Universidade Federal de São Carlos
Centro de Ciências Exatas e de Tecnologia
Departamento de Estatística

Modelo de Regressão Logística com Mistura de Distribuições - Estimadores de Máxima Verossimilhança e Bayesiano Utilizando o *Stan*

Luis Roberto Ferreira Junior

Trabalho de Conclusão de Curso

Universidade Federal de São Carlos
Centro de Ciências Exatas e de Tecnologia
Departamento de Estatística

Modelo de Regressão Logística com Mistura de Distribuições - Estimadores de Máxima Verossimilhança e Bayesiano Utilizando o *Stan*

Luis Roberto Ferreira Junior
Orientador(a): Prof. Dr. Luis Aparecido Milan

Trabalho de Conclusão de Curso apresentado ao Departamento de Estatística da Universidade Federal de São Carlos - DEs-UFSCar, como parte dos requisitos para obtenção do título de Bacharel em Estatística.

São Carlos
02 de Fevereiro de 2024

Luis Roberto Ferreira Junior

Modelo de Regressão Logística com Mistura de Distribuições - Estimadores de Máxima Verossimilhança e Bayesiano Utilizando o Stan

Este exemplar corresponde à redação final do trabalho de conclusão de curso devidamente corrigido e defendido por Luis Roberto Ferreira Junior e aprovado pela banca examinadora.

Aprovado em 02 de fevereiro de 2024.

Banca Examinadora:

- Prof. Dr. Luis Aparecido Milan (Orientador)
- Prof.ª Dra. Daiane Aparecida Zuanetti
- Prof. Dr. Alexsandro Giacomo Grimbert Gallo

Resumo

Neste trabalho foi explorado o modelo de regressão logística aplicado em um conjunto de dados reais a partir de duas perspectivas: realizando o ajuste do modelo logístico utilizando o estimador de máxima verosimilhança; e realizando o ajuste do modelo por meio de uma abordagem bayesiana com mistura de distribuições para $K = 1$ e 2 componentes, utilizando o Método de Monte Carlo Hamiltoniano, implementado no software *Stan* por meio do pacote *brms* do R. Realizamos a comparação entre os coeficientes estimados pelos modelos ajustados pelo estimador de máxima verossimilhança e o modelo bayesiano para $K = 1$ componente, em que observamos grande semelhança entre eles. No modelo logístico com mistura de $K = 2$ componentes, o modelo não convergiu. No estudo de simulação realizado, simulamos dados considerando as duas variações da mistura ($K = 1$ e $K = 2$). Na primeira variação ($K = 1$), o modelo convergiu apenas para $K = 1$ e não para $K = 2$. Concluímos, por saber a natureza dos dados simulados, que houve não identificabilidade dos parâmetros devido ao sobre ajuste. Nos dados simulados considerando $K = 2$, o modelo de mistura para $K = 2$ convergiu. Assim, após contatar o problema de não identificabilidade devido ao sobre ajuste, concluímos que o modelo mais adequado para o conjunto de dados reais analisado é o modelo para uma componente $K = 1$.

Palavras-chave: *Regressão Logística, Modelo de Mistura, Dados de Marketing, Instituição Financeira, Inferência Bayesiana, Stan.*

Lista de Figuras

Lista de Tabelas

Sumário

Capítulo 1

Introdução

A modelagem de uma variável categórica é um dos problemas mais comuns do mundo real, praticamente todos os setores da ciência, do mercado e da sociedade como um todo possuem problemas desse tipo.

Na área da saúde, pode haver o interesse em prever fatalidade em pacientes feridos ou prever a possibilidade de uma pessoa apresentar determinada doença. Por exemplo, doenças como diabetes e doenças cardíacas podem ser previstas com base em variáveis como idade, sexo, peso e fatores genéticos.

Na política, pode haver o interesse em prever eleições. Por exemplo, um candidato republicano será eleito nos Estados Unidos? Essas previsões podem ser feitas utilizando variáveis como idade, sexo, local de residência, posição social e padrões de votação anteriores para produzir uma previsão de voto.

No setor de marketing, pode haver o interesse em prever a possibilidade do cliente prospectado realizar uma assinatura de um determinado serviço ou fazer a compra de um determinado produto. Assim, com esse tipo de estudo, a empresa em questão pode utilizar os resultados para aperfeiçoar o direcionamento das campanhas de marketing, direcionando seus esforços para clientes que têm mais chances de se tornarem seus clientes de fato.

