y = 0$ indica que o evento não ocorre e $Y = 1$ indica que o evento ocorre. Seja ainda um vetor $\mathbf{X} = (X_1, X_2, \cdots X_k)$ de variáveis aleatórias. Para modelos Binários, Y representa a variável reposta e $\mathbf{X}$ um vetor de variáveis explicativas.

A forma do modelo logístico é dada por:

$$\pi(\mathbf{X}) = \frac{\exp^{\boldsymbol{\beta}'\mathbf{x}}}{1 + \exp^{\boldsymbol{\beta}'\mathbf{x}}} \tag{1.1}$$

A partir da Equação (1.1) é obtida a função de ligação logito.

$$g(\mathbf{X}) = \ln\left[\frac{\pi(\mathbf{X})}{1 - \pi(\mathbf{X})}\right] = \boldsymbol{\beta}'\mathbf{X} \tag{1.2}$$

Essa transformação $g(\mathbf{X})$ garante muitas propriedades de um modelo linear. A função logito $g(\mathbf{X})$ é linear nos parâmetros, pode ser contínua e está variando em $\Re$.

Na regressão logística binária a variável resposta é dicotômica e pode ser escrita como: $Y = \pi(\mathbf{X}) + \varepsilon \Longrightarrow E(Y|\mathbf{X}) = \pi(X)$. Neste caso a variável Y assume apenas dois valores 0 ou 1 sucesso ou fracasso. Portanto em n ensaios a probabilidade de obter y sucessos do evento Y é dada por uma distribuição binomial de parâmetro π.

$$P(Y = y) = \frac{1!}{y!(1-y)!}\pi(y)^y(1 - \pi(y))^{1-y} \quad y = 1, \ldots, n. \tag{1.3}$$

Sendo $\pi(y)$ a probabilidade de ocorrer o evento, ou seja, a probabilidade de sucesso. Note que a quantidade $\varepsilon = y - \pi.$ pode assumir dois valores, ou seja,

$$\begin{cases} y = 1 \longrightarrow \varepsilon = 1 - \pi(x), \text{com probabilidade} \quad \pi \\ y = 0 \longrightarrow \varepsilon = -\pi(x), \text{com probabilidade} \quad 1 - \pi \end{cases}$$

e assim,

$$E(Y|\varepsilon) = 0 \quad e \quad var(Y|\varepsilon) = \pi(x)\left[1 - \pi(x)\right] \tag{1.4}$$

resultando numa distribuição Binomial com média 0 e variância $\pi(x)\left[1 - \pi(x)\right]$.

1.5.1 Estimação dos parâmetros e interpretação

Considere os pares (y_i, x_i), $i = 1 \dots n$, observações independentes, a função de verossimilhança para os dados $Y_i = 1, \dots, n$ é dada por :

$$L(\beta) = \prod_{i=1}^{n} [\pi(x_i)]^{y_i} \prod_{i=1}^{n} [1 - \pi(x_i)]^{1-y_i}, \tag{1.5}$$

Aplicando logaritmo na Equação (1.6):

$$\ell(\beta) = \sum_{i=1}^{n} \{y_i \ln[\pi(x_i)] + (1 - y_i) \ln[1 - \pi(x_i)]\}. \tag{1.6}$$

Logo, as estimativas de máxima verossimilhança dos parâmetros são obtidas pelos valores que maximizam a função dada pela Equação (1.6).

Os estimadores de máxima verossimilhança para os $\beta^{'}s$, são encontrados derivando com relação a β_0 e β_1 a Equação (1.6).

$$\frac{\partial \ell(\beta)}{\partial(\beta_0)} = \sum [y_i - \pi(x_i)] \tag{1.7}$$

$$\frac{\partial \ell(\beta)}{\partial(\beta_1)} = \sum x_i \, [y_i - \pi(x_i)] \tag{1.8}$$

Igualando as Equações (1.7) e (1.8) a zero, encontram-se equações que resultam nos estimadores de verossimilhança para os parâmetros. Observa-se que não existem formulas fechadas para estas equações, que são não-lineares e portanto para obter os estimadores são utilizados métodos iterativos como o de Newton-Raphson (Ruggiero [17]).

A interpretação dos parâmetros normalmente é dada por meio de uma medida denominada *odds ratio* ou razão de chances. Quando a variável é dicotômica, esta medida quantifica a chance de ocorrer o evento quando a variável resposta pertence ao grupo de referência. Mas neste trabalho o interesse maior é na probabilidade de ocorrência de cada categoria da variável resposta.

1.5.2 Testes de significância dos coeficientes

Assim, como nos modelos lineares, é necessário verificar quais variáveis são significativas, ou seja, quais variáveis possuem influência no evento de interesse, pois nem todas as covariáveis devem ser utilizadas para estimar a variável resposta. Então testa-se a significância dos coeficientes através dos testes da Razão de Verossimilhança e Wald (Hosmer e Lemeshow [6]).

- **Teste da Razão de Verossimilhança**

  O teste da razão de verossimilhança baseia-se na comparação entre valores observa-

dos com os valores preditos, definido por:

$$D = -2\ln\left[\frac{verossimilhança\ do\ modelo\ ajustado}{verossimilhança\ do\ modelo\ saturado}\right] \tag{1.9}$$

Em regressão logística binária, a verossimilhança para o modelo saturado (combinação de todos os parâmetros do modelo) é igual 1 e $\hat{\pi} = y$, como mostra a Equação (1.10)

$$L(modelo\ saturado) = \prod_{i=1}^{n} y_i^{y_i}(1-y_i)^{(1-y_i)} = 1 \tag{1.10}$$

Logo,

$$D = -2\ln[\ verossimilhança\ do\ modelo\ ajustado] \tag{1.11}$$

Então para testar a significância da variável independente deve-se comparar o valor da estatística D na Equação (1.11) com a variável e o modelo sem a variável independente, com isso tem-se a estatística G dada por:

$$G = D\,(modelo\ sem\ a\ variável) - D\,(modelo\ com\ a\ variável) \tag{1.12}$$

– **Hipóteses**

As hipóteses do teste da razão de verossimilhança são: $\begin{cases} H_0) & \beta_1 = 0 \\ H_1) & \beta_1 \neq 0 \end{cases}$

Sob a hipótese nula, a estatística G, dada na Equação (1.12) tem distribuição Qui-Quadrado com 1 grau de liberdade.

- **Teste de Wald**

A estatística do teste de Wald é dada por:

$$W = \left(\frac{\hat{\beta}_i}{\hat{SE}(\hat{\beta}_i)}\right)^2 \tag{1.13}$$

$\hat{SE}(\beta_i)$ é a estimativa do erro padrão do valor estimado do parâmetro β_i, $(\hat{\beta}_i)$. O teste de Wald é estatisticamente equivalente a estatística G para testar significância das variáveis. Logo, as hipóteses são equivalentes as descritas na Equação (1.12), mas em outras palavras podem ser definidas como:

$\begin{cases} \text{H}_0) \text{ A variável independente } x_i \text{ não possui influencia significativa no modelo ,} \\ \text{ mantidas todas as outras constantes.} \\ \text{H}_1) \text{A variável independente } x_i \text{ possui influencia significativa no modelo,} \\ \text{mantidas todas as outras constantes.} \end{cases}$

– **Regra de Decisão**

Considere $W = z^2$, onde a variável aleatória z possui distribuição aproximadamente normal padrão sob $\beta_i = 0$. Portanto z^2 segue uma distribuição aproximadamente qui-quadrado com 1 grau de liberdade. Nos dois teste descritos acima a regra de decisão, aceitar ou não a hipótese nula (H_0), deve-se calcular o p-valor da estatística do teste, equivalente a dizer que a probabilidade de obter um valor mais extremo do que a estatística do teste para a amostra observada, ou seja, z^2 no caso do teste de Wald,

$$\text{p-valor} = p(\chi_1^2 > z^2). \tag{1.14}$$

Dado um nível de significância α que é definido antecipadamente, pelo pesquisador, quando o (p-valor $> \alpha$), não há evidências estatísticas suficientes para rejeitar a hipótese nula, portanto β_1 é estatisticamente igual a zero, logo a variável independente x_i não deve ser considerada no modelo. Inversamente, se p-valor $< \alpha$, rejeita-se a hipótese nula e conclui-se que a covariável x_i é estatisticamente significante, logo deve ser considerada no modelo.

- **Intervalos de Confiança para os parâmetros**

Uma outra maneira de verificar a significância dos coeficientes das variáveis dos modelos é construindo intervalos de confiança para os parâmetros que são baseados no teste de Wald.

Para:

$$\beta_1 : \hat{\beta}_1 \pm z_{1-\alpha/2}\hat{SE}(\hat{\beta}_1) \quad e \quad \beta_0 : \hat{\beta}_0 \pm z_{1-\alpha/2}\hat{SE}(\hat{\beta}_0)$$

Na seção seguinte, será abordada a técnica de componentes principais, que é uma técnica de estatística multivariada muito utilizada para redução de dimensão em estudos em que o número de variáveis é grande. Dessa forma, nesse trabalho será utilizada essa técnica para reduzir a dimensão das covariáveis em estudo. Além disso, as componentes principais resultam em vetores de combinações lineares das variáveis originais não correlacionadas entre si, evitando o problema de multicolinearidade.

1.6 Análise de Componentes Principais

A análise de componentes principais (ACP), é uma técnica utilizada para explicar a estrutura de variância e covariância de um vetor com p-variáveis aleatórias por meio de combinações lineares das variáveis originais. Tais combinações lineares são chamadas de componentes principais e são não correlacionadas entre si (Johnson e Wichern [9]).

As combinações lineares resultam em um vetor $\mathbf{U}$ construído à partir de variáveis aleatórias não correlacionadas entre si que explica a estrutura de covariância do vetor $\mathbf{X}$, formado pelas variáveis originais (Mingoti [12]). O objetivo é diminuir o número de

variáveis que serão analisadas e as interpretações das combinações lineares construídas. Assim, a informação contida nas p-variáveis originais é substituída pela informação das k componentes principais não correlacionadas. Portanto, passa-se do espaço **p** para o **k**, onde **k**<**p**. Isto é, a dimensão, de um conjunto de dados, é reduzido pelas componentes principais, que retêm a maior parte da variação dos dados originais Wichern [9]).

Os vetores aleatórios **X** e **U**, possuem a mesma variância total, mas o vetor **U** tem a vantagem de ser formado por variáveis não correlacionadas, o que implica em maior facilidade de interpretação e de aplicação de algumas técnicas estatísticas (Mingoti [12]).

O número de componentes principais construído é igual ao número de variáveis aleatórias originais, ou seja, quando se tem p-variáveis originais pode-se obter p componentes principais (Mingoti[12]).

1.6.1 Componentes principais via matriz de covariância

As componentes principais podem ser estimadas por meio da matriz de covariâncias, $\sum_{p\times p}$, que na prática é desconhecida, e portanto, estimada através dos dados amostrais pela matriz de covariâncias amostral, $\mathbf{S}_{p\times p}$.

Suponha um conjunto de dados com **p** colunas e **m** linhas, onde as colunas representam as variáveis e as linhas são referentes à uma mesma observação. Essa estrutura pode ser vista como uma matriz com **p** colunas e **m** linhas, sendo **p** o número de variáveis e **m** o número de observações de cada variável. Onde X_j indica a j-ésima variável, para $j = 1, \ldots, p$ [9, p. 5].

Considere a matriz X que representa o conjunto de dados

$$X_{m\times p} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{m1} & x_{m2} & \cdots & x_{mp} \end{bmatrix} = [\ X_1 \quad X_2 \quad \cdots \quad X_p\ ]. \tag{1.15}$$

Suponha que $[\ X_1 \quad X_2 \quad \cdots \quad X_p\ ]$ seja um vetor aleatório com vetor média determinado por :

$$\bar{\boldsymbol{x}} = \begin{bmatrix} \bar{x}_1 \\ \bar{x}_2 \\ \vdots \\ \bar{x}_p \end{bmatrix}, \tag{1.16}$$

onde

$$\bar{X}_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij}, \tag{1.17}$$

para $j = 1, \ldots, p$ [9, p. 6:7].

Com base nesse vetor de médias obtêm-se a matriz de covariâncias amostrais, $\mathbf{S}_{p\times p}$

que é uma medida de dispersão importante pois indica o grau dependência linear entre duas variáveis ou a variabilidade conjunta.

$$S_{p\times p} = \begin{bmatrix} s_{11} & s_{12} & \cdots & s_{1p} \\ s_{21} & s_{22} & \cdots & s_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ s_{m1} & s_{m2} & \cdots & s_{mp} \end{bmatrix}. \tag{1.18}$$

Para $i \neq k$,

$$s_{ik} = s_{ki} = \frac{1}{n}\sum_{i=1}^{n}(X_{ik} - \bar{X}_i)(X_{jk} - \bar{X}_k) \quad i = 1, 2, \cdots, m, \quad k = 1, 2 \cdots, p. \tag{1.19}$$

Para o caso onde i=k, tem-se as variâncias.

$$s_{ik} = s_i^2 = s_k^2 = \frac{1}{n}\sum_{i=1}^{n}(X_{ij} - \bar{X}_i)^2 = \frac{1}{n}\sum_{i=1}^{n}(X_{ij} - \bar{X}_k)^2, \tag{1.20}$$

[9, p. 7:8].

Os autovalores da matriz $S_{p\times p}$ estão disposto de forma que $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p$ e os autovetores, ão dados por $e_i, i = 1, 2 \cdots p$. Visto isso, pode-se construir as combinações lineares (U_i) das variáveis X_i, para $i = 1, ..., p$, de forma que as componentes principais sejam definidas por:

$$U_1 = e_1{}'X = e_{11}X_1 + e_{12}X_2 + \cdots + e_{1p}X_p, \tag{1.21}$$

$$U_2 = e_2{}'X = e_{21}X_1 + e_{22}X_2 + \cdots + e_{2p}X_p, \tag{1.22}$$

$$\vdots \tag{1.23}$$

$$U_p = e_p{}'X = e_{p1}X_1 + e_{p2}X_2 + \cdots + e_{pp}X_p. \tag{1.24}$$

[9, p. 431].

As CPs U_i podem ser representadas por qualquer combinação linear de forma que valores e_i's maximizem as variâncias e covariâncias dessas combinações lineares, expressas respectivamente por:

$$Var(U_i) = e_i{}'Se_i, i = 1, 2 \cdots, p; \tag{1.25}$$

$$Cov(U_i, U_k) = e_i{}'Se_k, i, k = 1, 2 \cdots, p. \tag{1.26}$$

[9, p. 431].

Algumas propriedades das componentes principais são descritas a seguir:

1. A variância estimada de $\hat{U}_i$ é dada por $\sum_{i=1}^{p} \hat{\lambda}_i$, $i = 1, 2 \cdots, p$;

2. A $Cov(\hat{U}_i, \hat{U}_k) = 0$, para $i \neq k$. Portanto, as componentes principais são não correlacionadas;

3. A proporção total da variância explicada pela k-ésima componente é dada por $\frac{\lambda_k}{\sum_{i=1}^{p} \lambda_i}$, $i = 1, 2 \cdots p$. Ressalta-se que a primeira componente principal é a que possui a maior proporção de explicação da variância original, seguida pela segunda componente principal, que explica a segunda maior parte da variância restante, e assim sucessivamente.

4. A correlação estimada entre as componentes e as variáveis originais é dada por: $\hat{\rho}_{U_i, X_k} = \frac{e_{ik}\sqrt{\hat{i}}}{\sqrt{s_{kk}}}$.

1.6.2 Componentes principais via matriz de correlação

Destaca-se que as componentes principais também são estimadas pela matriz de correlação em casos onde as variâncias das variáveis originais são muito discrepantes. Tal situação ocorre, na maioria das vezes, devido às diferenças entre as unidades de medidas das variáveis originais (Mingoti [12]).

De acordo com Johnson e Wichern [9], pode-se diminuir esse problema realizando uma transformação nas variáveis originais para deixá-las na mesma escala. A padronização mais utilizada envolve a subtração pela média e divisão pelo desvio padrão. Como à seguir:

$$Z_i = \frac{(X_i - \mu)}{\sqrt{\sigma_{ij}^2}}.$$

Normalmente, a matriz de correlação ρ é desconhecida. Utiliza-se, então a matriz de correlação amostral, $R_{p \times p}$, que é calculada dividindo-se cada s_{ik} pelos desvios padrões s_i e s_k.

$$R_{p \times p} = \begin{bmatrix} 1 & r_{12} & \cdots & r_{1p} \\ r_{21} & 1 & \cdots & r_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ r_{m1} & r_{m2} & \cdots & 1 \end{bmatrix}. \tag{1.27}$$

Para $i \neq k$, tem-se que

$$r_{ik} = \frac{s_{ik}}{s_i^2 s_k^2}. \tag{1.28}$$

Já para o caso em que $i = k$,

$$r_{ik} = \frac{s_i^2}{s_i^2} = \frac{s_k^2}{s_k^2} = 1, [9, p.\ 8]. \tag{1.29}$$

A propriedade de simetria da correlação afirma que:

$$r_{ik} = r_{ki} \tag{1.30}$$

Considerando a matriz $R_{p\times p}$ dos dados $X = [\ X_1 \quad X_2 \quad \cdots \quad X_m\ ]$, o seus respectivos autovalores estão organizados de forma que $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p$ e os autovetores correspondentes dados por $e_1, e_2 \cdots, e_p$. Portanto, a i-ésima componente principal é dada por:

$$U_i = e_i' Z = e_{i1}Z_1 + e_{i2}Z_2 \cdots + e_{ip}Z_p.$$

Abaixo, cita-se algumas propriedades:

1. A variância estimada de $\hat{U}_i$ é dada por $\hat{\lambda}$, $i = 1, 2 \cdots, p$;
2. A $Cov(\hat{U}_i, \hat{U}_k) = 0$, para $i \neq k$, o que implica que as componentes principais não são correlacionadas;
3. A proporção total da variância explicada pela i-ésima componente é dada por: $\frac{\hat{\lambda}_i}{p}, i = 1, 2 \cdots p$, em que p é o número de variáveis originais;
4. A correlação entre U_i e Z_i é dada por: $r_{U_i,Z_i} = e_{ik}\sqrt{\lambda_i}$.

Ressalta-se que os coeficientes das componentes principais estimadas pela matriz de correlação não são numericamente iguais às estimadas pela matriz de covariância. Em geral, é necessário utilizar um número maior de componentes para se explicar a mesma quantidade da variabilidade total obtida quando se usa a matriz de $\sigma_{p\times p}$ (Mingoti [12]).

Na Seção 1.7 será abordada a técnica de análise discriminante, utilizada para a classificação de grupos de elementos. Neste trabalho a regra matemática para a discriminação dos grupos será fornecida pela regressão logística.