No setor financeiro, por exemplo, pode haver o interesse de uma empresa de cartão de crédito em prever a probabilidade de um cliente não realizar seus pagamentos dentro do prazo estipulado, ou seja, se tornar inadimplente. Essa tarefa vem se tornando cada vez mais uma tarefa importante para que bancos e outras instituições financeiras possam tomar as melhores decisões na hora de decidir para quem irão oferecer seus serviços de empréstimo, por exemplo. Hoje em dia praticamente todas essas empresas utilizam modelos estatísticos para estudar essa e outras questões e estão sempre em busca de métodos mais precisos para mapear perfis de clientes e melhorar cada vez mais suas previsões, com o intuito de buscar sempre analisar o perfil de cada cliente estatisticamente e buscar conceder crédito a clientes que têm mais chance de pagar suas dívidas em dia e evitar os clientes que têm tendências a se tornarem inadimplentes.

Para resolver esses tipos de problemas, podemos utilizar modelos estatísticos para descrever a relação entre a variável resposta de interesse (dependente) e uma ou mais variáveis explicativas (independentes). Um dos modelos mais utilizados para tal é a regressão logística que, segundo Agresti (2002), é o modelo mais importante para dados de resposta categórica. No modelo logístico, a variável resposta de interesse é categórica e geralmente binária, assumindo portanto, dois valores, que são geralmente tratados como "sucesso" e "fracasso" conforme o estudo de caso em questão.

1.1 Objetivos

O objetivo deste trabalho é abordar o modelo de regressão logística utilizando os estimadores de máxima verossimilhança e o bayesiano. Nesta abordagem pretendemos atingir os seguintes objetivos:

- Testar o estimador de máxima verossimilhança utilizando a função *glm()*, nativa do R.
- Testar o estimador bayesiano disponibilizado pelo software *Stan* implementado no R por meio do pacote *brms*.
- Comparar os estimadores de máxima verossimilhança e bayesiano obtidos pelas ferramentas R e Stan, respectivamente.
- Testar o ajuste obtido pelo método implementado no *Stan* utilizando um simulador de dados para modelos de mistura em que as componentes seguem o modelo logístico ($K = 1$ e $K = 2$), este simulador foi desenvolvido especificamente para esta finalidade.
- Aplicar as metodologias citadas em um conjunto de dados real. Neste conjunto de dados, a variável resposta indica se um cliente de uma instituição financeira realizou a compra do produto "depósito a prazo", oferecido pela mesma por uma campanha de marketing realizada por meio de ligações telefônicas.

Capítulo 2

Metodologia

Neste capítulo serão apresentados os métodos utilizados no trabalho.

2.1 Análise Descritiva de Dados

Para Reis (2002), a Análise Descritiva é a fase inicial do processo de estudo dos dados coletados. São utilizados métodos de estatística descritiva para organizar, resumir e descrever os aspectos importantes de um conjunto de características observadas ou comparar tais características entre dois ou mais conjuntos.

Esse tipo de análise proporciona uma visão do comportamento geral do conjunto de dados em relação ao objetivo do estudo. Desta forma consegue-se um entendimento básico dos dados e permite também avaliar se existem empecilhos na realização de análises posteriores, dando margem ao pesquisador para realizar mudanças pertinentes, evitando conclusões equivocadas.

A descrição dos dados também tem como objetivo identificar anomalias, até mesmo resultante do registro incorreto de valores, e dados dispersos, aqueles que não seguem a tendência geral do restante do conjunto.

Não só nos artigos técnicos direcionados para pesquisadores, mas também nos artigos de jornais e revistas escritos para o público leigo, é cada vez mais frequente a utilização destes recursos de descrição para complementar a apresentação de um fato, justificar ou referendar um argumento.

2.2 Regressão Logística

Segundo Gonzalez (2002), a regressão logística é uma técnica estatística de modelagem que tem como objetivo construir, por meio de um conjunto de observações, um modelo que permita realizar a predição de valores de uma variável categórica, geralmente binária, em função de uma ou mais variáveis explicativas quantitativas ou qualitativas. Então, a partir desse modelo gerado é possível calcular ou prever a probabilidade de um evento ocorrer, dado uma observação aleatória.

A variável resposta Y na regressão logística é frequentemente binária, logo, nestes casos ela segue a distribuição de Bernoulli, tendo uma probabilidade de sucesso desconhecida p.