1.7 Regressão logística como função Discriminante

A análise discriminante é uma técnica estatística multivariada usada para a classificação de grupos de elementos de uma amostra ou população. Os grupos em que será classificado cada elemento amostral, ou populacional, devem ser predefinidos. Esse conhecimento *a priori* sobre os grupos permite a obtenção de uma função matemática denominada regra de classificação ou discriminação (Mingoti [12]). A regra de classificação é baseada na teoria de probabilidades.

A análise de regressão logística, portanto, também pode ser utilizada como função de classificação ou discriminação, pois por meio deste modelo é possível se estimar as probabilidade de um elemento pertencer a uma determinada categoria. As categorias são os grupos predefinidos. A regra de classificação é determinada pelos valores preditos, definidos a partir do modelo logístico e determinados por $P(Y_i = k|x_{i1}, x_{12} \cdots, x_{ip})$.

Os valores preditos são utilizados para a separação dos grupos, permitindo a estimação da probabilidade de classificação incorreta de uma determinada observação pertencer a uma determinada classe, ou das probabilidades de classificação incorretas de cada observação pertencer à cada classe.

Suponha que se deseja estimar a probabilidade de classificação incorreta da avaliação de textura do arroz quanto a pegajosidade, sendo atribuído "1" para solto e,"0" para pegajoso. Nesse caso, com apenas duas categorias, seria usado o modelo de regressão logística binário. Portanto, cada observação tem dois valores preditos, sendo um para cada categoria. Quando o modelo logístico binário é utilizado é preciso se determinar um ponto de corte, um valor **c** entre 0 e 1 tal que, quando valores preditos são menores que **c**, deve-se classificar como "Pegajoso" e, acima, como "Solto". Geralmente o valor **c** é 0,5 , mas não existe uma regra. Esse valor pode variar de acordo com algum conhecimento que se tenha *a priori*.

Quanto ao modelo politômico tem-se k categorias possíveis, sendo k maior que 2. Por exemplo, considere a variável Pegajosidade com 5 níveis, ou 5 categorias, ou seja, $k = 5$. Nessa situação, tem-se 5 valores preditos para cada observação, um para cada categoria. Dessa forma, não é necessário determinar um ponto de corte. Portanto, classifica-se cada observação como pertencente a uma determinada categoria cuja probabilidade é maior.

Ressalta-se que muitas vezes a verificação com precisão da escolha da categoria não é simples. Por exemplo, quando, para alguns valores da variável explicativa em que as 2 curvas se cruzam, a diferença entre os valores preditos é muito pequena. Desse modo, deve-se utilizar métodos que avaliem a qualidade da regra de discriminação (Rios [16]).

De acordo com Mingoti [12], um dos métodos consiste em se calcular a taxa de acerto de classificação pelo método de validação cruzada. Existem dois erros associados à tal método descritos a seguir:

- Erro 1: a observação pertence à categoria $k = 1$, mas a regra de classificação o classifica como da categoria $k = 2$;
- Erro 2: a observação pertence à categoria $k = 2$, mas a regra de classificação o classifica como pertencente à categoria $k = 1$.

Segue abaixo a representação dos erros de classificação:

Tabela 1.1: Classificação observada versus classificação prevista através dos modelos logísticos

		Classificação prevista	
	k	$y = 1$	$y = 0$
Classificação real	$y = 1$	n_{1c}	n_{1m}
	$y = 0$	n_{2m}	n_{2c}
		n_1	n_2

A Tabela 1.1 é denominada matriz de confusão. Portanto, considerando $k = 2$, segue que n_{1c} representa o número de observações da categoria 1 que foram classificadas corretamente na categoria 1; n_{1m} é o número de itens da categoria 1 que foram classificadas em $k = 2$; n_{2c} é o número de observações da categoria 2 que foram classificadas corretamente e n_{2m} é o número de observações da categoria dois que foram classificados incorretamente (Johnson e Wichern [9]).

Com base nisso, calcula-se a taxa de erro aparente (classificação) dada por:

$$APER : \frac{n_{1c} + n_{2c}}{n_1 + n_2}.$$

A $APER$ representa a proporção de observações classificadas erroneamente. O processo para o cálculo desta desta taxa é feito por validação cruzada. Esse método consiste em separar o total $= n_1 + n_2$ em duas amostras, uma de treinamento, que será usada na construção do modelo preditivo, e outra de validação, usada para a avaliação da qualidade da predição. Os três passos do método, baseados em Mingoti [12], estão descritos abaixo:

1. Primeiramente retira-se uma observação do total e utiliza-se as $n_1 + n_2 - 1$ observações totais restantes para se construir o modelo logístico preditivo, ou seja, a regra de discriminação;

2. Utiliza-se a regra de discriminação do passo 1 para a classificação das observações restantes, ou seja, a amostra de validação. Com isso é possível averiguar se a regra de classificação conseguiu discriminar corretamente as observações nas suas categorias de origem.

3. Em seguida recoloca-se a observação que foi retirada no passo 1 na amostra de treinamento e retira-se uma observação diferente. Os passos 1 e 2 são repetidos até que todas $n_1 + n_2$ observações sejam classificadas.

1.8 Regressão Logística Bayesiana

Modelos clássicos, como regressão binomial são baseados em inferências para grandes amostras (teoria assintótica). Os modelos Bayesianos permitem inferências mesmo quando as amostras são pequenas, e podem ser aplicados por meio de aproximações e técnicas computacionais simples. Os problemas que dificultam a análise Bayesiana vêm sendo resolvidos por métodos Monte Carlo via Cadeias de Markov-$MCMC$ (Mendonça [11]).

Dentro dos modelos lineares generalizados, o mais utilizado é o modelo para dados binários ou binomial (Gelman [5]).

O paradigma Bayesiano tem a estrutura descrita a seguir. Considere por exemplo, que o estudo de interesse seja estimar θ com base nos dados $Y = [y_1, y_2, \cdots y_n]$ através de um modelo estatístico descrito pela densidade $p(\theta)$. A filosofia Bayesiana afirma que θ não é

fixo, e o grau de incerteza sobre este parâmetro é declarado por meio de uma distribuição de probabilidade (Kinas [10]).

Os elementos essênciais da inferência Bayesiana são:

- A distribuição de probabilidade para θ é formulada como $\pi(\theta)$, que é conhecido como a distribuição *a priori*, ou simplesmente *priori*. A distribuição *a priori* é um conhecimento que se tem externo ao conjunto de dados, com respeito ao parâmetro.
- Considerando Y como o vetor de dados observados, é escolhido um modelo estatístico, para $p(Y|\theta)$, com a finalidade de se descrever a distribuição de Y dado θ.
- Atualiza-se a distribuição *a priori*, combinando a informação da distribuição *a priori* e os dados por meio do cálculo da distribuição *a posteriori*. Isto é feito levando-se em consideração o Teorema de Bayes do qual se deriva o nome deste ramo da estatística.

Suponha que Y seja uma variável aleatória discreta binomial, $y_i \sim Bin(n_i, \theta)$, com $n_i = 1$ para todo i, e suponha:

$$\eta_i = g(\theta_i) = \ln\left(\frac{\theta i}{1-\theta_i}\right) \qquad (1.31)$$
$$\eta_i = \beta_0 + \beta_1 x_i.$$

A função de ligação logito definida na Equação (1.31) transforma o parâmetro θ_i restrito em $[0,1]$, sendo η_i definida em $(-\infty, +\infty)$. A distribuição para os dados y e a função de log verossimilhança são respectivamente dadas por:

$$p(y|\boldsymbol{\beta}) = \prod_{i=1}^{n} \begin{pmatrix} n_i \\ y_i \end{pmatrix} \left(\frac{\exp^{\eta_i}}{1+\exp^{\eta_i}}\right)^{y_i} \left(\frac{1}{1+\exp^{\eta_i}}\right)^{\eta_i - y_i} e \qquad (1.32)$$

$$\ell(\boldsymbol{\beta}|Y) = \sum_{i=1}^{n} \left[y_i \log\left(\frac{\exp^{\eta_i}}{1+\exp^{\eta_i}}\right) + (\eta_i - y_i)\log\left(\frac{1}{1+\exp^{\eta_i}}\right)\right] = \sum_{i=1}^{n} \left[y_i\eta_i - \eta_i \log\left(1+\exp^{\eta_i}\right)\right]. \qquad (1.33)$$

Dada uma distribuição *a priori* para β, $p(\beta)$, é encontrada a distribuição *a posteriori*:

$$p(\boldsymbol{\beta}|Y) = \frac{L(\boldsymbol{\beta}|Y)p(\boldsymbol{\beta})}{\int L(\boldsymbol{\beta}|Y)p(\boldsymbol{\beta})d\boldsymbol{\beta}}. \qquad (1.34)$$

A distribuição *a posteriori* é atualização da priori, condicionada a todos os dados do estudo. O cálculo da primeira normalmente envolve a resolução de integrais numéricas. Através do Método de Monte Carlo via cadeias de *Markov* são obtidas aproximações para as integrais numéricas por meio de amostras da distribuição *a posteriori*.

A ideia geral por trás do método *MCMC* envolve a transformação do problema estático de interesse em um problema dinâmico. Isso é realizado construindo um processo

estocástico artificial fácil de simular e que possa convergir para distribuição de interesse, nesse caso, para a distribuição a posteriori. Dessa forma, utiliza-se técnicas de simulação baseadas em cadeias de Markov, resultando em valores gerados não independentes (Paulino et.al [15]).

Existem muitas maneiras de se encontrar amostras *a posteriori* utilizando o método $MCMC$. Dois algoritmos muito utilizados são Metropolis-Hasting e o amostrador de Gibbs (Mendonça [11]). Porém, será utilizado apenas o algoritmo de Metropolis-Hasting, que é uma classe de algoritmos de MCMC cuja a ideia é gerar uma cadeia de Markov com distribuição estacionária. No algoritmo geramos pontos de uma distribuição proposta. Se o ponto gerado é aceito, a cadeia move para um estado posterior. Se o ponto não é aceito a cadeia permanece no estado atual. Esse critério garante a convergência da cadeia para a distribuição de equilíbrio, que é *a posteriori* (Mendonça [11]).

Quando a convergência da cadeia não é alcançada, pode-se alterar alguns procedimentos que envolvem a geração da cadeia, tais como:

- Aquecimento: é a fase antes da convergência, onde os valores gerados não tem um comportamento estacionário e estão altamente correlacionados. Contudo, um aquecimento de tamanho 1000 significa que os 1000 primeiros valores gerados da amostra a partir de um ponto inicial antes da convergência serão descartados da cadeia. Isto ajuda na convergência do algoritmo.

- Salto: é o número de iterações dada para que um valor seja selecionado para amostra depois do período de aquecimento, ou seja, um salto de tamanho 30 implica que será selecionada a observação gerada após 30 iterações para a amostra depois do período de aquecimento.

Existem várias maneiras de verificar a convergência da cadeia gerada para distribuição de interesse. (Paulino et.al [15]). Aqui serão utilizados apenas dois gráficos para avaliar a convergência da cadeia de $Markov$:

- O primeiro, mostra o número de iterações $MCMC$ excluindo as iterações de aquecimento versus os valores das variáveis aleatórias ($\beta(s)$). Espera-se que os valores estejam distribuídos em torno de um certo ponto, ou seja, deve-se observar a estacionariedade.

- o segundo gráfico, representa as autocorrelações versus *lags*, significa, avaliar as correlações entre os valores gerados na iteração t e na iteração t+1. Nesse caso o ideal é verificar um decaimento rápido nas autocorrelações dos valores gerados pela cadeia.

Na regressão logística Bayesiana os parâmetros são variáveis aleatórias e são descritos através da distribuição *a posteriori*.

O principal objetivo da posteriori é se conseguir uma amostra da distribuição de probabilidade associada a cada um dos parâmetros de interesse. Esses parâmetros podem ser estimados por meio de medidas descritivas da distribuição *a posteriori*, como por exemplo a média, a mediana e os quartis.

O intervalo de credibilidade possui interpretação direta sobre os parâmetros, ou seja, o parâmetro tem $(1-\alpha)$ de probabilidade de estar entre os limites inferior e superior do intervalo. Quanto menor for o comprimento do intervalo, maior será a concentração em torno do parâmetro estimado. No caso Bayesiano também, quando o zero está dentro do intervalo de credibilidade, há um indício de que a variável não é significativa no modelo (Mendonça [11]).

O método de $MCMC$ está implementado nos softwares estatísticos R e SAS, facilitando assim, o uso de modelos Bayesianos, visto que o tempo computacional não é mais um problema em modelos relativamente simples. O software SAS utilizado nesse estudo apresenta as seguintes PROCS:

- $MCMC$: utiliza como *default* o algoritmo de Metropolis-Hasting para geração da cadeia. O modelo logístico pode ser escrito na opção *model* onde escolhe-se a função de ligação e a distribuição. As prioris são inseridas pela opção *prior*. Para o modelo logístico binário, pode-se escolher a distribuição como *binomial* ou *bernouli* dependo de como estão os dados. Uma descrição das opções utilizadas é encontrada no apêndice.

- $GENMOD$: também pode ser usada para modelagem Bayesiana por meio da opção *bayes* e utiliza como *default* para geração da cadeia o algoritmo amostrador de Gibbs. Para mais detalhes ver Manual do SAS [19].

Capítulo 2

Resultados e Discussão

2.1 Análise Descritiva

As informações sobre a análise da textura do arroz e cultivo para o ano de 2013 são apresentadas por 189 observações. Observa-se na Tabela 2.1, que os valores referentes à classificação sensorial da variável dureza estão muito concentrados em uma única categoria ("Macio") correspondente à 73,54% das observações. A categoria "Extremamente macio" possui a segunda maior frequência, porém, representa apenas 15,87% dos dados. As demais categorias referentes a análise da textura sensorial da dureza apontam frequências bem inexpressivas, sendo que as categorias "Muito firme" e "Extremamente Firme" nem foram analisadas.

Tabela 2.1: Classificação sensorial de dureza

Dureza	Frequência	Porcentagem
Extremamente firme	0	0%
Muito firme	0	0%
Levemente firme	1	0,5%
Macio com centro firme	14	7,4%
Ligeiramente macio	5	2,7%
Macio	139	73,4%
Extremamente macio	30	16%

Na Tabela 2.2 estão representados os valores referentes à classificação sensorial da pegajosidade. Observa-se que existe uma melhor distribuição dos dados entre as categorias para a variável pegajosidade do que para a variável dureza. A categoria "Levemente solto" obteve a maior representatividade com 38% das observações, seguida pela categoria "Pegajoso" com 33%, pela "Muito pegajoso" com 23%, e a categoria "Solto" com apenas 6,4% dos dados. Apesar da classificação sensorial da pegajosidade estar melhor distri-

buída, nenhuma observação foi classificada como "Extremamente solto", "Muito solto" ou "Extremamente pegajoso".

Tabela 2.2: Classificação sensorial de pegajosidade

Pegajosidade	Frequência	Pegajosidade
Extremamente pegajoso	0	0%
Muito pegajoso	43	23%
Pegajoso	63	33%
Levemente solto	71	38%
Solto	12	6,4%
Muito solto	0	0%
Extremamente solto	0	0%

Na Figura 2.1, verifica-se a classificação sensorial da dureza de acordo com o tipo de terreno. Nota-se que, para arroz de terrenos irrigados, a classificação sensorial está bem mais concentrada que para arroz de terras altas.

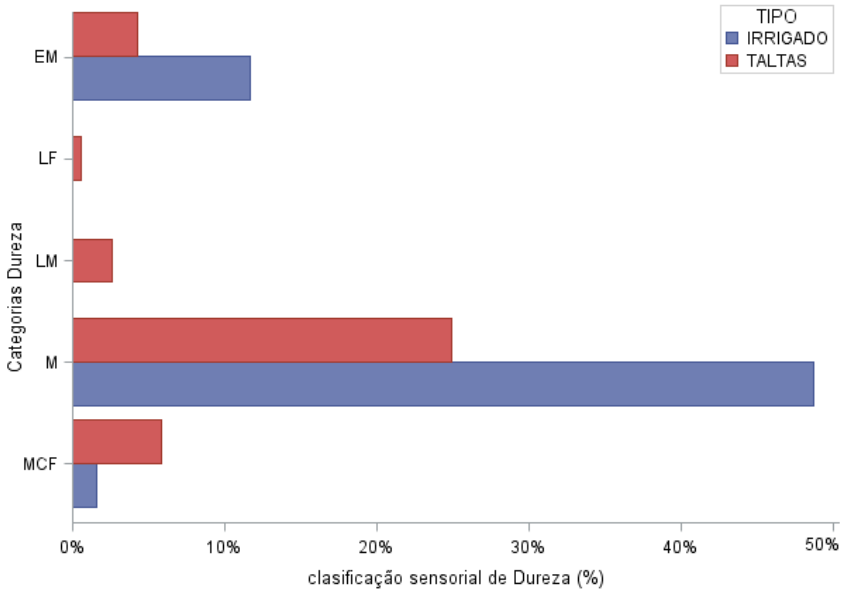

Figura 2.1: Classificação sensorial da dureza sensorial segundo o tipo de terreno

[1] A Figura 2.1 apresenta quatro categorias: EM (EXTREMAMENTE MACIO), LF (LEVEMENTE FIRME), LM (LIGEIRAMENTE MACIO), M (MACIO), MCF (MACIO com CENTRO FIRME). Além disso, na legenda do gráfico, Tipo refere-se ao Tipo de Cultivar; IRRIGADO refere-se a Terreno Irrigado; e TALTAS diz respeito à Terras Altas.

A Figura 2.2 mostra a classificação sensorial para a variável pegajosidade. Verifica-se que uma melhor distribuição dos dados entre as categorias para terrenos irrigados do que para terras altas. Ressalta que para terrenos irrigados a categoria "pegajoso" apresenta a maior frequência enquanto que para terras altas é a categoria "Levemente Solto".