Portanto, a variável resposta Y é definida como

$$Y = \begin{cases} 1, & \text{se ocorrer sucesso;} \\ 0, & \text{se ocorrer fracasso.} \end{cases}$$

A probabilidade de sucesso é $0 \leq p \leq 1$ e a probabilidade de fracasso é $q = 1 - p$.

Na regressão logística, é feita a estimação da probabilidade desconhecida p, dada uma combinação linear de variáveis independentes.

2.2.1 Função Logit

Na Seção 2.2, foi dito que a variável resposta binária na regressão logística possui distribuição de Bernoulli, portanto é necessário conectar as variáveis explicativas à essa distribuição na variável resposta. A função mais comum para satisfazer essa necessidade é chamada de logit. Assim como na maioria dos problemas de distribuição de Bernoulli, na regressão logística nós não conhecemos a probabilidade de sucesso p. Assim, o objetivo do modelo logístico é estimar p para uma combinação linear das variáveis explicativas do modelo (Gonzalez, 2002).

Para ligar essa combinação linear à distribuição de Bernoulli, é necessário uma função que as una, ou mapeie a combinação linear de variáveis que poderia retornar qualquer valor em uma distribuição de probabilidade de Bernoulli com domínio de 0 a 1. A função logit nada mais é que o logaritmo natural da razão de probabilidades, que é chamada de chance ou *odds* em inglês. A função de ligação logit é dada por

$$logit(p) = ln(odds) = ln\left(\frac{p}{1-p}\right), \tag{2.1}$$

e está representada graficamente na Figura 2.1.

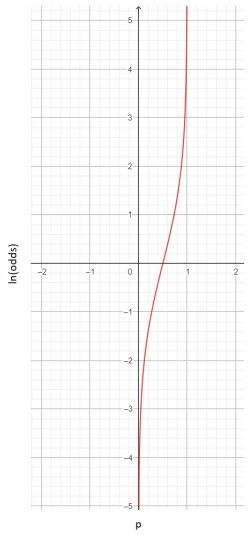

Figura 2.1: Comportamento da função logit(p).

Ao analisar o comportamento da função logit nota-se que existem assíntotas verticais nos pontos $p = 0$ e $p = 1$, ou seja, quando p tende a 0 o valor da função é $-\infty$ (menos infinito) e quando p tende a 1 o valor da função é ∞ (infinito), ou seja, a função só está definida quando $0 \leq p \leq 1$, que é exatamente o que buscamos, pois a probabilidade também varia somente nesse intervalo. Dessa forma, podemos ter certeza que, para valores de $p < 0$ e $p > 1$ a função não está definida, isto é, pela função logit não é possível obter uma probabilidade inferior a 0% ou superior a 100% .

Como o nosso interesse é estimar a probabilidade da variável dependente Y ser igual a 1 ($Y = 1$), dado uma combinação linear das variáveis independentes, devemos ter as probabilidades no eixo y do gráfico. Para isso, basta inverter a função logito descrita acima. Portanto, a partir da Equação (2.1), temos que

$$logit^{-1}(\alpha) = \frac{1}{1+e^{-\alpha}} = \frac{e^{\alpha}}{1+e^{\alpha}}, \tag{2.2}$$

em que α é a combinação linear das variáveis independentes e seus coeficientes.

A Figura 2.2 exibe a representação gráfica da função inversa da função logit(p). Portanto, agora temos uma função de ligação com domínio no intervalo de 0 a 1, assim como precisávamos para estimar a probabilidade do evento de interesse $Y = 1$, dado uma combinação linear de variáveis independentes.

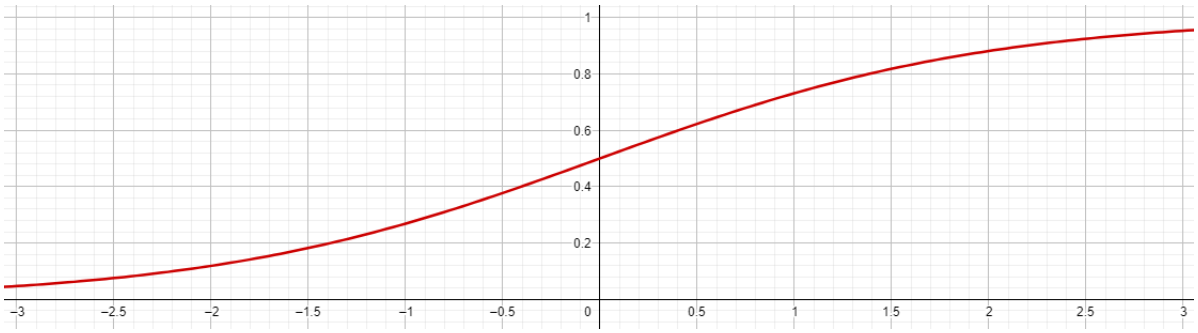

Figura 2.2: Comportamento da função inversa do logit(p).

2.2.2 Regressão Logística Simples

Segundo Gonzalez (2002), a regressão logística simples consiste em um modelo utilizado para modelar uma variável dependente binária Y com base em uma combinação linear de uma única variável independente. A forma geral do modelo de regressão logística simples é dada por

$$g(x) = \boldsymbol{\beta X} = \beta_0 + \beta_1 x_1, \tag{2.3}$$

em que $\boldsymbol{\beta} = (\beta_0, \beta_1)$ é o vetor de coeficientes e x_1 é a única variável independente.

Igualando a função de ligação logit apresentada na Equação (2.1) à função $g(x)$ apresentada na Equação (2.3), temos que

$$ln\left(\frac{p}{1-p}\right) = \boldsymbol{\beta X} = \beta_0 + \beta_1 x_1. \tag{2.4}$$

Como o objetivo do modelo logístico é estimar p, devemos isolar p utilizando a função exponencial, de maneira que

$$\frac{p}{1-p} = e^{\beta_0 + \beta_1 x_1}, \tag{2.5}$$

e, posteriormente,

$$\hat{p} = \frac{e^{\hat{\beta}_0 + \hat{\beta}_1 x_1}}{1 + e^{\hat{\beta}_0 + \hat{\beta}_1 x_1}} = \frac{1}{1 + e^{-(\hat{\beta}_0 + \hat{\beta}_1 x_1)}}. \tag{2.6}$$

A Equação (2.6) é chamada de regressão estimada e é, essencialmente, a função que representa o objetivo do modelo de regressão logística, haja vista que $\hat{p}$ é a probabilidade estimada para quaisquer valores de coeficientes e variáveis que venhamos a inserir nesta equação. O vetor de coeficientes $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \hat{\beta}_1)$ é obtido pelo método de estimação de máxima verossimilhança descrito na subseção seguinte.

2.2.3 Estimação dos coeficientes do modelo

Como visto em Gonzalez (2002), para ajustar um modelo de regressão logística é necessário estimar os parâmetros do modelo. Para isso, utiliza-se o método de estimação por máxima verossimilhança. A partir do conjunto de observações, este método irá calcular os estimadores dos parâmetros β_0 e β_1, que serão denotados aqui por $\hat{\beta}_0$ e $\hat{\beta}_1$, que maximizam a função de máxima verossimilhança. Em outras palavras, o método de estimação por máxima verossimilhança permite encontrar os estimadores dos parâmetros do modelo que têm maior probabilidade de replicar o padrão de observações dos dados da amostra.

Seja $\boldsymbol{\beta} = (\beta_0, \beta_1)$ o vetor de coeficientes, e sejam as probabilidades $P(Y_i = 1|x_i) = \pi(x_i)$ e $P(Y_i = 0|x_i) = 1 - \pi(x_i)$. Então, para os pares (x_i, y_i) tais que $y_i = 1$, a contribuição para a função de verossimilhança é $\pi(x_i)$, e para os pares (x_i, y_i) tais que $y_i = 0$, a contribuição para a função de verossimilhança é $1 - \pi(x_i)$, onde $\pi(x_i)$ denota o valor de $\pi(x)$ avaliado em x_i.

A função de verossimilhança é dada por

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} \pi(x_i)^{y_i} [1 - \pi(x_i)]^{1-y_i}. \tag{2.7}$$

Para obter a função log-verossimilhança, aplica-se o logaritmo natural em ambos os lados. Assim, temos que

$$ln[L(\boldsymbol{\beta})] = l(\boldsymbol{\beta}) = \sum_{i=1}^{n} [y_i ln[\pi(x_i)] + (1 - y_i) ln[1 - \pi(x_i)]] . \tag{2.8}$$