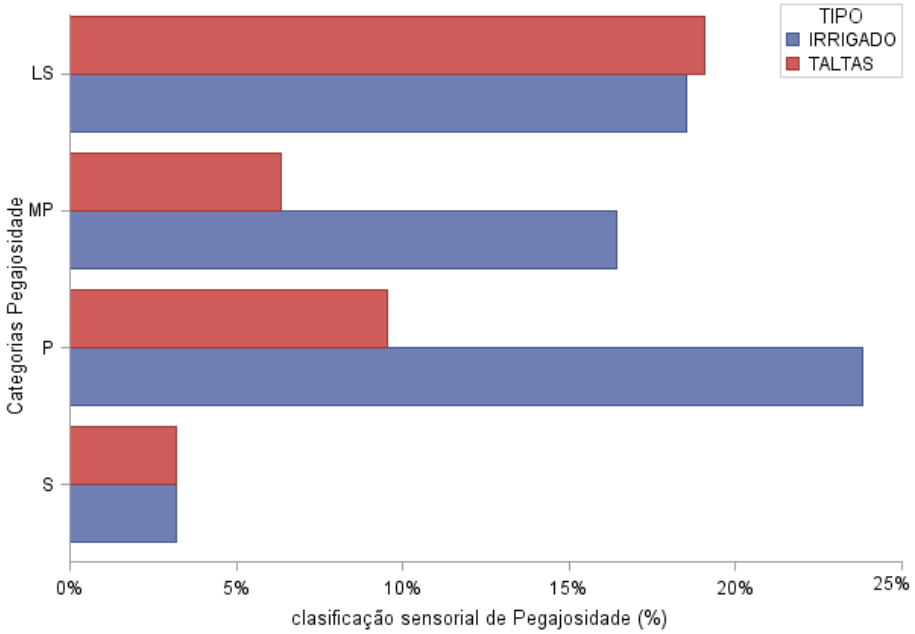

Figura 2.2: Classificação sensorial da pegajosidade sensorial segundo o tipo de terreno

[2] A Figura 2.2 apresenta quatro categorias: LS (LEVEMENTE SOLTO), MP (MUITO PEGAJOSO), P (PEGAJOSO), S (SOLTO). Além disso, na legenda do gráfico, Tipo refere-se ao Tipo de Cultivar; IRRIGADO refere-se a Terreno Irrigado; e TALTAS diz respeito à Terras Altas.

A Tabela 2.3 apresenta algumas medidas descritivas para as variáveis quantitativas referentes às medidas de textura instrumental ("PEGAJT" e "DURAZAT") e as demais variáveis representam as medidas instrumentais de viscosidade (Perfil viscoamilográfico). Ressalta-se que a variável PEGAJAT apresenta alta variabilidade em torno da média (coeficiente de variação -60%), e a variável DUREZAT contém valores mais concentrados em torno da média, com apenas 15% de dispersão. Essas duas variáveis possuem assimetria forte e negativa, ou seja, os dados estão mais deslocados para esquerda. Os coeficientes de assimetria são respectivamente $-0,7$ e $-0,6$.

Tabela 2.3: Medidas resumo das variáveis quantitativas

	Média	Variância	Coef.Var	1° Quartil	Mediana	3° Quartil	Coef. AS
TAAFIA	15,1	52,1	0,47	11,4	19	20,3	-1,1
TAASEC	16,0	67,2	0,51	10,3	19,9	22,8	-0,8
TG	4,7	2,6	0,34	3,0	4,4	6,4	0,2
PEAK	230,2	5538,1	0,32	179,1	247,6	289,1	-0,5
BREAKDOWN	90,3	240,6	0,52	53,6	80,2	127,8	0,4
FINAL	280,5	13486,9	0,41	209,5	264	385,9	<0,01
SETBACK	140,7	4961,7	0,50	99,8	131,7	197,6	<0,03
DUREZAT	140,9	453,8	0,15	125,8	143	156,7	-0,6
PEGAJT	-9,2	30,6	-0,60	-13,0	-7,3	-4,5	-0,7

As variáveis referentes às medidas instrumentais de viscosidade apresentam variabilidades altas, principalmente para as variáveis PEAK ($var = 5538,1$), BREAKDOWN (240,6) e FINAL (13486,9). A variável BREAKDOWN apresenta a maior dispersão em torno da média (cv =52%). Seguida pela variável FINAL ($cv = 41\%$) e por PEAK ($cv = 0,32\%$). Os coeficientes de assimetria para as variáveis PEAK, BREAKDOWN e FINAL que são respectivamente ($AS = -0,5$), ($AS = 0,4$) e ($AS = 0,029$). Portanto, nota-se que as variáveis PEAK e BREAKDOW possuem assimetria moderada, à esquerda e à direita, respectivamente. As variáveis que possuem assimetria forte com caudas mais pesada à esquerda são respectivamente TAAFIA ($AS = -1,1$), TAASEC ($AS = -0,8$), PEGAJT ($AS = -0,7$) e DUREZAT ($AS = -0,6.$). Ressalta-se que DUREZAT e PEGJAT refere-se as medidas instrumentais de textura.

A Tabela 2.4 apresenta o teste de hipótese para o coeficiente de correlação spearman, em que H_0: $\rho = 0$ (não existe correlação linear entre as variáveis) e $\rho \neq 0$ (existe correlação linear entre as variáveis). Considerando $\alpha = 0,01$, verifica-se que existem evidências para se rejeitar a hipótese nula para algumas variáveis referentes às medidas de viscosidade. Portanto existem evidências estatísticas de que não existe correlação linear entre a variável PEAK e as variáveis TAAFIA (p-valor = $0,0124$), TAASEC (p-valor = $0,0488$), ou as medidas instrumentais DUREZAT (p-valor = $0,2189$) e PEGAJT (p-valor = $0,0956$). Esse fato se repete com variável BREAKDOWN e as variáveis FINAL (p-valor = $0,2321$), SETBACK (p-valor = $0,6784$) e a DUREZAT (p-valor = $0,0331$).

Tabela 2.4: Matriz de correlação entre as varíaveis quantitativas seguido do seu p-valor para a hipótese nula $\rho = 0$

	TAAFIA	TAASEC	TG	PEAK	BREAKDOWN	FINAL	SETBACK	DUREZAT	PEGAJT
TAAFIA	1								
TAASEC	0,97 **<0,0001**	1							
TG	0,389 **<0,0001**	0,47 **<0,0001**	1						
PEAK	0,18 **0,01**	0,14 **0,05**	-0,29 **<0,0001**	1					
BREAKDOWN	-0,2562 **0,0003**	-0,31 **<0,0001**	-0,64 **<0,0001**	0,74 **<0,0001**	1				
FINAL	0,69 **<0,0001**	0,69 **<0,0001**	0,22 **0,0023**	0,70 **<0,0001**	0,09 **0,2321**	1			
SETBACK	0,77 **<0,0001**	0,77 **<0,0001**	0,24 **0,0007**	0,59 **<0,0001**	0,03 **0,6784**	0,97 **<0,0001**	1		
DUREZAT	0,52 **<0,0001**	0,53 **<0,0001**	0,20 **0,0063**	0,09 **0,22**	-0,07 **0,03**	0,31 **<0,0001**	0,37 **<0,0001**	1	
PEGAJT	0,90 **<0,0001**	0,91 **<0,0001**	0,44 **<0,0001**	0,12 **0,0956**	-0,37 **<0,0001**	0,69 **<0,0001**	0,76 **<0,0001**	0,38 **<0,0001**	1

O coeficiente de correlação de Spearman varia entre -1 e 1. Valores mais próximos de 1 indicam correlação positiva e forte, e valores próximos de -1 indicam correlação negativa e forte. Com base na Figura 2.3 e na Tabela 2.4, observa-se que as medidas instrumentais da textura do arroz referentes à avaliação pegajosidade e da dureza estão correlacionadas de maneira fraca e positiva,($r = 0,38$), indicando que quando aumenta-se a dureza instrumental aumenta-se também levemente a pegajosidade instrumental. Nas medidas instrumentais da textura, a variável PEGAJAT (Pegajosidade) está correlacionada de maneira positiva e muito forte com as medidas instrumentais de viscosidade TAAFIA

($r = 0,90$), TAASEC ($r = 0,91$),SETBACK ($r = 0,76$). Já a variável DUREZAT (Avaliação instrumental da dureza) está correlacionada com as variáveis TAAFIA ($r = 0,52$), TAASEC ($r = 0,53$), SETBACK($r = 0,37$)de maneira positiva. Sendo a correlação moderada e fraca respectivamente. As correlações mais representativas entre as medidas de viscosidades são apresentadas pelas variáveis TAAFIA, TAASEC, FINAL, SETBACK, BREAKDOW e TG. As variáveis TAAFIA e TAASEC são altamente correlacionadas positivamente ($r = 0,97$). Ocorre o mesmo para as variáveis FINAL e SETBACK ($r = 0,97$). A variável SETBACK apresenta correlação forte e positiva quanto às variáveis TAAFIA ($r = 0,77$) e TAASEC ($r = 0,77$). Já as variáveis TG e BREAKDOWN apresentam correlação negativa e forte ($r = -0,64$).

Figura 2.3: Matriz de dispersão das medidas instrumentais de viscosidade e instrumentais de textura

É possível dizer, analisando-se a Figura 2.3 e Tabela 2.4, que muitas variáveis referente às medidas instrumentais de viscosidades estão bastante correlacionadas. O que pode ajuste dos modelos.

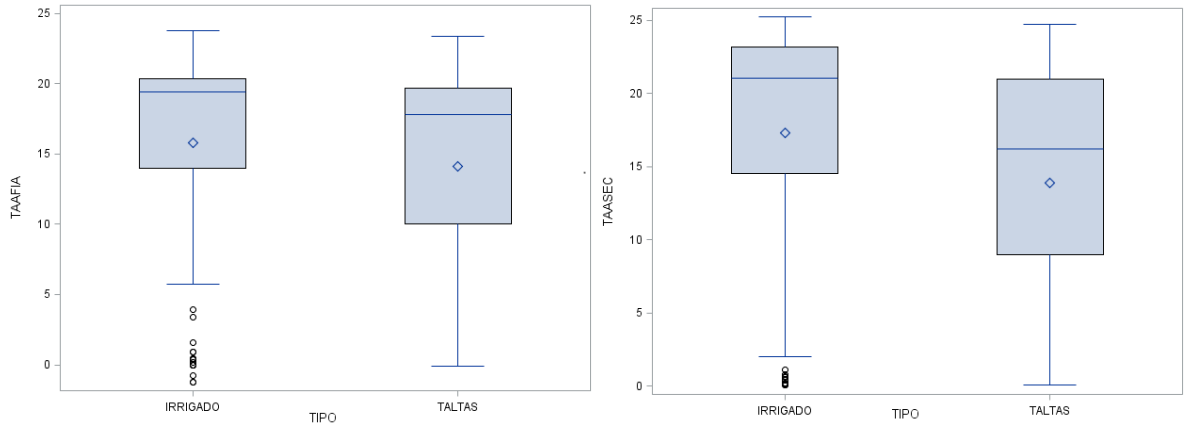

Figura 2.4: Medida instrumentais de viscosidade: variáveis TAAFIA e TAASEC

Observa-se na Figura 2.4 que a variável TAAFIA apresenta muitos valores discrepantes e baixos para arroz de cultivares irrigados e valores mais concentrados para terras altas. Enquanto a variável TAASEC apresenta maior dispersão dos valores para ambos os tipos de cultivares. Porém, comportamento dessas variáveis segundo o tipo se cultivar de arroz são semelhantes.

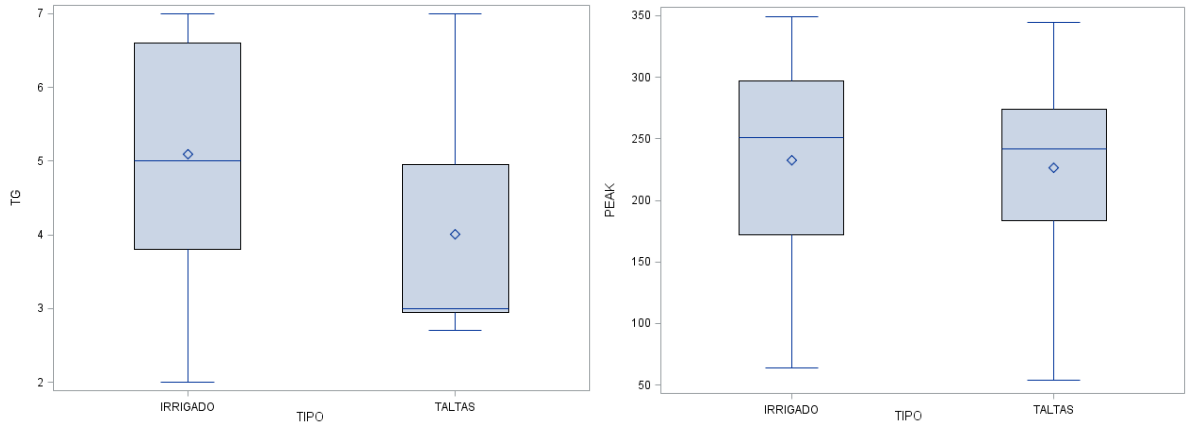

Figura 2.5: Medidas instrumentais de viscosidade: variáveis TG e SPEAK

Com base na Figura 2.5 verifica-se que a variável TG apresenta uma grande dispersão para os dois cultivares, porém os valores para cultivares irrigados são bem menores do que para terras altas.

[1]

[1]Nos gráficos dessa seção no eixo das abscissas, Tipo refere-se a TIPO de Cultivar, IRRIGADO refere-se a Terrenos Irrigados e TALTAS refere-se a Terras Altas.

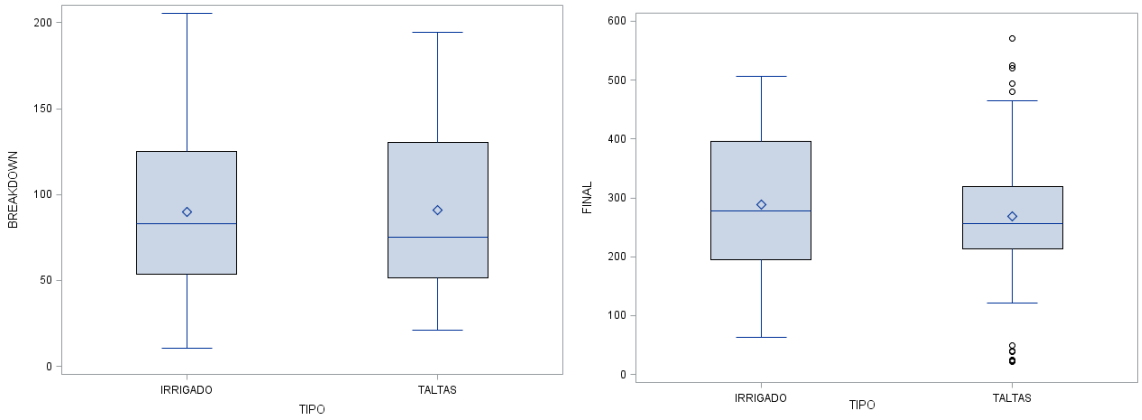

Figura 2.6: Medidas instrumentais de viscosidade: variáveis BREAKDOWN e FINAL

A Figura 2.6 mostra que as medidas de viscosidade BREAKDOWN e FINAL estão com valores concentrados em torno de um mesmo valor. Para terrenos irrigados, estas variáveis possuem uma maior variabilidade, enquanto que, para cultivar de terras altas, a variável FINAL possui pouca variabilidade, mas com alguns pontos discrepantes.

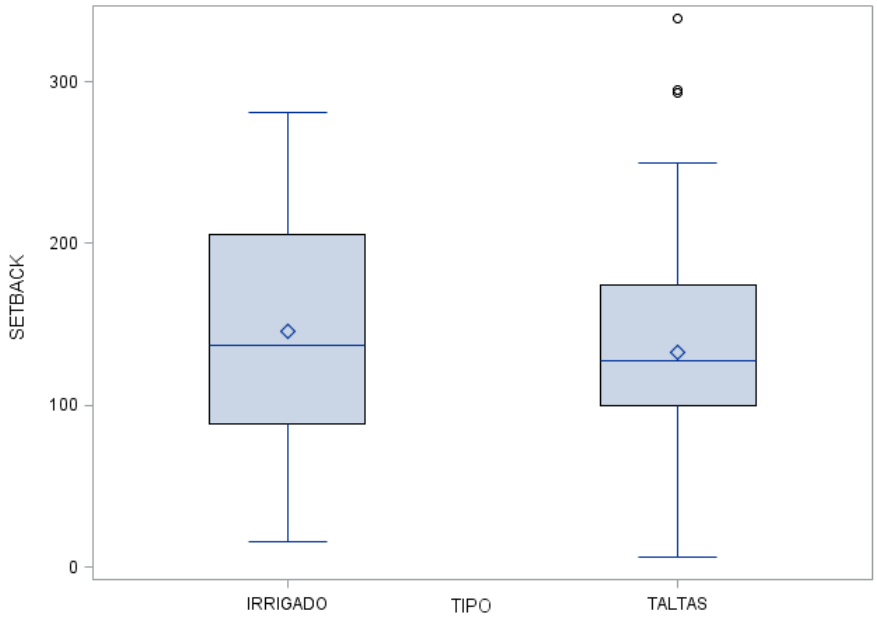

Figura 2.7: Medidas instrumentais de viscosidade:variável STBEAK

Observa-se na Figura 2.7 que, para cultivar de terras altas, 'a variável STBEAK apresenta baixa variabilidade e valores bem concentrados para cultivar irrigado.

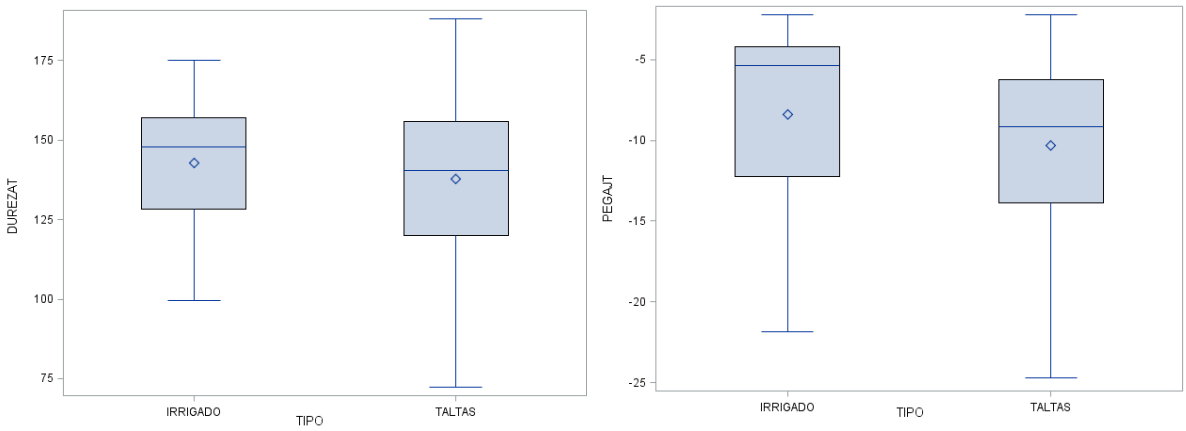

Figura 2.8: Medidas Instrumentais da textura:variáveis DUREZAT E PEGAJT

A Figura 2.8 apresenta as variáveis referentes *à* análise instrumental da textura. Verifica-se que a variabilidade é bem próxima para os dois tipos de cultivares, porém existem valores maiores para terrenos irrigados.