A quantidade $\boldsymbol{\beta}$ que maximiza $l(\boldsymbol{\beta})$ é obtida após derivar $l(\boldsymbol{\beta})$ em relação aos parâmetros $\boldsymbol{\beta} = (\beta_0, \beta_1)$:

$$\frac{\partial l(\boldsymbol{\beta})}{\partial \beta_0} = \sum_{i=1}^{n} [y_i - \pi(x_i)] \tag{2.9}$$

e

$$\frac{\partial l(\boldsymbol{\beta})}{\partial \beta_1} = \sum_{i=1}^{n} x_i [y_i - \pi(x_i)] . \tag{2.10}$$

Os estimadores de (β_0, β_1) denotados por $(\hat{\beta}_0, \hat{\beta}_1)$, são as soluções das Equações (2.9) e (2.10) quando igualadas a 0. Estes estimadores medem a taxa de variação do logit para uma unidade de variação na variável independente, isto significa que eles são de fato, a inclinação da linha de regressão entre a variável dependente y_i e a sua variável

independente x_i.

Como estas equações são não-lineares nos parâmetros, é necessário a utilização de um procedimento iterativo para resolvê-las, geralmente o método de Newton-Raphson. Este método escolhe, sucessivamente, novos conjuntos de parâmetros que produzam maiores log-verossimilhanças e melhores ajustes aos dados observados. O processo continua através de iterações até atingir a maximização da função log-verossimilhança.

2.2.4 Regressão Logística Múltipla

Como visto em Gonzalez (2002), na regressão logística múltipla, ao invés de termos apenas uma variável independente, temos um vetor de k variáveis independentes. A forma geral do modelo de regressão logística múltipla é dada por

$$g(x) = \boldsymbol{\beta X} = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k, \tag{2.11}$$

em que $\boldsymbol{\beta} = (\beta_0, \beta_1, ..., \beta_k)$ é o vetor de coeficientes e $\boldsymbol{X} = (x_1, ..., x_k)^T$ é o vetor de variáveis independentes.

Analogamente à regressão logística univariada, igualamos a função logito à função $g(x)$ descrita na Equação (2.11) e aplicamos a função exponencial, conforme a Equação (2.12), de modo que

$$\frac{p}{1-p} = e^{\beta_0 + \beta_1 x_1 + ... + \beta_k x_k}. \tag{2.12}$$

Posteriormente, seguimos com o procedimento para isolar p, conforme a Equação (2.13).

$$\hat{p} = \frac{e^{\hat{\beta_0} + \hat{\beta_1} x_1 + ... + \hat{\beta_k} x_k}}{1 + e^{\hat{\beta_0} + \hat{\beta_1} x_1 + ... + \hat{\beta_k} x_k}} = \frac{1}{1 + e^{-(\hat{\beta_0} + \hat{\beta_1} x_1 + ... + \hat{\beta_k} x_k)}}. \tag{2.13}$$

A função de verossimilhança é a mesma da Equação (2.7), com a diferença de que $\pi(x_i)$ é dado como $\pi(i)$, em função da Equação (2.11), representando o conjunto de variáveis independentes em $g(x)$ e seus respectivos coeficientes. Portanto, a função log-verossimilhança é dada por

$$ln[L(\boldsymbol{\beta})] = l(\boldsymbol{\beta}) = \sum_{i=1}^{n} \left[y_i ln[\pi(i)] + (1 - y_i) ln[1 - \pi(i)] \right]. \tag{2.14}$$

As expressões das equações a partir das derivadas parciais são dadas pelas Equações (2.15) e (2.16).

$$\frac{\partial l(\boldsymbol{\beta})}{\partial \beta_0} = \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} \hat{\pi}_i = 0 \qquad \text{e} \tag{2.15}$$

$$\frac{\partial l(\boldsymbol{\beta})}{\partial \beta_j} = \sum_{i=1}^{n} x_{ij} y_i - \sum_{i=1}^{n} x_{ij} \hat{\pi}_i = 0, \tag{2.16}$$

para j $\in (1, ..., k)$.

2.3 Abordagem Bayesiana

A abordagem bayesiana para a estimação dos parâmetros em uma regressão logística envolve a aplicação do Teorema de Bayes para obter a distribuição a posteriori dos parâmetros, dada a distribuição a priori e os dados observados. Vamos descrever isso matematicamente.

Suponhamos que temos um modelo de regressão logística com k preditores (ou variáveis independentes) e n observações. Os parâmetros do modelo consistem em um vetor de intercepto β_0 e um vetor de coeficientes de inclinação $\beta = (\beta_1, \beta_2, ..., \beta_k)$, tal como em (2.11).