2.2 Análise de Componentes principais para os anos de 2013 e 2014 segundo o tipo de arroz

Nesse trabalho foi realizada a análise de componentes principais unindo os dados de 2013 e 2014, visando criar um score geral das CPs para obtenção da análise Bayesiana.

2.2.1 Terras Altas

A análise de componentes principais foi aplicada para todas as medidas instrumentais de viscosidade com O objetivo de reduzir a dimensão do estudo e eliminar problemas de multicolinearidade nos dados. Inicialmente, utilizou-se as variáveis TAAFIA, TAASEC, TG, PEAK, SETEBACK, BREAKDOWN e FINAL. Com base no resultado da ACP e da análise de correlação mostrada na (Tabela 2.4 e Figura 2.3), decidiu-se por retirar as variáveis TAASEC e SETBACK do estudo, pois estão altamente correlacionadas com as variáveis TAAFIA e FINAL, respectivamente. A escolha da variável TAASEC foi por conveniência do laboratório, pois o processo de mensuração é caro e difícil. A a variável SETBACK foi escolhida por convenção dos pesquisadores.

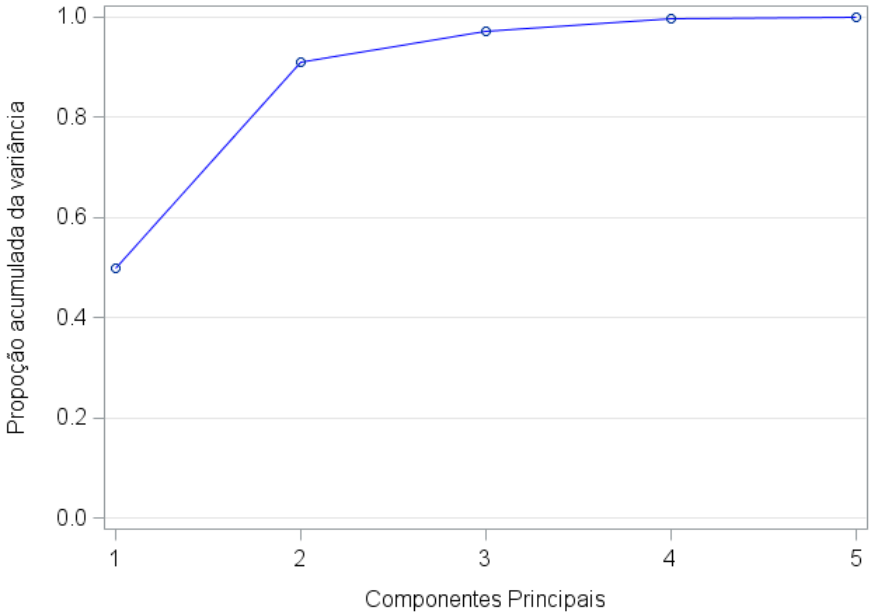

Figura 2.9: Proporção acumulada da variância explicada por cada componente principal para arroz de Terras Altas.

Observa-se na Tabela 2.5 e na Figura 2.9 que a primeira componente ($\hat{U}_1$) explica apenas 50% da variabilidade total dos dados. Portanto, é necessário utilizar pelo menos mais uma componente para se resumir as informações dos dados. Além disso, $\hat{U}_2$=41%. Logo, as duas primeiras CPs juntas explicam aproximadamente 91% da variância total. Dessa forma, optamos por utilizar apenas as duas primeiras componentes, pois elas contém as informações mais relevantes dos dados, uma vez que a terceira componente contribui pouquíssimo.

Tabela 2.5: Desvio padrão das componentes principais para arroz de Terras Altas e porcentagem da contribuição de cada uma dessas variâncias para a variância total.

Componentes Principais	Desvio Padrão	Proporção da variância	Proporção acumulada da variância
Primeira	1,57	0,50	0,50
Segunda	1,43	0,41	0,91
Terceira	0,55	0,06	0,97
Quarta	0,36	0,03	1
Quinta	0,12	<0,01	1

Verifica-se, na Tabela 2.6, que as variáveis TAAFIA, TG, FINAL E PEAK estão contribuindo de forma similar na primeira componente se opondo à contribuição da variável BREAKDOWN. Ressalta-se que a variável PEAK é a que menos contribui na primeira componente. Porém, é a que mais contribui na segunda componente. As variáveis TG e Final contribuem de maneira oposta às demais na segunda componente.

Tabela 2.6: Contribuição de cada variável nas duas primeiras componentes principais para arroz de Terras Altas e coeficiente de correlação entre as variáveis e as componentes principais selecionadas

Variáveis	Primeira Componente	Coeficiente de Correlação	Segunda Componente	Coeficiente de Correlação
TAAFIA	0,57	0,90	0,10	0,15
TG	0,51	0,73	-0,31	0,50
PEAK	0,16	0,25	0,66	0,95
BREAKDOWN	-0,28	-0,40	0,60	0,95
FINAL	0,57	0,88	0,30	0,42

2.2.2 Terrenos Irrigados

Visando a redução de dimensão no estudo utilizou-se análise de componentes principais para todas as medidas de viscosidade já relatadas na Seção 2.2.1 desse capítulo. Essa análise e a análise de correlação (Seção 2.1) mostram que as variáveis TAASEC e SETBACK podem ser retiradas do estudo, pois estão muito correlacionadas com TAAFIA e FINAL respectivamente. A justificativa para a escolha dessas duas variáveis encontra-se nessa Seção 2.2.1.

Tabela 2.7: Desvio padrão das componentes principais para arroz de Terrenos Irrigados e porcentagem da contribuição de cada uma dessas variâncias para a variância total.

Componentes Principais	Desvio Padrão	Proporção da variância	Proporção acumulada da variância
Primeira	1,60	0,51	0,51
Segunda	1,33	0,36	0,87
Terceira	0,67	0,09	0,96
Quarta	0,46	0,04	1
Quinta	0,11	<0,01	1

A Tabela 2.7 e a Figura 2.10 mostram que a primeira componente $\hat{U}_1$ explica 51% da variância total dos dados. Portanto, é necessário se utilizar mais uma componente para explicar os dados originais. Juntas, as duas primeiras componentes principais retêm 87% da variância total dos dados originais.

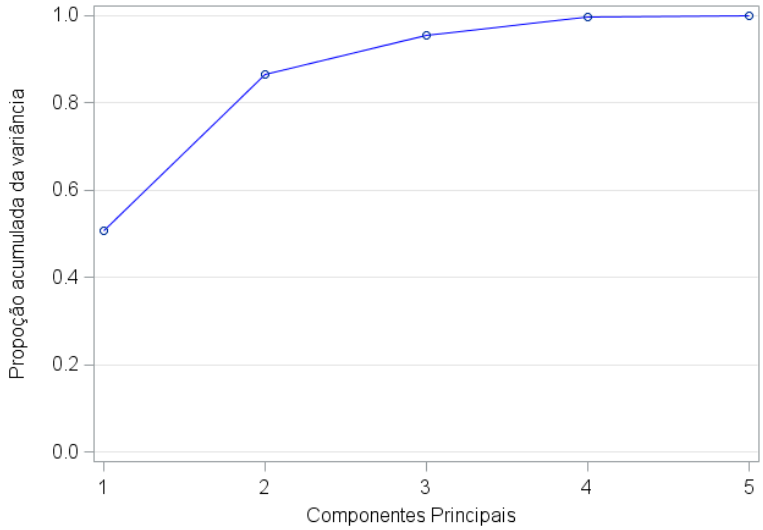

Figura 2.10: Proporção acumulada da variância explicada por cada componente principal para arroz de Terrenos Irrigados.

Tabela 2.8: Contribuição de cada variável nas duas primeiras componentes principais para arroz de Terrenos Irrigados e coeficiente de correlação entre as variáveis e as componentes principais selecionadas

Variáveis	Primeira Componente	Coeficiente de Correlação	Segunda Componente	Coeficiente de Correlação
TAAFIA	0,045	0,07	0,67	0,90
TG	-0,45	-0,60	0,35	0,56
PEAK	0,60	0,95	-0,13	0,17
BREAKDOWN	0,54	0,72	-0,29	-0,48
FINAL	0,38	0,60	0,56	0,75

Observa-se na Tabela 2.8 que as variáveis PEAK, BREAKDOWN e FINAL estão contribuindo de maneira semelhante na primeira componente, opondo-se à contribuição da variável TG nessa componente. A variável TAAFIA é a que menos contribui na primeira componente. Porém, é a que apresenta a maior contribuição na segunda componente. A segunda variável que mais contribui na segunda componente é a FINAL, seguida da TG.

2.3 Resultados: Regressão Logística binária Clássica

2.3.1 Predição da avaliação sensorial de Pegajosidade através de medidas de viscosidade para Terras Altas no ano de 2013

Esse trabalho se restringe apenas ao estudo dos modelos binários, para os anos de 2013 e 2014 por tipo de Arroz. Ressalta-se que foram analisados somente os dados refe-

rentes análise instrumental de pegajosidade mostrada na Tabela 2.2. A análise sensorial de dureza dada (Tabela 2.1), disposta em 7 categorias, não será analisada, pois os dados estão muito concentrados na categoria, “Macio", o que impossibilita a dicotomização das classes, transformando-as em “Macio"e “Firme". Portanto, na avaliação de pegajosidade considerando-se, apenas quatro categoria (Muito Pegajoso(MP), Pegajoso (P), Ligeiramente Macio(LM) e Solto (S)), utilizou-se apenas duas categorias: Pegajoso e Solto.

Tabela 2.9: Frequências da avaliação sensorial de pegajosidade binária para Terras Altas no ano de 2013

Pegajosidade	Frequência	Pegajosidade
Solto (S*)	42	58,33%
Pegajoso(P*)	30	41,67%

A Tabela 2.9 representa a nova classificação da variável pegajosidade, em que a categoria Solto (S*) abrange as observações classificadas nas categorias Levemente Solto (LS) e Solto (S). A categoria Pegajoso (P*) abrange as observações das categorias Muito Pegajoso (MP) e Pegajoso (P*).

Tem-se na Tabela 2.6 os pesos que cada medida de viscosidade representa na construção das duas primeiras componentes principais para Terras Altas considerando todo conjunto de dados. Portanto, foram usadas como variáveis explicativas os escores das componentes, calculados multiplicando-se estes pesos por cada valor das variáveis originais para os dados de Terras Altas somente do ano 2013. Essas novas variáveis obtidas foram usadas como variáveis explicativas no modelo de regressão logística binária. Os escores serão denominados de $C1$ e $C2$, e representam combinações lineares das variáveis originais e formados pelos coeficientes da primeira e segunda componente.

O modelo ajustado primeiro sugere a predição da avaliação sensorial de pegajosidade por meio das variáveis $C1$ e $C2$ de arroz de Terras Altas para o ano de 2013. Contudo, a variável $C2$ não foi significativa para um nível de significância de 5%, sendo o p-valor 0,6493. Dessa forma, foi utilizada apenas a variável $C1$, pois esta possui efeito significativo no modelo ($p - valor = 0,01$) menor do que o nível de significância de 5%.

Através desse modelo, a probabilidade de se obter a avaliação sensorial na categoria solto é dada por :

$$P(Y_i = 1|C1_i) = \frac{\exp[-2,0785 + 0,0148C1_i]}{1 + \exp[-2,0785 + 0,0148C1_i]}. \tag{2.1}$$

Tem-se que i-ésimo valor da variável $C1$ é dado por $C1_i$ e $\hat{\beta}_1 = 0,0148$ representa o efeito dessa variável na classificação da avaliação sensorial de pegajosidade.

A probabilidade do arroz pertencer à categoria Pegajoso (P*) é dada por:

$$P(Y_i = 0|C1_i) = 1 - P(Y_i = 1|C1_i). \tag{2.2}$$

Por meio do método de validação cruzada, obteve-se a matriz de classificação sensorial observada versus a classificação sensorial prevista de acordo com os modelos 2.1 e 2.2.

Tabela 2.10: Classificação sensorial de pegajosidade para Terras Altas no ano de 2013 versus a classificação prevista

		Classificação prevista	
		P*	S*
Classificação real	P*	16	14
	S*	8	34
Taxa de Erro de Classificação = 30,56%			

A classificação da avaliação sensorial de pegajosidade para o ano de 2013 versus a classificação prevista foi realizada por meio do modelo de regressão logística binária utilizando os scores das componentes principais das medidas de viscosidade para arroz de Terras Altas 2013.

Baseado na Tabela 2.10, calculou-se a taxa de erro de classificação que foi de 30,56% ((8+14)/(16+14+8+34)). Ou seja, quase $\frac{1}{3}$ das observações referentes à pegajosidade sensorial estão sendo classificadas erroneamente, o que pode ser considerado uma taxa bastante elevada.

Ressalta-se que no estudo de Rios [16] para o modelo 2.1 calculou-se essa mesma taxa de erro, todavia as componentes principais foram calculas de outra maneira como já descrito anteriormente.

A Figura 2.11 é a ilustra a Equação (2.1) e relaciona diferentes valores da variável $C1$ com a chance do arroz ser classificado como Solto (S*). Logo, à medida que os escores das medidas de viscosidades na primeira componente principal aumentam, a probabilidade do arroz receber avaliação sensorial como Solto (S*) aumenta. Portanto, valores de C1 entre 12 e 150 seriam classificados por meio da avaliação sensorial como Pegajoso (P*) e valores no intervalo de 150,5 a 363, como Solto (S*).

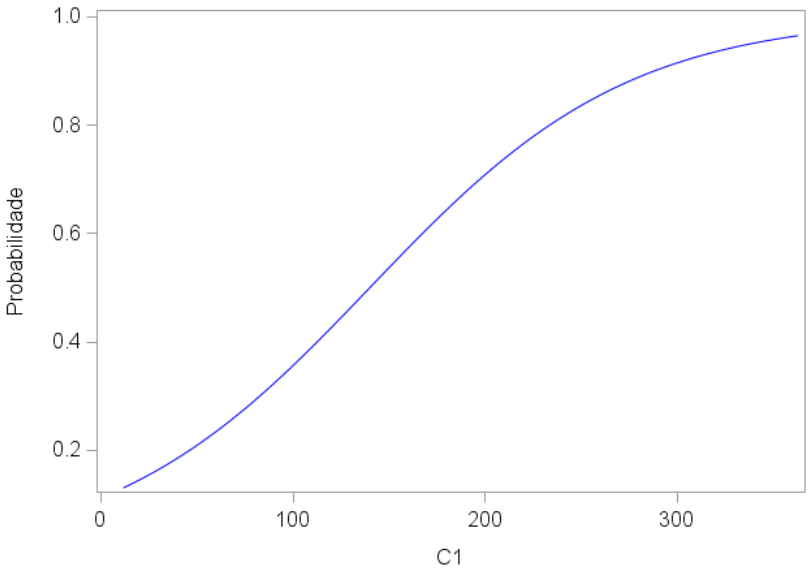

Figura 2.11: Probabilidade do arroz receber avaliação sensorial como Solto (S*) considerando diferentes valores da variável $C1$ de arroz de Terras Altas para o ano de 2013

2.3.2 Predição da avaliação sensorial de Pegajosidade através de medidas de viscosidade para Terrenos Irrigados no ano de 2013

Considerando 117 observações da avaliação sensorial do arroz de Terrenos Irrigados do ano de 2013, ajustou-se um modelo logístico binário propondo-se a predição da avaliação sensorial de pegajosidade por meio dos escores das CPs das medidas instrumentais de viscosidade. Considerou-se nesse modelo as variáveis $C1$ e $C2$, pois ambas possuem efeito significativo no modelo, ou seja, o p-valor de $C1$ foi $0,0007$ e o de $C2$ foi menor que $0,0001$. Portanto, ambos são menores que o nível de significância, $\alpha = 5\%$, para cada variável.

Baseado nesse modelo modelo, a probabilidade de obter-se avaliação sensorial na categoria solto é dada por:

$$P(Y_i = 1|C1_i, C2_i) = \frac{\exp[-0,8701 - 0,0134C1_i + 0,0226C2_i]}{1 + \exp[-0,8701 - 0,0134C1_i + 0,0226C2_i]}, \tag{2.3}$$

onde o i-ésimo valor da variável $C1$ é dado por $C1_i$, $\hat{\beta}_1 = -0,0134$ e representa o efeito dessa variável na classificação da avaliação sensorial de pegajosidade, o i-ésimo valor da variável $C2$ é dado por $C2_i$ e $\hat{\beta}_2 = 0,0226$.

A probabilidade do arroz pertencer à categoria Pegajoso (P*) é dado por:

$$P(Y_i = 0|C1_i, C2_i) = 1 - P(Y_i = 1|C1_i, C2_i). \tag{2.4}$$

A Tabela 2.11 mostra a matriz de classificação sensorial observada versus a classificação sensorial prevista por meio dos modelos 2.3 e 2.4.

Tabela 2.11: Classificação sensorial de pegajosidade para Terrenos Irrigados no ano de 2013 versus a classificação prevista

		Classificação prevista	
		P*	S*
Classificação real	P*	63	13
	S*	27	14
Taxa de Erro de Classificação = 34,18%			

A Classificação da avaliação sensorial de pegajosidade para o ano de 2013 versus a classificação prevista foi realizada por meio do modelo de regressão logística binária utilizando-se os escores das medidas de viscosidade nas componentes principais de arroz de Terrenos irrigados.

Verifica-se na Tabela 2.11 que a taxa de erro de classificação é de 34,18%. Portanto, tem-se que 34,18% das observações da análise sensorial de pegajosidade de arroz de Terrenos Irrigados estão sendo classificada erroneamente quando avaliadas a partir dos escores das medidas de viscosidade ($C1$ e $C2$), taxa considerada alta.

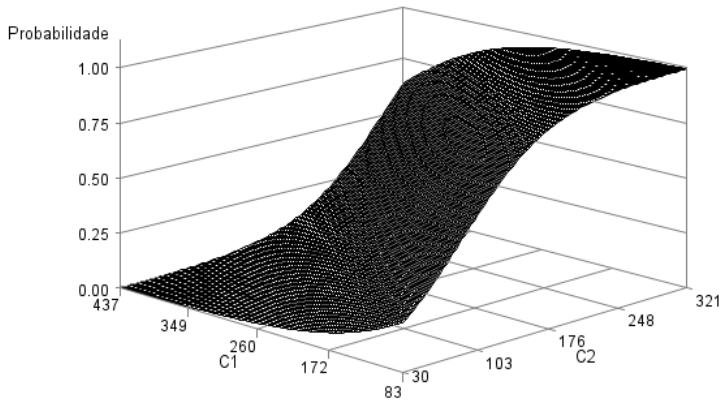

Figura 2.12: Probabilidade do arroz de Terrenos Irrigados para o ano de 2013 receber avaliação sensorial como Solto (S*) considerando diferentes valores das variáveis $C1$ e $C2$.

Observa-se na Figura 2.12, que para valores altos de $C1$ e baixos de $C2$, o arroz terá alta probabilidade de ser classificado como Pegajoso (P*), já para valores altos de $C2$ e baixos de $C1$ tem-se uma maior probabilidade do arroz ser classificado como Solto (S*). Para valores intermediários de $C1$ e $C2$ não é possível concluir com precisão se a avaliação sensorial seria mais propensa a ser do tipo Pegajoso (P*) ou Solto (S*).

2.3.3 Predição da avaliação sensorial de Pegajosidade através de medidas de viscosidade para Terras Altas no ano de 2014

A predição da avaliação sensorial de pegajosidade (com 72 avaliações) utilizando medidas de viscosidades para Terras Altas no ano de 2014 foi realizada ajustando-se um modelo que relaciona a avaliação sensorial de pegajosidade arroz de Terras Altas com os escores das componentes principais das medidas de viscosidade.

O modelo ajustado considerou-se as variáveis $C1$ e $C2$, pois essas possuem efeito significativo no modelo. Destaca-se que ambas as variáveis possuem p-valores menores que o nível de significância $\alpha = 5\%$. O p-valor de $C1$ foi menor que $0,0001$ é de $C2$ menor que $0,00004$.

O i-ésimo valor da variável $C1$ é dado por $C1_i$ e $\hat{\beta}_1 = 0,0843$ representa o efeito dessa variável na classificação da avaliação sensorial de pegajosidade. O i-ésimo valor da variável $C2$ é dado por $C2_i$ e $\hat{\beta}_2 = -0,0366$.

Para esse modelo, a probabilidade do arroz ser classificado na categoria Solto (S*) é dada por:

$$P(Y_i = 1|C1_i, C2_i,) = \frac{\exp[-2,65830 - 0,0843C1_i + 0,0226C2_i]}{1 + \exp[-2,65830 - 0,0843C1_i + 0,0226C2_i]}. \tag{2.5}$$

A de pertencer à categoria Pegajoso é dada por:

$$P(Y_i = 0|C1_i, C2_i) = 1 - P(Y_i = 1|C1_i, C2_i). \tag{2.6}$$

Tabela 2.12: Classificação da avaliação sensorial de pegajosidade para Terras Altas no ano de 2014 versus a classificação prevista

		Classificação prevista	
		P*	S*
Classificação real	P*	33	1
	S*	3	35
Taxa de Erro de Classificação = 5,56%			

A classificação da avaliação sensorial da pegajosidade para o ano de 2014 versus a classificação prevista foi desenvolvida, por meio do modelo de regressão logística binária, utilizando os escores das componentes principais das medidas de viscosidade de arroz de Terras Altas.

Observa-se na Tabela 2.11 a classificação com base nos modelos 2.5 e 2.6. Obteve-se a taxa de erro de classificação de 5,56%. Indica que um percentual pequena das observações na avaliação sensorial de pegajosidade são classificadas erroneamente para arroz de Terras Altas.

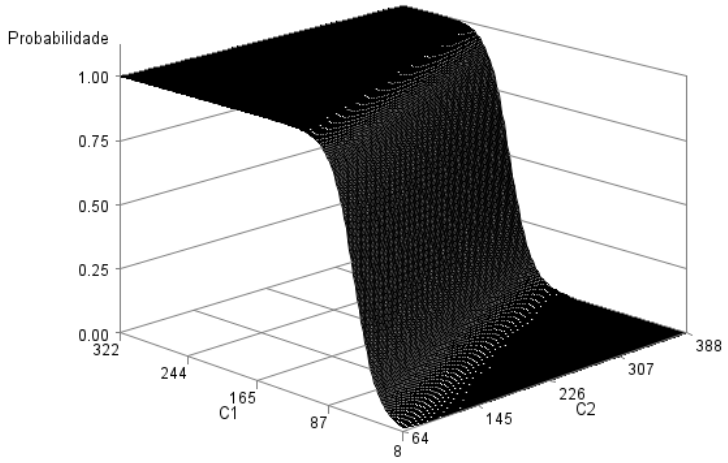

Figura 2.13: Probabilidade do arroz de Terras Altas para o ano de 2014 receber avaliação sensorial como Solto (S*) considerando diferentes valores das variáveis $C1$ e $C2$

De acordo com a Figura 2.14, a probabilidade do arroz receber avaliação sensorial como Solto (*) diminui a medida que a variável $C1$ diminui. Portanto para valores altos de $C1$ o arroz recebe avaliação como Solto, e para valores baixos de $C1$ o arroz recebe avaliação como Pegajoso (P*). Note que para os valores intermediários da variável C1 é impossível concluir com segurança se o arroz seria classificado como Solto ou Pegajoso.

2.3.4 Predição da avaliação sensorial de Pegajosidade através de medidas de viscosidade para Terrenos Irrigados no ano de 2014

A predição da avaliação sensorial de Pegajosidade através de medidas de viscosidade para Terrenos Irrigados no ano de 2014 foi desenvolvida ajustando um modelo que relaciona a avaliação sensorial de pegajosidade de arroz de Terrenos Irrigados no ano de 2014 com os escores das medidas de viscosidades nas componentes principais, considerando-se 75 observações. As variáveis $C1$ e $C2$ são significativas, ambas com p-valor menores que $0,0001$, ou seja, menor que o nível de significância de 5%.

Como abordado anteriormente, o i-ésimo valor da variável $C1$ é dado por $C1_i$, $\hat{\beta}_1 = -0,0459$ representa o efeito dessa variável na classificação da avaliação sensorial de pegajosidade. E, o i-ésimo valor da variável $C2$ é dado por $C2_i$ e $\hat{\beta}_2 = 0,06939$.

A probabilidade do arroz ser classificado na categoria Solto (S*) é dada por:

$$P(Y_i = 1|C1_i, C2_i) = \frac{\exp[0,7962 - 0,0459C1_i + 0,0693C2_i]}{1 + \exp[0,7962 - 0,0459C1_i + 0,0693C2_i]}. \tag{2.7}$$

E, a probabilidade de pertencer a categoria Pegajoso é dado por:

$$P(Y_i = 0|C1_i, C2_i) = 1 - P(Y_i = 1|C1_i, C2_i). \tag{2.8}$$

Tabela 2.13: Classificação da avaliação sensorial de pegajosidade para Terrenos Irrigados no ano de 2014 versus a classificação prevista

		Classificação prevista	
		P*	S*
Classificação real	P*	26	5
	S*	2	42
Taxa de Erro de Classificação = 9,33%			

A classificação sensorial de pegajosidade para o ano de 2014 versus a classificação prevista por meio do modelo de regressão logística binária utilizando os escores das componentes principais das medidas de viscosidade de arroz de Terrenos Irrigados. Com base na Tabela 2.13, tem-se a taxa do erro de classificação calculado por validação cruzada de $9,33\%$.

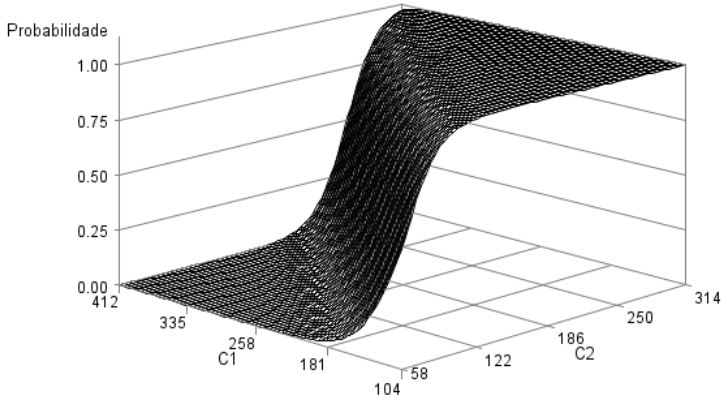

Figura 2.14: Probabilidade do arroz de Terrenos Irrigados para o ano de 2014 receber avaliação sensorial como Solto (S*) considerando diferentes valores das variáveis $C1$ e $C2$

Verifica-se na Figura 2.14 que quando as variáveis $C1$ e $C2$ aumentam, a probabilidade do arroz ser classificado como Solto(*) é maior, já para valores baixos de $C2$ e altos de $C1$ o arroz seria classificado como Pegajoso (P*). E para os valores intermediários da variável $C1$ é impossível concluir com segurança que o arroz seria preferencialmente classificado como Solto ou Pegajoso.

2.4 Resultados da Regressão Logística Bayesiana

Nessa seção, serão mostrados alguns resultados dos modelos logísticos Bayesianos para os tipos de cultivar de arroz, do ano de 2014, que foram ajustados usando prioris não informativas e prioris baseadas na informação dos modelos clássicos do ano de 2013.

Devido à esse processo de atualização, ou seja, utilizar a informação dos modelos de 2013 como priori para 2014, optou-se por obter as componentes principais para todo conjunto de dados (2013 e 2014) e depois utilizar esses coeficientes para criar escores das medidas de viscosidade do ano de 2013 e também para o ano de 2014 para os dois tipos de cultivares de arroz.

2.4.1 Modelo Binário para Terras altas no ano de 2014 usando priori não informativa

[2]

Primeiramente, ajustou-se modelos para os dados de Terras altas 2014, usando as mesmas variáveis do modelo clássico, supondo que cada parâmetro segue uma distribuição uniforme. Foi realizada uma análise de sensibilidade, na qual, ajustou-se um primeiro modelo com um intervalo mais abrangente e em seguida vários modelos diminuindo 0.5 em cada um dos intervalos. Aqui será apresentado apenas o melhor resultado obtido.

O modelo foi ajustado no software SAS, por meio da `Proc MCMC`, usando o algoritmo de Metropolis-Hasting para obter a distribuição a posteriori. As prioris não informativas para β_0, β_1 e β_2 foram $U(-14,5; 14,5)$, $U(-5,5; 5,5)$ e $U(-4,5; 4,5)$ respectivamente. Logo, a distribuição *a posteriori* é proporcional a verossimilhança. Portanto, espera-se que os parâmetros ajustados pela metodologia Bayesiana sejam próximos aos ajustados pela metodologia clássica, dado que os dois estão usando apenas a informação contida na verossimilhança. Assim, temos que:

$$\pi(\beta|Y) \propto L(\beta|Y)$$

A obtenção da cadeia usou um aquecimento de 4000000, uma amostra a partir de 10000 gerações e um salto de tamanho 15. O tamanho do aquecimento e do salto foram tomados de forma que a convergência da cadeia fosse garantida. A cadeia foi gerada tomando como pontos iniciais o valor 0.

Inicialmente, será feito o diagnóstico da cadeia de *Markov*, pois antes de observar as distribuições *a posteriori* marginais para os parâmetros é necessário averiguar as condições de convergência e de autocorrelação para as cadeias. Essas verificações asseguram que as cadeias convirjam para as distribuições de interesse. A convergência para a distribuição *a posteriori* é garantida quando observa-se que os valores das variáveis aleatórias estão

[2] Deve-se observar que as prioris utilizadas nessa Seção (2.4.1) foram formuladas com uma pequena variação (intervalo muito pequeno), portanto, podem ser consideradas como informativas. Porém, foi verificado que o uso de intervalos grandes não alterou significativamente os resultados.

distribuídos aleatoriamente em torno de um valor fixo. Além disso, deve-se verificar um decaimento rápido das autocorrelações.

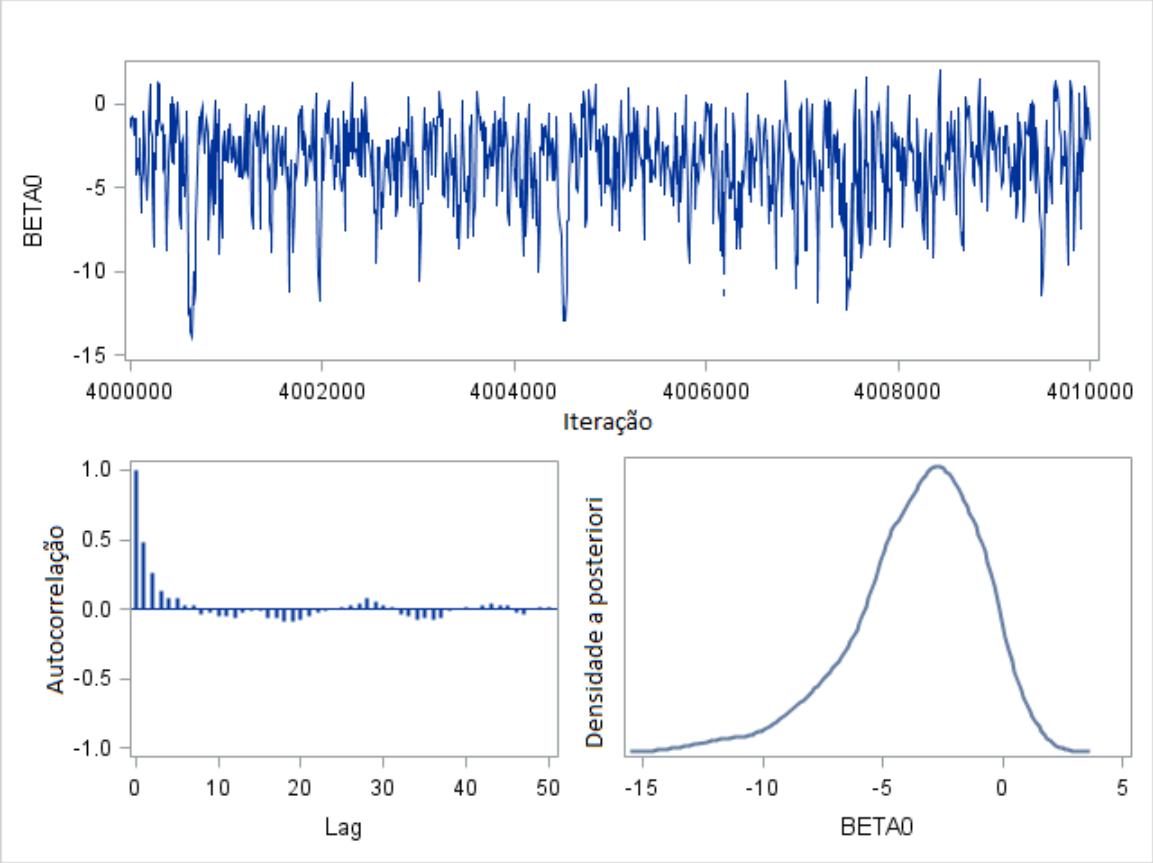

Figura 2.15: Gráficos de diagnóstico da cadeia para o parâmetro β_0 e sua distribuição a posteriori considerando arroz de Terras Altas, ano 2014.

A Figura 2.15 representa o gráfico de diagnóstico de convergência, ou seja, os valores da variável aleatória β_0 versus o número de iterações excluindo as iterações de aquecimento. Nota-se que os valores da v.a β_0 está variando entorno de um valor fixo mostrando um comportamento estacionário. Dessa forma, é possível dizer que esta cadeia está convergindo para a distribuição *a posteriori* desejada. Ressalta-se que por esse gráfico pode-se ter uma ideia da significância dos parâmetros.

Verificasse, ainda na Figura 2.15 o gráfico das autocorrelações versus *lags*. Nota-se que, tomando observações com espaçamento de 15 em 15 é observado um decaimento rápido das autocorrelações evidenciando que os valores da cadeia não estão correlacionados principalmente após o *lag* 15, indicando que os valores gerados após o aquecimento são independentes.

O terceiro gráfico da Figura 2.15 mostra a distribuição *a posteriori* para o β_0, que é assimétrica à esquerda. Nota-se que esta distribuição inclui o valor *zero* próximo ao limite da cauda.

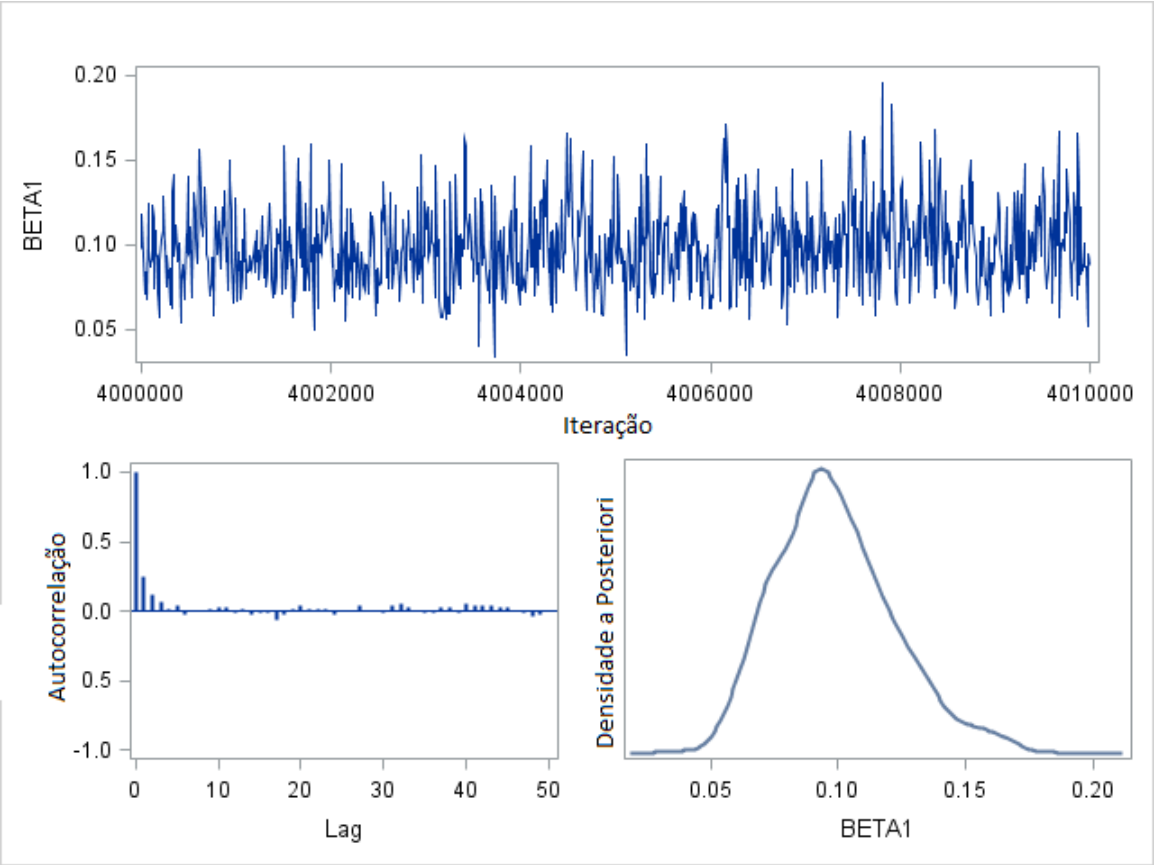

Figura 2.16: Gráficos de diagnóstico da cadeia para o parâmetro β_1 e sua distribuição *a posteriori* considerando arroz de Terras Altas, ano 2014.