A função de probabilidade para a regressão logística é dada por

$$P(Y = 1|\mathbf{X}, \beta) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + ... + \beta_k x_k)}} \tag{2.17}$$

onde Y é a variável de resposta binária, $\mathbf{X}$ é o vetor de preditores e β é o vetor de parâmetros.

Especificamos uma distribuição a priori para os parâmetros, denotada por $P(\beta)$. Esta distribuição expressa as crenças iniciais ou conhecimentos prévios sobre os valores dos parâmetros.

O teorema de Bayes é utilizado para calcular a distribuição a posteriori dos parâmetros:

$$P(\beta|\mathbf{X}, \mathbf{Y}) \propto P(\mathbf{Y}|\beta, \mathbf{X}) P(\beta) \tag{2.18}$$

em que $P(\mathbf{Y}|\beta, \mathbf{X})$ é a verossimilhança e $P(\beta)$ é a distribuição a priori.

A distribuição a posteriori $P(\beta|\mathbf{X}, \mathbf{Y})$ é explorada para realizar inferências sobre os parâmetros. Pode-se calcular estimativas pontuais (como a média a posteriori) e intervalos de credibilidade.

Em termos matemáticos, a implementação prática desse processo pode envolver métodos computacionais, como a amostragem de Monte Carlo, especialmente quando a solução analítica não é possível. Ferramentas como o software *Stan* são frequentemente empregadas para realizar esses cálculos de forma eficiente e precisa.

2.3.1 Software *Stan*

O Stan é uma plataforma para modelagem estatística, sobretudo bayesiana, e computação estatística de alta performance, desenvolvida em C++ e que possui interface com as linguagens de programação R, Python, shell, MATLAB, Julia e Stata.

De acordo com Wikipédia, o *Stan* implementa múltiplos algoritmos *Markov Chain Monte Carlo* para realizar modelagens bayesianas, como o *Hamiltonian Monte Carlo (HMC)* e *No-U-Turn sampler (NUTS)*, que é uma variação do HMC.

Segundo o Manual de Referência oficial do Stan, esta plataforma também conta com métodos para codificação de modelos de probabilidade, algoritmos de inferência para ajustar modelos e fazer previsões, e ferramentas de análise posterior para avaliar os resultados.

Neste trabalho, será utilizado o pacote *brms* do *R* que implementa o *Stan* para realizar todos os ajustes de modelos, gráficos e análise dos resultados da abordagem bayesiana do modelo de regressão logística estudado.

2.3.2 Algoritmos *Markov Chain Monte Carlo (MCMC)*

Assim como foi dito na Seção 2.3.1, o *Stan* utiliza dois algoritmos de *Markov Chain Monte Carlo (MCMC)* para realizar o ajuste dos modelos bayesianos, o Método de Monte Carlo Hamiltoniano (HMC) e sua variante adaptativa *No-U-Turn sampler (NUTS)*.

Nesta Seção, será dada uma breve introdução a esses algoritmos juntamente com detalhes de sua implementação e configuração.

Monte Carlo Hamiltoniano (HMC)

Segundo o Manual de Referência do Stan, o Monte Carlo Hamiltoniano (HMC) é um método de *Markov Chain Monte Carlo (MCMC)* que usa derivadas da função de densidade sendo amostrada para gerar transições eficientes abrangendo a distribuição a priori. Este algoritmo utiliza uma simulação dinâmica hamiltoniana aproximada baseada na integração numérica que é então corrigida executando uma etapa de aceitação Metropolis.

Função densidade alvo

O objetivo da amostragem é extrair uma amostra aleatória de uma função densidade de probabilidade $p(\theta)$ para os parâmetros θ, geralmente a distribuição a posteriori $p(\theta|y)$ dado os dados y codificada como um programa Stan.

Variável auxiliar de momento

O HMC incorpora variáveis auxiliares de momento ρ e extrai de uma densidade conjunta

$$p(\rho, \theta) = p(\rho|\theta)p(\theta).$$

Na maioria das aplicações do HMC, incluindo o *Stan*, a variável auxiliar tem distribuição Normal Multivariada que não depende dos parâmetros θ, de maneira que

$$\rho \sim \text{MultiNormal}(0, M),$$

em que M é chamada de "matriz de massa", uma matriz simétrica, positiva definida e conhecida como métrica Euclidiana. Segundo Radford (2011), essa matriz é tipicamente diagonal, e é frequentemente um múltiplo escalar da matriz identidade. Pode ser vista como uma transformação do espaço de parâmetros que torna a amostragem mais eficiente.

Por padrão, o *Stan* define M^{-1} como uma estimativa diagonal da covariância calculada durante o aquecimento (*warmup*).

O Hamiltoniano

A densidade conjunta $p(\rho, \theta)$ define um Hamiltoniano

$$\begin{aligned} H(\rho, \theta) &= -log p(\rho, \theta) \\ &= -log p(\rho|\theta) - log p(\theta) \\ &= T(\rho|\theta) + V(\theta) \end{aligned}$$

em que o termo

$$T(\rho|\theta) = -logp(\rho|\theta)$$

é chamado de "energia cinética" e o termo

$$V(\theta) = -logp(\theta)$$

é chamado de "energia pontencial". A energia potencial é especificada pelo programa *Stan* através da definição de uma log-densidade.

Gerando transições

Iniciando do valor atual dos parâmetros θ, uma transição para um novo estado é gerada em duas etapas antes de ser submetida a uma etapa de aceitação Metropolis, que será explicado posteriormente.

Primeiro, um valor para o momento é extraído independentemente dos valores atuais dos parâmetros,

$$\rho \sim \text{MultiNormal}(0, M).$$

Dessa forma, o momento não persiste ao longo das iterações.

Em seguida, o sistema conjunto (θ, ρ), composto pelos valores atuais dos parâmetros θ e o novo momento ρ é evoluído pelas equações de Hamilton,

$$\frac{d\theta}{dt} = +\frac{\partial H}{\partial \rho} = +\frac{\partial T}{\partial \rho}$$

$$\frac{d\rho}{dt} - \frac{\partial H}{\partial \theta} = -\frac{\partial T}{\partial \theta} - \frac{\partial V}{\partial \theta}.$$

Com a densidade do momento sendo independente da densidade alvo, isto é, $p(\rho|\theta) = p(\rho)$, o primeiro termo na derivada do tempo do momento $\frac{\partial T}{\partial \theta}$ é zero, produzindo o par de derivadas do tempo

$$\frac{d\theta}{dt} = +\frac{\partial T}{\partial \rho}$$

$$\frac{d\rho}{dt} = -\frac{\partial V}{\partial \theta}.$$

Integrador *Leapfrog*

A Seção anterior deixa uma equação diferencial de dois estados para resolver. O *Stan*, como a maioria das outras implementações do HMC, usa o integrador *Leapfrog*, que é um algoritmo de integração numérica especificamente adaptado para fornecer resultados estáveis para sistemas de equações hamiltonianos.

Como a maioria dos integradores numéricos, o algoritmo de salto realiza etapas discretas de algum pequeno intervalo de tempo ϵ. O algoritmo *Leapfrog* começa extraindo um novo termo do momento independentemente dos valores dos parâmetros θ ou do valor do momento anterior

$$\rho \sim \text{MultiNormal}(0, M).$$

Em seguida, ele alterna atualizações de meio passo do momento e atualizações de passo completo da posição.

$$\begin{aligned} \rho &\longleftarrow \rho - \frac{\epsilon}{2}\frac{\partial V}{\partial \theta} \\ \theta &\longleftarrow \theta + \epsilon M^{-1}\rho \\ \rho &\longleftarrow \rho - \frac{\epsilon}{2}\frac{\partial V}{\partial \theta}. \end{aligned}$$

Aplicando L etapas *Leapfrog*, um total de $L\epsilon$ de tempo é simulado. O estado resultante no final da simulação (L repetições dos três passos acima) será denotado por (ρ^*, θ^*).

O erro do integrador *Leapfrog* é da ordem de ϵ^3 por passo e ϵ^2 globalmente, e ϵ é o intervalo de tempo (também conhecido como tamanho do passo).

Passo de aceitação Metropolis

Se o integrador *Leapfrog* fosse numericamente perfeito, não haveria necessidade de fazer mais nenhuma aleatorização por transição do que gerar um vetor de momento aleatório. Em vez disso, o que é feito na prática para contabilizar erros numéricos durante a integração é aplicar um passo de aceitação Metropolis, onde a probabilidade de manter a proposta (ρ^*, θ^*) gerada pela transição de (ρ, θ) é

$$\min(1,\ \exp(H(\rho, \theta) - H(\rho^*, \theta^*))).$$

Caso a proposta não seja aceita, o valor do parâmetro anterior é retornado para o próximo sorteio e utilizado para inicializar a próxima iteração.