Na Figura 2.16 verifica-se que o primeiro gráfico apresenta os valores da cadeia depois do período de aquecimento. O histórico da cadeia mostra que os valores gerados para β_1 estão variando aleatoriamente em torno de um valor fixo e permanecem constante ao longo das iterações, indicando a convergência da cadeia.O segundo gráfico da Figura 2.16 mostra que houve um decaimento rápido das autocorrelações quando tomadas com espaçamento de 15 em 15. Logo há indicio que os valores gerados da distribuição a *posteriori* são independentes.

Observa-se no último gráfico a distribuição *a posteriori* para β_1, que é assimétrica à direita.

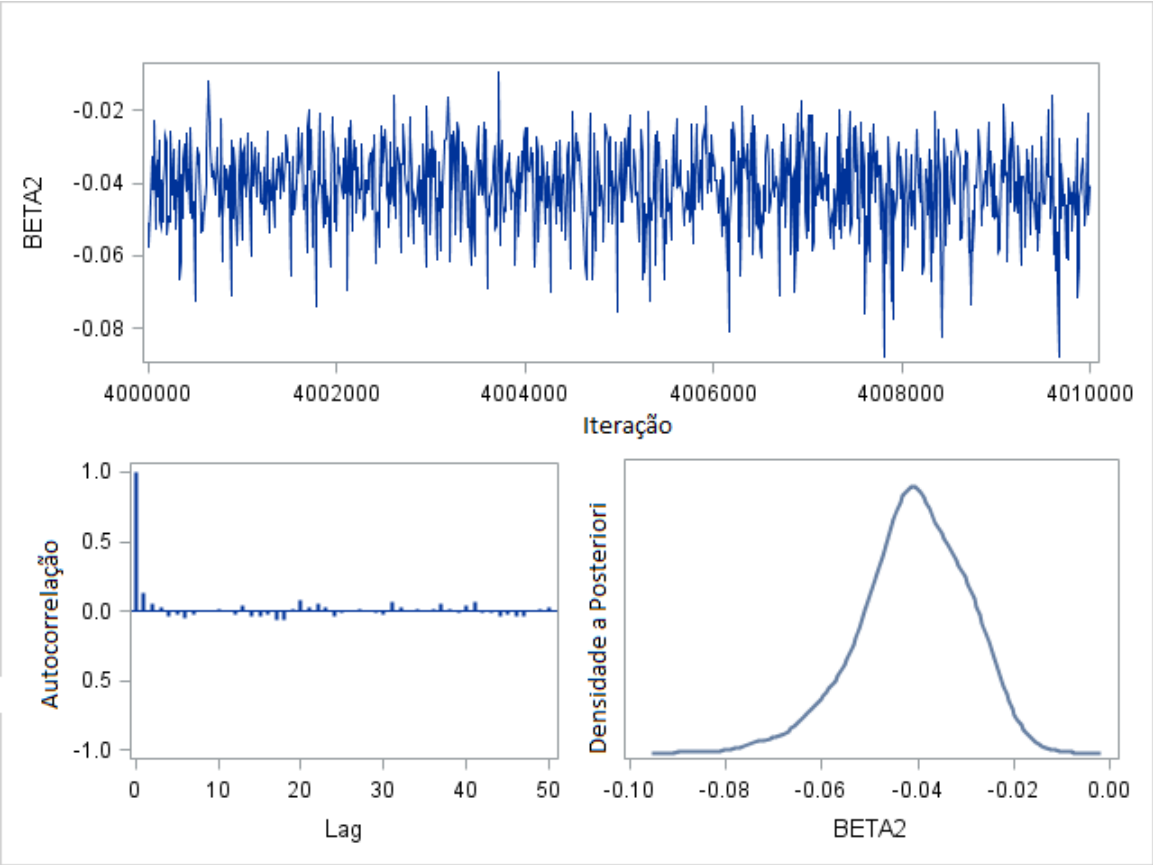

Figura 2.17: Gráficos de diagnóstico da cadeia para o parâmetro β_2 e sua distribuição a posteriori considerando arroz de Terras Altas, ano 2014.

Com base na Figura 2.17, verifica-se pelo primeiro gráfico que as condições de convergência da cadeia são garantidas e que a autocorrelação é desprezível. A distribuição *a posteriori* possui caudas mais pesadas para a esquerda.

Tabela 2.14: Estimadores Bayesianos para Terras altas para o ano de 2014

Parâmetro	Média	Erro Padrão	2.5%	97,5%
$\hat{\beta}_0$	-3,66	2,73	-8,97	1,04
$\hat{\beta}_1$	0,09	0,02	0,06	0,14
$\hat{\beta}_1$	-0,04	0,01	-0,06	-0,03

Na Tabela 2.14 estão as estimativas dos parâmetros Bayesianos, onde a média *a posteriori* e o erro padrão são calculados a partir da amostra *a posteriori* de tamanho $n = 667$.

Observa-se que as médias *a posteriori* dos parâmetros são muito próximas das estimativas dos parâmetros estimados pelo método de verossimilhança na Seção 2.3.3. Isso já era esperado, visto que foi utilizado uma priori não informativa, ou seja, *a posteriori* é proporcional à verossimilhança.

Houve ganho, portanto, apenas em termos do intervalo de credibilidade. Esse possui uma interpretação direta sobre os parâmetros. Por exemplo, o intervalo de credibilidade para β_0 significa que a estimativa de β_0 está variando entre $[-8,97; 1,04]$ com 95% de probabilidade. Nota-se que esse intervalo inclui o zero indicando que o intercepto não é significativo nesse modelo. A interpretação para os outros parâmetros é similar. E, pode ser visto na Tabela 2.14 que o intervalo de credibilidade para β_1 e β_2 não incluem o zero, logo as variáveis $C1$ e $C2$ são significativas. O comprimento desse intervalo é pequeno, indicando, assim, que a distribuição é concentrada em torno do valor estimado.

Após essas verificações, foi realizada análise de discriminante logística, com o cálculo das probabilidades preditivas da classificação sensorial de pegajosidade. A probabilidade do arroz receber avaliação sensorial como Solto (S*) é dada por:

$$P(y = 1|C, D) = \int p(y = 1|C, \beta)p(\beta|D)p(\beta), \tag{2.9}$$

onde C representa as variáveis explicativas D é o conjunto de todos os dados disponíveis no problema β o vetor de parâmetros.
E a probabilidade do arroz receber avaliação sensorial como Pegajoso (P*) é dado por:

$$P(y = 0|C, D) = 1 - p(y = 1|C, D) \tag{2.10}$$

quando $P(y = 1|C, D) > 0,5$ então o arroz é classificado como Solto (S*)

Tabela 2.15: Classificação da avaliação sensorial de pegajosidade versus a classificação prevista, por meio do modelo logístico Bayesiano

		Classificação prevista	
		P*	S*
Classificação real	P*	33	1
	S*	3	35
Taxa de Erro de Classificação = 5,56%			

A Tabela 2.19, apresenta a matriz de classificação da avaliação sensorial de pegajosidade para o ano de 2014 versus o modelo de logístico Bayesiano utilizando os escores das componentes principais das medidas de viscosidade de arroz de Terras Altas . Ressalta-se que o erro de classificação por esse modelo foi de 5,56%. Logo, foi exatamente igual ao erro obtido no modelo clássico. Porém, essa taxa de erro não foi calculada por validação cruzada, podendo está superestimada.

2.4.2 Modelo Binário para Terrenos Irrigados no ano de 2014 usando priori não informativa

[3]

O modelo foi ajustado considerando-se as mesmas variáveis do caso clássico da Seção 2.3.4. Considerou-se uma priori com distribuição uniforme para os coeficientes: $U(-4,5;4,5)$, $U(-3,5;3,5)$ e $U(-2,5;2,5)$. Ressalta-se que escolha dessas prioris já foram justificadas na Seção 2.4.

Esse modelo utilizou um aquecimento de tamanho 400000 e obteve-se a amostra com 10000 gerações e salto de tamanho 10. Foram ajustados diversos modelos com diferentes tamanhos de cadeias e saltos. De acordo com os gráficos de diagnóstico o modelo usando as prioris em tela foi o melhor modelo, na qual, obteve-se as maiores adequações das suposições da cadeia, conforme os gráficos abaixo:

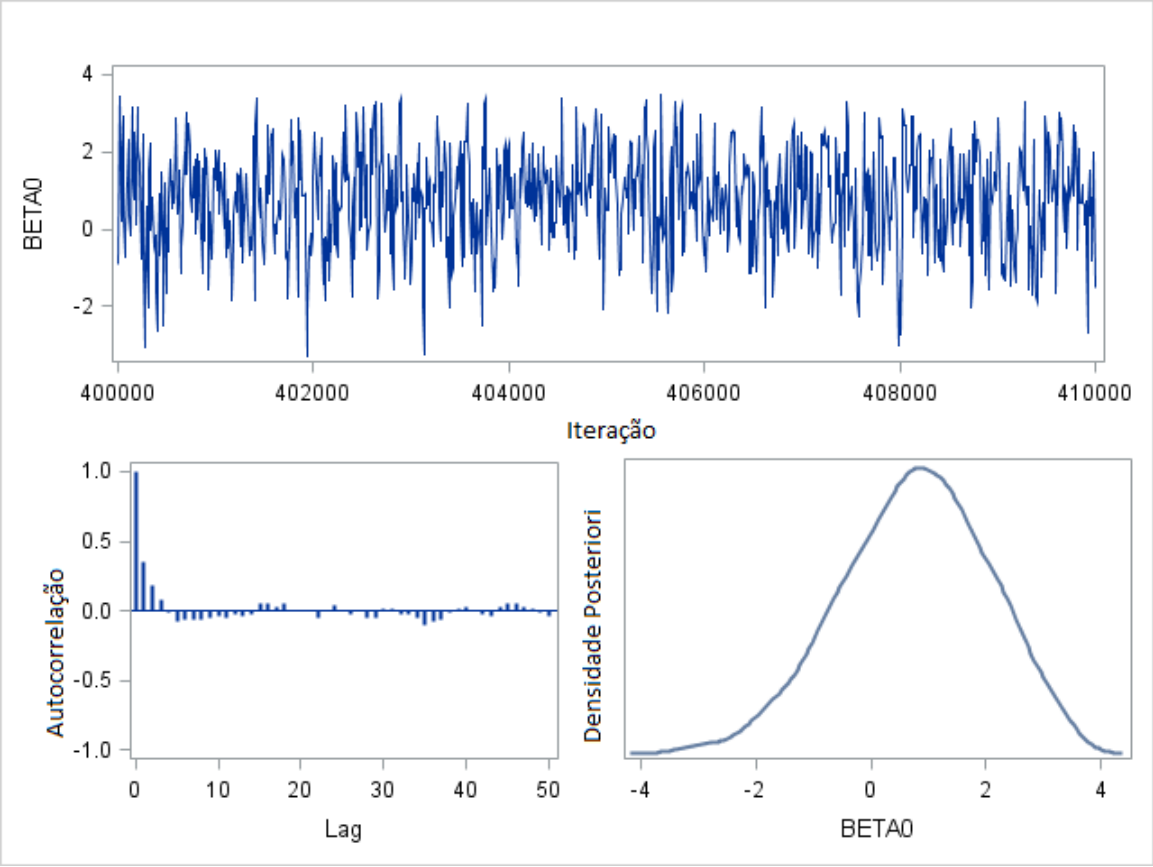

Figura 2.18: Gráficos de diagnóstico da cadeia para o parâmetro β_0 e sua distribuição a posteriori considerando arroz de Terrenos Irrigados, ano 2014.

Verifica-se, na Figura 2.18, que o primeiro gráfico mostra a trajetória *MCMC* ex-

[3]Deve-se observar que as prioris utilizadas nessa Seção (2.4.2) foram formuladas com uma pequena variação (intervalo muito pequeno), portanto, podem ser consideradas como informativas. Porém, foi verificado que o uso de intervalos grandes não alterou significativamente os resultados.

cluindo as iterações de aquecimento versus os valores da variável aleatória β_0. O histórico dessa cadeia mostra que no geral os valores permanecem aleatoriamente em torno de um valor fixo ao longo das iterações, indicando que cadeia convergiu. Por esse gráfico também é possível se ter uma ideia da significância do parâmetro, ou seja, quando o valor 0 está entre os valores gerados é um indicativo de que mesmo a cadeia convergindo, o parâmetro não é significativo.

Na Figura 2.18 ainda o gráfico 2 representa as autocorrelações versus *lags*. Esse é o caso onde, mesmo pegando as observações com espaçamento de 10 em 10, ainda há indicio de dependência. Porém, nota-se um decaimento rápido, e a partir do *lag* 20 pode-se dizer que existe pouca correlação.

O último gráfico da Figura 2.18 mostra a distribuição *a posteriori* para o parâmetro β_0, nota-se uma leve assimetria para a esquerda e o valor zero incluso é mais próximo da cauda da distribuição, evidenciando a não significância desse parâmetro.

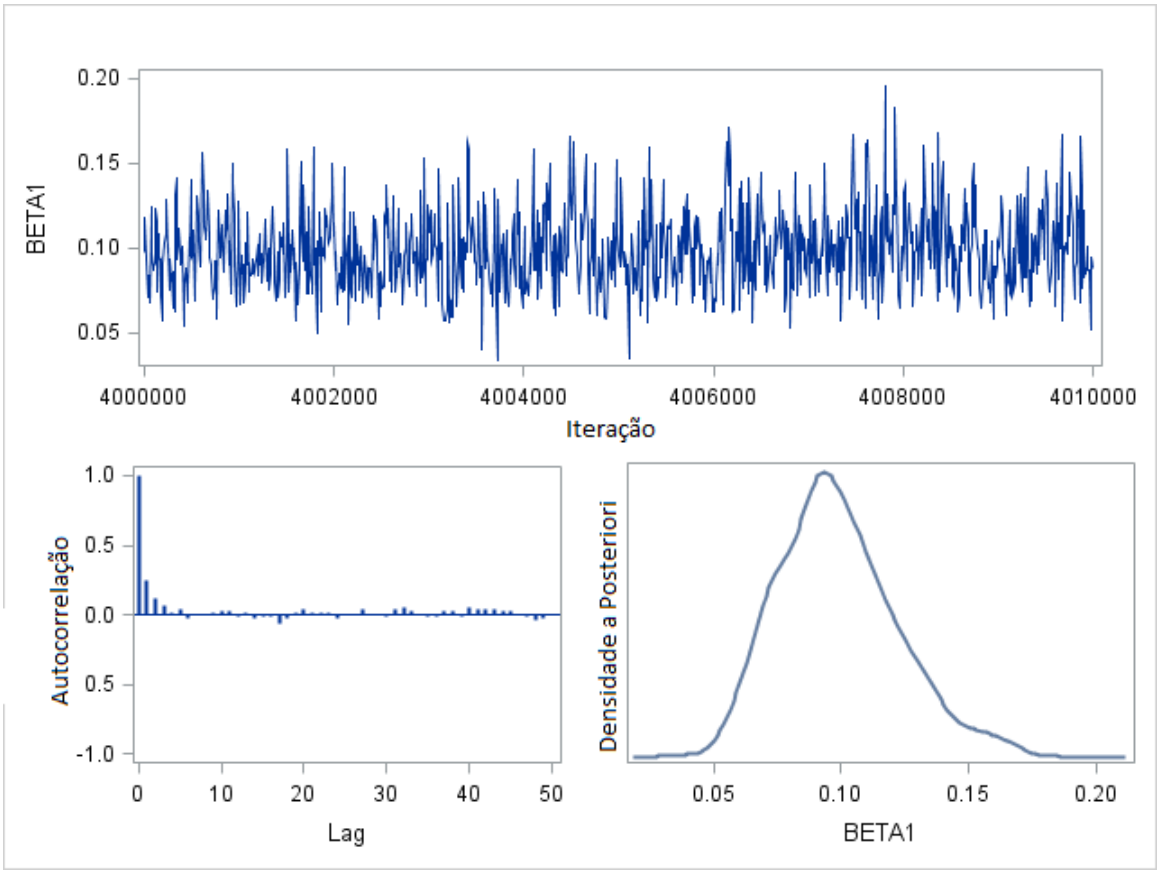

Figura 2.19: Gráficos de diagnóstico da cadeia para o parâmetro β_1 e sua distribuição *a posteriori* considerando arroz de Terras Irrigados, ano 2014.

Observa-se, na Figura 2.19, que os dois primeiros gráficos são de diagnóstico da cadeia para β_1 e o terceiro mostra a distribuição *a posteriori*. No entanto, o primeiro gráfico indica que a cadeia de *Markov* gerada permanece estacionária, ou seja, está variando em torno de um valor fixo ao longo das iterações, indicando a convergência da cadeia.

O segundo gráfico da Figura 2.19 mostra que houve um decaimento muito rápido das autocorrelações. Portanto, pegando as observações com espaçamento de 10 em 10 há um indicativo que está sendo gerados valores independentes (pseudo aleatórios).

A distribuição *a posteriori* obtida para β_1 é assimétrica. Nota-se que a distribuição não inclui o valor 0, indicando que este parâmetro foi significativo no modelo.

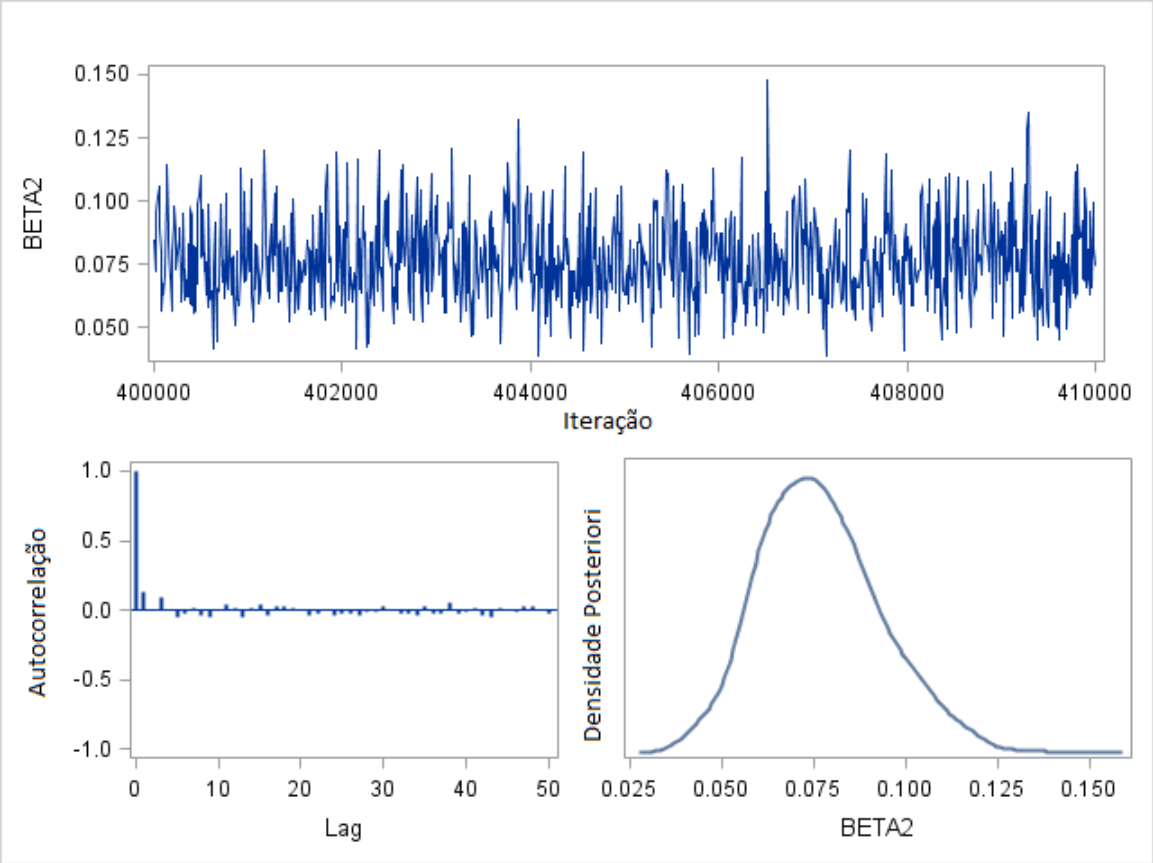

Figura 2.20: Gráficos de diagnóstico da cadeia para o parâmetro β_2 e sua distribuição a posteriori considerando arroz de Terrenos Irrigados para o ano de 2014.

Os gráficos de diagnóstico da cadeia para o parâmetro β_2 podem ser vistos na Figura 2.20. O primeiro gráfico indica que os valores gerados para β_2 está variando em torno de um ponto fixo conforme o número de iterações. Isso indica que a cadeia pode ser estacionária e que atingiu a convergência. O gráfico referente às autocorrelações indica um decaimento rápido. Portanto, a indício de que as observações geradas são independentes (pseudo aleatórios).

Tabela 2.16: Estimadores Bayesianos para Terrenos Irrigados, ano de 2014

Parâmetro	Média	Erro Padrão	2.5%	97,5%
$\hat{\beta}_0$	0,76	-1,64	-1,66	3,91
$\hat{\beta}_1$	-0,05	0,01	-0,07	-0,03
$\hat{\beta}_2$	0,07	0,01	0,05	0,12

Observa-se que as estimativas dos parâmetros obtidos a partir de amostras de tamanho 1000 das distribuições *a posteriori* são muito próximas das estimativas dos parâmetros estimados pelo método de máxima verossimilhança na Seção 2.3.4. Por exemplo, o intervalo de credibilidade para β_0 significa que a estimativa de β_0 está variando entre $[-1,66; 1,04]$ com 95% de probabilidade. Nota-se que esse intervalo inclui o zero, indicando que o intercepto não é significativo nesse modelo. A interpretação para os outros parâmetros é similar. Porém, as variáveis $C1$ e $C2$ são significativas, pois o intervalo de credibilidade para a distribuição *a posteriori* de β_1 e β_2 não incluem o zero.

Tabela 2.17: Classificação da avaliação sensorial de pegajosidade versus a classificação prevista, por meio do modelo logístico Bayesiano para Terrenos Irrigados

		Classificação prevista	
		P*	S*
Classificação real	P*	26	2
	S*	5	42
Taxa de Erro de Classificação=9,33%			

Observa-se na Tabela 2.17 que o erro de classificação de acordo com modelo logístico Bayesiano usando uma priori não informativa considerando os escores das componentes principais das medidas de viscosidade. Logo, apenas 9,33% das observações foram classificadas erroneamente, portanto, não houve redução no erro de classificação se comparado com o modelo ajustado por verossimilhança. Ressalta-se que não foi utilizado o método de validação cruzada no cálculo dessa taxa, porém, foi utilizado para o modelo clássico.

2.4.3 Modelo Binário para arroz de Terras Altas, no ano de 2014, usando priori informativa

A literatura de inferência Bayesiana deixa bem destacado que a principal vantagem da abordagem é utilizar um conhecimento externo ao banco de dados, ou seja, o conhecimento a priori. Foram realizadas algumas sugestões de prioris para modelos logístico, porém, não houve adequação para esses dados.

A priori utilizada nesse estudo foi baseada numa sugestão de Sullivan e Greenland [21], que propõe utilizar como priori para os $\beta(s)$ a distribuição normal com média e variância baseadas no intervalo de confiança da razão de chances (OR)=e^{β} (Agresti [1]). Essa medida foi obtida por meio dos modelos ajustados por verossimilhança para os dados do ano de 2013.

Considerando-se o intervalo de confiança de 95% para a razão de chances, os limites do IC são dados por: OR_{inf} e OR_{sup} que são respectivamente o limite inferior e o limite superior do intervalo.

A média é dada por:

$$\beta_{prior} = \ln\left(OR_{prior}\right) = \frac{\ln(OR_{sup}) + \ln(OR_{inf})}{2}. \tag{2.11}$$

E, a variância é dada por:

$$v_{prior} = \left[\frac{\ln(OR_{sup}) - \ln(OR_{inf})}{2 * 1,96}\right]^2. \tag{2.12}$$

Ajustou-se um modelo para o cultivar de Terras Altas do ano de 2014 com as informações do modelo ajustado por verossimilhança para os dados de arroz de Terras Altas do ano de 2013. Esse modelo é dado na Equação 2.1 e apenas a variável $C1$ possui efeito significativo no modelo, considerando um nível de significância de 5%.

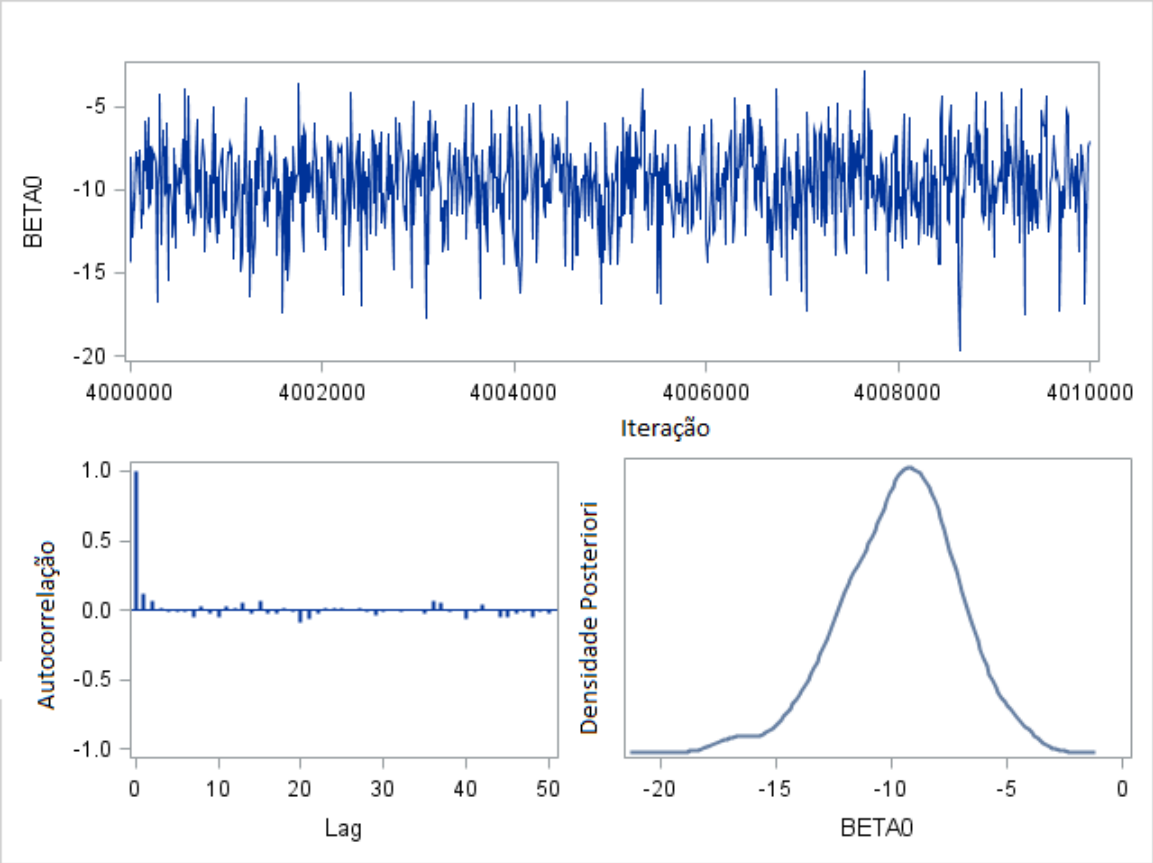

Figura 2.21: Gráficos de diagnóstico da cadeia para o parâmetro β_0 e sua distribuição *a posteriori* considerando uma priori informativa para arroz de Terras Altas para o ano de 2014.

No modelo visto na Equação (2.1), a variável $C2$ não é significativa. Portanto obteve-se uma priori apenas para β_1. Onde a $OR = 1,015$, no entanto, encontrou-se uma distribuição $N(0,01481660, 0000217)$. Nota-se que obteve uma variância muito pequena.

Ressalta-se, que o modelo foi ajustado considerando a variância igual a 1 e mantendo a média. Para o intercepto utilizou-se uma priori não informativa, considerando uma variância muito grande $N(0, 1000)$.

As Figuras 2.21 e 2.22 apresentam os gráficos de diagnóstico para os parâmetros β_0 e β_1. Observou-se que, para ambos os casos, a cadeia permanece estacionária ao longo das iterações após o período de aquecimento, indicando a convergência da cadeia. No geral, considerando um salto de tamanho 10, nota-se um decaimento rápido nas autocorrelações, apontando que os valores gerados para a cadeia não são correlacionados. As distribuições *a posteriori* são assimétrica, porém, não inclui o valor zero.

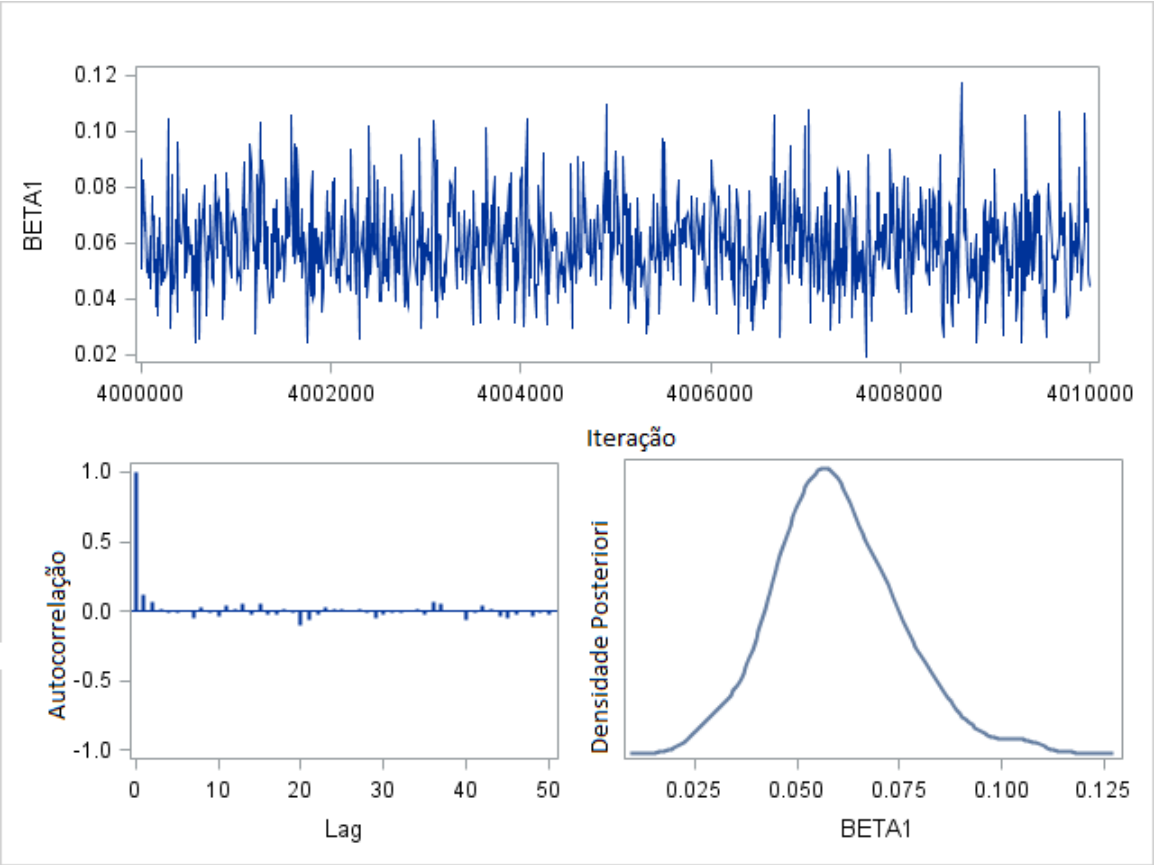

Figura 2.22: Gráficos de diagnóstico da cadeia para o parâmetro β_1 e sua distribuição a posteriori considerando uma priori informativa para arroz de Terras Altas, ano 2014.

Tabela 2.18: Estimadores Bayesianos usando uma priori informativa para Terras Altas para o ano de 2014

Parâmetro	Média	Erro Padrão	2.5%	97,5%
$\hat{\beta}_0$	-9,19	2,45	-14,40	-4,85
$\hat{\beta}_1$	0,05	0,01	0,02	0,08

Observa-se na tabela 2.18 que o estimador para para β_0, que é média *a posteriori*, é bem menor que a estimativa por verossimilhança e o estimador de β_1 foi maior. Pode-se observar que todos os parâmetros são significativos pelo intervalo de credibilidade das distribuições *a posteriori*.

Tabela 2.19: Classificação da avaliação sensorial de pegajosidade *versus* a classificação prevista, por meio do modelo logístico Bayesiano para Terras Altas para o ano de 2014

		Classificação prevista	
		P*	S*
Classificação real	P*	29	6
	S*	5	32
Taxa de Erro de Classificação = 15,28%			

Verifica-se na Tabela 2.19 o erro de classificação de acordo com modelo logístico Bayesiano usando a ideia de uma priori informativa considerando os escores das componentes principais das medidas de viscosidades. Nota-se que 15, 28% das observações foram classificadas erroneamente, portanto, houve um aumento brusco no erro se comparado com o modelo ajustado por verossimilhança. A principal causa desse fator é que a priori escolhida não foi adequada para esses dados e também essa taxa pode está superestimada, pois não está sendo calculada pelo método de validação cruzada.

2.4.4 Modelo Binário para arroz de Terrenos Irrigados no ano de 2014 usando priori informativa

Aqui, tomou-se como informação para construção das prioris, os dados de arroz de Terrenos Irrigados do ano de 2013. As razões de chance estimadas para 2013 são: $OR = e^{\hat{\beta}_1} = 0,987$ e $OR = e^{\hat{\beta}_2} = 1,023$. Isso retornou intervalos de confiança com comprimentos pequenos, portanto, as variâncias estimadas para os parâmetros foram $v_{\beta_1} = 0,0000158$ e $v_{\beta_1} = 0,0000321$, extremamente pequenas. Então, manteve-se a média e usou variância igual a 1.

As prioris para esse modelo foram a distribuição Normal para os parâmetros β_1 e β_2 respectivamente $N(-0,013; 0,1)$ e $N(0,023; 1)$

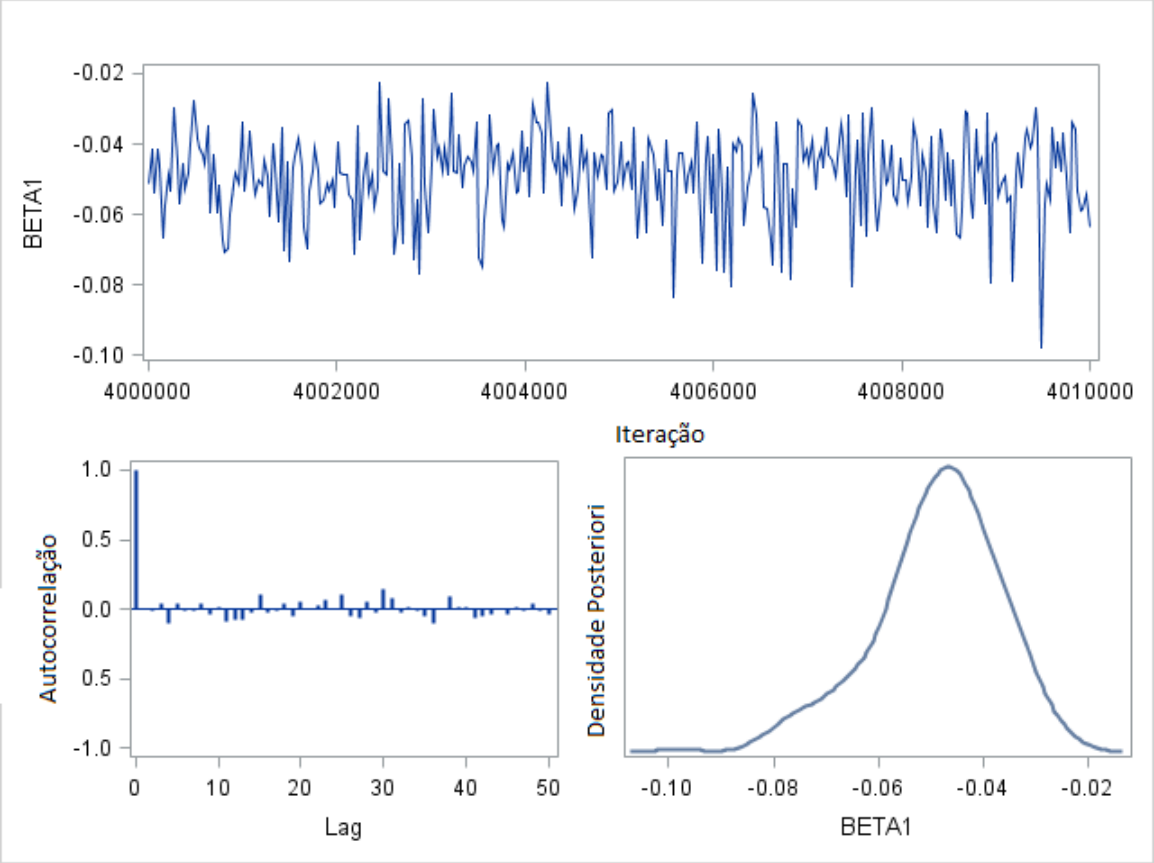

Figura 2.23: Gráficos de diagnóstico da cadeia para o parâmetro β_1 e distribuição *a posteriori* considerando uma priori informativa para arroz de Terrenos Irrigado, ano 2014.

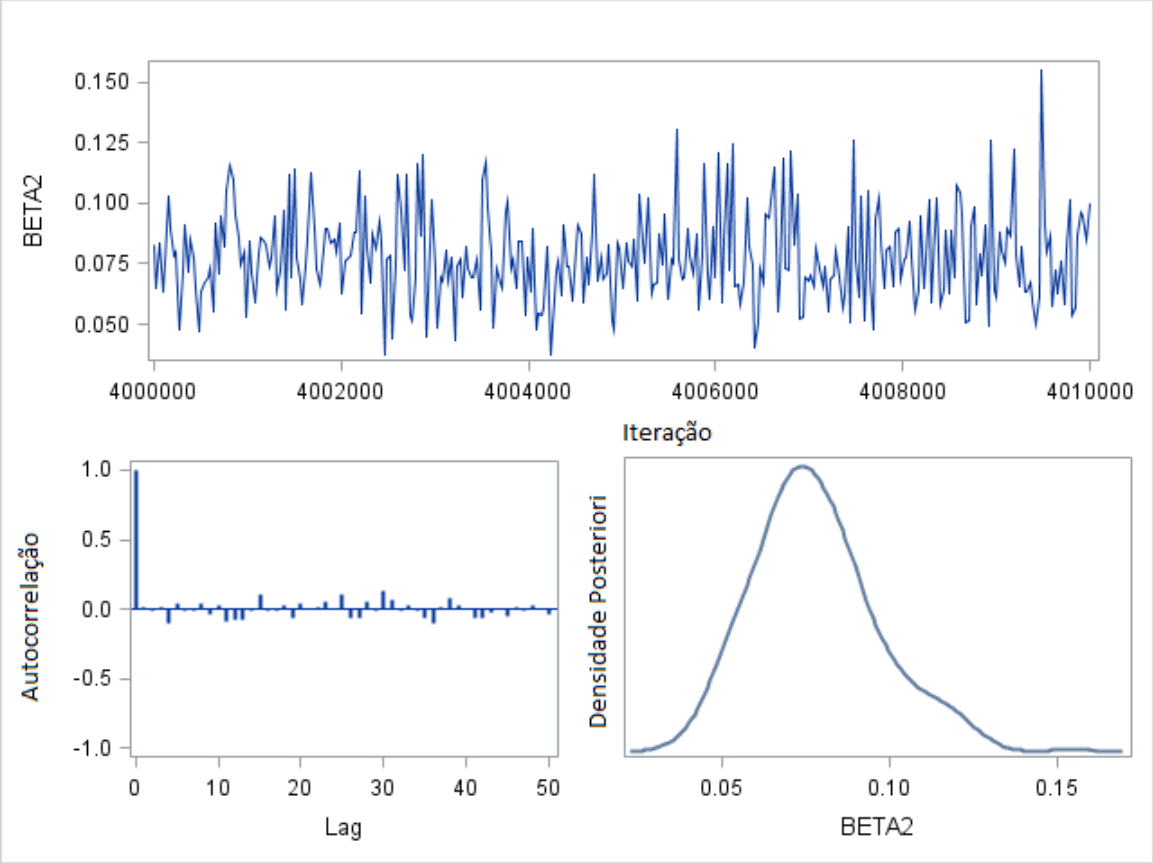

Figura 2.24: Gráficos de diagnóstico da cadeia para o parâmetro β_2 e distribuição *a posteriori* considerando uma priori informativa para arroz de Terrenos Irrigado, ano 2014.

No ajuste desse modelo optou por não considerar o intercepto, visto que esse não é significativo no modelo da Equação 2.7. Foi usado aquecimento de tamanho 4000000 e gerações a partir de 100000 amostras e um salto de 30. Esses parâmetros foram escolhidos de forma que garantisse a convergência da cadeia. Portanto, aumentou-se muito o salto, visando diminuir as autocorrelações, mas diminuiu-se o tamanho da amostra *a posteriori* gerada, tendo, $n = 336$ observações.

Há possibilidade de dizer, conforme as Figuras 2.23 e 2.24 que as cadeias convergiram. Porém, apresenta alguns picos. As autocorrelações possuem um decaimento rápido. As distribuições *a posteriori* são muito assimétricas.

Tabela 2.20: Estimadores Bayesianos usando uma priori informativa para Terrenos Irrigados para o ano de 2014

Parâmetro	Média	Erro Padrão	2.5%	97,5%
$\hat{\beta}_1$	-0,04	0,01	-0,07	-0,02
$\hat{\beta}_2$	0,07	0,01	0,04	0,1

Com base na Tabela 2.20, observa-se que as estimativas dos parâmetros são próximas dos estimados por máxima verossimilhança. Pelo intervalo de credibilidade, tem-se que os parâmetros são significativos.

Tabela 2.21: Classificação da avaliação sensorial de pegajosidade versus a classificação prevista, por meio do modelo logístico Bayesiano para Terrenos Irrigados

		Classificação prevista	
		P*	S*
Classificação real	P*	26	2
	S*	5	42
Taxa de Erro de Classificação = 9,33%			

Verifica-se na Tabela 2.21, a taxa de erro de classificação usando como função discriminante o modelo Bayesiano com uma priori informativa para Terrenos irrigados. Assim, mesmo usando informação dos dados de 2013, a taxa de erro não reduziu, permanecendo 9,33%, igual ao modelo clássico, para esse mesmo cultivar.

2.4.5 Resumo dos resultados para os modelos analisados

As Tabelas (2.22) e (2.23) apresentam os resultados de todos modelos que avaliam pegajosidade em função das medidas de viscosidade segundo o tipo de cultivar. As caselas simbolizadas pelo sinal de - indicam que o parâmetro não foi significativo naquele modelo, que a categoria em questão não foi definida para o modelo por falta de observações ou que não houve ajustamento dos dados ao modelo proposto. As colunas Binária C., Binária B.1 e Binária B.2 referem-se aos modelos binários clássicos e Bayesianos com prioris não informativas e prioris informativas, respectivamente.

Tabela 2.22: Parâmetros estimados dos modelos logísticos Binários: Clássicos e Bayesianos com priori não informativa e informativa usando medidas instrumentais de viscosidade

Ano	Tipo de terrenos	Parâmetros	Binária C.	Binária B.1	Binária B.2
2013	Terras Altas	β_0	-2,07	-	-
		β_1	0,01	-	-
		β_2	-	-	-
	Irrigados	β_0	-0,87	-	-
		β_1	-0,01	-	-
		β_2	0,02	-	-
2014	Terras Altas	β_0	-2,65	-3,65	-9,19
		β_1	-0,08	0,09	0,05
		β_2	0,02	-0,04	-
	Irrigados	β_0	0,79	0,76	-
		β_1	-0,04	-0,05	-0,04
		β_2	0,069	0,07	0,07

Tabela 2.23: Taxa de erro de classificação para todos os modelos da Tabela 2.22

Ano	Tipo de Terreno	Binária C.	Binária 1.	Binária 2.
2013	Terras Altas	31%	-	-
	Irrigados	34%	-	-
2014	Terras Altas	5,6%	5,6%	14%
	Irrigados	9,3%	9,3%	9,3%

[1] As taxas do erro de classificação fazem referências as tabelas de classificação da avaliação sensorial versus a classificação prevista pelo modelos logísticos. Onde Binária C. refere-se, aos modelos clássicos e Binária B.1 modelos Bayesianos usando priori não informativa e Binária B.2 com priori informativa

Baseado nas Tabelas 2.22 e 2.23, nota-se que não se obteve melhoras na taxas de erro de classificação da pegajosidade sensorial quando utilizou-se modelos Bayesianos. Ressalta-se que as taxas de classificação foram obtidas por validação cruzada apenas para os modelos clássicos. Observa-se que as taxas de erro permanecem iguais às previstas por meio dos modelos clássicos, mesmo quando foram usadas prioris informativas. Ressalta-se que a

classificação prevista por meio do modelo Bayesiano para Terras Altas 2014 com priori informativa resultou em uma taxa de erro maior do que a obtida pelo modelo clássico. Um problema é o fato dessa taxa não ter sido obtida por meio da técnica de validação cruzada, portanto pode estar superestimada. Alem disso, as prioris não parecem adequadas. Isso acontece porque quando há poucas observações a priori ganha maior peso nos cálculos. Quando há um grande número de observações, é dado um peso maior para a verossimilhança. Apesar se ter poucas observações nesse estudo, as prioris estabelecidas nesse problema não foram suficientemente adequadas, fazendo com que a verossimilhança recebesse mais importância. Assim, o resultado Bayesiano foi aproximadamente equivalente ao clássico.

Tabela 2.24: Resultados dos modelos logísticos Binários apresentados por Rios [16]

Ano	Tipo de Terreno	Parâmetros	Mod. Binários	Classificação
2013	Terras Altas	β_0 β_1 β_2	2,57 -0,21 -	31%
	Irrigados	β_0 β_1 β_2	-0,86 -0,01 0,02	40%
2014	Terras Altas	β_0 β_1 β_2	-2,62 -0,08 0,02	5,6%
	Irrigados	β_2 β_1 β_2	0,77 0,03 0,06	12%

Verifica-se na Tabela 2.24 os resultados da regressão logística binária e as taxas de classificação da avaliação sensorial de pegajosidade através de medidas de viscosidade para cada cultivar de arroz apresentados por Rios [16]. Nota-se os dois trabalhos utilizaram componentes principais, porém, essas foram mensuradas de formas diferentes. No entanto, os resultados são próximos de forma que para arroz de Terras Altas dos anos de 2013 e 2014 as taxas de erro de classificação foram as mesmas verificadas na Tabela 2.23. Quanto aos cultivares de Terrenos Irrigados houve uma leve redução na taxa de erro de classificação quando comparados com a Tabela 2.23.

Capítulo 3

Conclusão

As duas técnicas utilizadas nesse trabalho (regressão logística clássica e regressão logística Bayesiana) apresentaram resultados muito semelhantes para a predição da avaliação sensorial de pegajosidade para dois tipos de cultivares de arroz nos anos de 2013 e 2014.

Os modelos Bayesianos foram propostos com a intenção de melhorar a previsão da avaliação sensorial por medidas de viscosidade, visto o problema de se ter poucas observações. Porém, não trouxeram resultados melhores que os modelos estimados por verossimilhança, uma vez que foram utilizadas prioris não informativas. O principal problema de termos resultados para os modelos Bayesianos equivalentes aos modelos clássicos foi dificuldade de utilizar informações dos anos anteriores na construção de prioris do modelo. Dessa forma, a priori informativa proposta não pareceu de grande relevância nos modelos propostos.

A literatura Bayesiana indica que há uma grande vantagem em utilizar uma informação externa ao banco de dados como conhecimento prévio, mas a abordagem sobre como identificar essa informação à priori e associá-la a uma distribuição de probabilidade ainda é bem tímida e insuficiente.

A vantagem no uso da abordagem Bayesiana neste estudo foi ter obtido intervalos de credibilidade que proporcionam uma interpretação mais prática e completa.

É possível afirmar que o ajuste de modelos binários trouxe melhoras somente para avaliação sensorial de pegajosidade. Os modelos politômicos apresentam problemas devido a falta de observações em algumas categorias. Um exemplo é a classificação de dureza sensorial que apresenta uma grande concentração de dados numa única categoria e resulta em erro de classificação elevado para os modelos clássicos e Bayesianos propostos.

Logo, caso não seja viável aos técnicos da Embrapa CNPAF modificar a escala da avaliação sensorial, sugerimos modificar o processo de preparação da amostra de arroz através de algum método de cozimento que permita avaliação de todas as categorias e obtenção de suas respectivas medidas de perfil viscoamilográfico.

Os procedimentos aplicados sugerem que é possível fazer a substituição da análise sensorial por medidas de viscosidade, mas ainda precisam de ajustes e mais observações das diferentes categorias de pegajosidade e dureza do arroz.

Uma sugestão visando contornar o problema da escolha de prioris informativas é usar a

técnica denominada *Power Prior*. A ideia de *Power Prior* é usar informações passadas, como a informação a priori, nos dados atuais da análise. Além disso, o desenvolvimento dessa técnica proporciona a análise de sensibilidade para os modelos. (Ibrahim [7]).

Referências Bibliográficas

[1] AGRESTI, A. **An Introduction to Categorical Data Analysis**. 2. ed. Florida: Wiley, 2007. 47

[2] BUENO, P.D.F. **Viscoamilografia na estimativa do teor de amilose e características de consumo de arroz**. Universidade Federal de Pelotas. Pelotas, 2008. 6

[3] FITZGERALD, M. A.; McCOUCH, S. R.; HALL, R. D. **Not just a grain of rice: the quest for quality. Trends in Plant Science**. Oxford, vol. 14. n.3 p.133-139, 2009. 7

[4] FERREIRA, C.M et. al.**Qualidade do arroz no Brasil: Evolução e Padronização**. Embrapa Arroz e Feijão, 61p. Santo Antônio de Goiás, Goiânia, 2005. 3, 6

[5] GELMAN, A. **Bayesian Data Analysis**. 2. ed. FCRC Press Taylor & Francis Group, 2014. 17

[6] HOSMER, D.W. ; LEMESHOW, S. **Applied Logistic Regression**. 2. ed. John Wiley & Sons, 2000. 2, 8, 9

[7] IBRAHIM, JOSEPH G.; CHEN, MINGU-HUI**Power Prior Distributions for Regression Models**. Statistical Science 2000, Vol. 15, No. 1, 46?60 58

[8] ISO. **Sensory Analisys : Vocabulary international organization for standartization ISO 5492, 1992**. 1

[9] JOHNSON, R.A. ; WICHERN, D. W. **Applied Multivariate Statistical Analysis**. 6. ed. Prentice Hall, 2002. 11, 12, 13, 14, 17

[10] KINAS, P.G. ; ANDRADE, H.A. **Introdução à Análise Bayesiana(comR)**. 1.ed, maisQnada Editora, 2010. 18

[11] MENDONÇA, T. S. **Modelo de Regressão Logística Clássica, Bayesiana e redes neurais para modelo de Credit Scoring**.177p. Dissertação (Mestre em Estatística)- Universidade Federal de São Carlos - UFScar, São Paulo, 2008. 17, 19, 20

[12] MINGOTI, S. A. **Análise de Dados Através de Métodos de Estatística Multivariada: Uma Abordagem Aplicada**. 1. ed. Editora: UFMG, 2005, 2013-2º reimpressão. 11, 12, 14, 15, 16, 17

[13] MINIM, P. V. R. **Análise sensorial: estudos com consumidores**. 2. ed. Viçosa: UFV, 2006. 4

[14] MOREIRA, A.M. **Proposed Methodology for Quality Pre-Selection of Rice Populations**. Cereal Chemistry. 91. vol. 2. n.2 p.201-206, 2014. 6

[15] PAULINO, C.D.M. et. al. **Estatística bayesiana**. 1ed. Fundação Calouste Gulbenkian, Portugal 2003 19

[16] RIOS, E. S.**Modelos Estatísticos para Avaliação da Qualidade Culinária de Arroz: Textura e Propriedades Viscoamilográficas**. Monografia apresentado ao Departamento de Estatística da Universidade de Brasília como requisitos para obtenção do grau Bacharel em Estatística - Departamento de Estatística, Universidade de Brasília, Brasília, Julho de 2015. xiv, 1, 2, 4, 6, 7, 16, 33, 55

[17] RUGGIERO, M.A.G. ; DA ROCHA LOPES, V.L **Cálculo numérico: aspectos teóricos e computacionais**. 2 ed. Makron Books do Brasil, 1996. 9

[18] SANTOS, T.P.B. et al. **Efeito dos grãos gessados nos teores de amilose e propriedades de pasta do arroz**. Anais da 63ª Reunião Anual da SBPC. Goiânia, 2011.

[19] SAS Institute Inc. 2013 **SAS/ESTAT 12.3 User's Guide.** 20

[20] SESMAT, A.; MEULLENET, J.F. **Prediction of Rice sensory texture attributes from a single compression test, multivariate regression, and a stepwise model optimization method**. Journal of Food Science, vol. 66. n.1 p.124- 131, 2001. 5

[21] SULLIVAN, S.G. ; GREENLAND, S. **Bayesian regression in SAS software** *International Journal of Epidemiology* 2013;42:308-317. 47

[22] TEBA, C.S. et al. **Efeito dos parâmetros de extrusão sobre as propiedades de pasta de massas alimentícias pré-cozidas de arroz e feijão**. Alimentos e Nutrição Araquara. 20. vol. 3. n. 2009. 7

Apêndice A

Códigos SAS

```
################################################
####### Código SAS Proc MCMC]#########
################################################
ODS GRAPHICS ON;
PROC MCMC DATA=ATALTAS14 NBI=4000000 NMC=10000 SEED=1966 NTHIN=10;
  PARMS BETA0 0 BETA1 0 BETA2 0;
  PRIOR BETA0 ~ UNIFORM(-14.5,14.5);
  PRIOR BETA1 ~ UNIFORM(-5.5,5.5);
  PRIOR BETA2 ~ UNIFORM(-4.5,4.5);
  P = LOGISTIC(BETA0 + BETA1*COL1+BETA2*COL2);
  MODEL PEGAJB ~ BINARY(P);
RUN;
```

A PROC MCMC utiliza como padrão para obter a amostra a posteriori o algorítimo de Metropolis-Hastings.

Estão descritos abaixo argumentos que são usados no procedimento MCMC:

- **P=logistic**- É a identificação do modelo, a opção *logistic* indica a função de ligação do modelo logístico.
- **Parms**-Especifica os valores iniciais para cada parâmetro.
- **Prios**-Especifica os parâmetros e sua distribuição a priori.
- **Model**-Especifica a variável resposta e sua distribuição, no caso da regressão logística binária, pode-se utilizar as opções *binomial* ou *binary*, conforme a disposição dos dados.
- **NMC**-Especifica o número de iterações MCMC, excluindo as $burn-in$ iterações, o $default$=1000.
- **NBI**-Especifica o número de iterações de $burn-in$, o $default = 1000$.
- **Nthin**-Especifica o tamanho do salto.

- **Semente**-Especifica a semente aleatória para simulação. Pois caso não estabeleça uma semente sempre que refazer o procedimento será gerado uma distribuição a posteriori diferente.