Resumo do algoritmo

De forma resumida, o Monte Carlo Hamiltoniano começa em um conjunto inicial especificado de parâmetros θ; no *Stan*, esse valor é especificado pelo usuário ou gerado aleatoriamente. Então, para um determinado número de iterações, um novo vetor de momento é amostrado e o valor atual do parâmetro θ é atualizado usando o integrador *Leapfrog* com tempo de discretização ϵ e número de passos L de acordo com a dinâmica hamiltoniana. Em seguida, um passo de aceitação Metropolis é aplicado e é tomada a decisão de atualizar para o novo estado (ρ^*, θ^*) ou manter o estado existente.

Capítulo 3

Modelo de mistura de distribuições

Segundo Erlandson (2009), modelos de mistura de distribuições são utilizados para modelar fenômenos ou experimentos cujas observações são provenientes de uma população composta por K subpopulações, em que K pode ser conhecido ou desconhecido. A utilização de tais modelos permite o pesquisador representar a existência de K subpopulações sem que seja necessário as mesmas estarem explícitas no conjunto de dados. Ou seja, a técnica pode ser aplicada em um conjunto de dados que não necessariamente identifica a subpopulação à qual certa unidade experimental pertence. Além disso, modelos com mistura de distribuições fornecem uma forma conveniente de modelar dados que podem não ser adequadamente modelados por qualquer família paramétrica de distribuições padrão.

Segundo Chauveau (2018), a primeira aplicação conhecida dos modelos de mistura de distribuições foi realizada em 1894, quando Karl Pearson analisou dados morfométricos de uma população de caranguejos, disponibilizados pelo zoólogo Walter F. Weldon. Nesse estudo Pearson identificou, graficamente, a evidência da mistura e abordou o problema por meio do método de momentos ajustando um modelo com mistura de duas distribuições normais, escolhendo os cinco parâmetros da mistura de forma que os momentos empíricos correspondessem aos do modelo. Segundo Wikipedia, Pearson obteve êxito em identificar a existência de duas potenciais espécies diferentes nessa população de caranguejos. Tal informação era desconhecida no conjunto de dados.

Podemos observar na Figura 3.1 uma representação gráfica de um modelo de mistura.

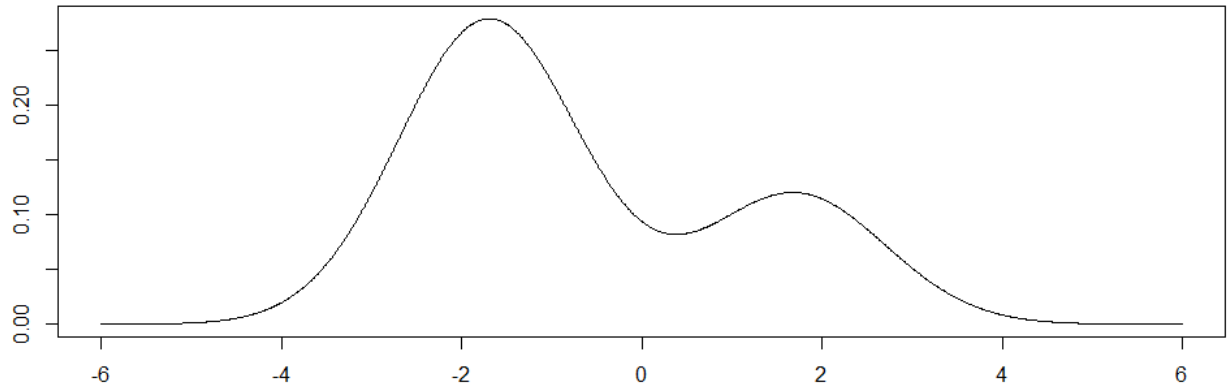

Figura 3.1: Exemplo de mistura de duas distribuições Normais.

Seguindo a notação de Meira (2014), o modelo de mistura considerando o número de componentes, K, conhecido pode ser expresso por

$$\begin{aligned} f(y) &= w_1 f_1(y) + ... + w_K f_K(y) \\ &= \sum_{k=1}^{K} w_k f_k(x), \end{aligned} \tag{3.1}$$

em que

- $f(y)$ é a função de densidade de probabilidade da mistura;
- K é o número de componentes na mistura;
- w_k é a probabilidade de pertencer ao componente k; $w_k > 0$ e $\sum_{k=1}^{K} w_k = 1$;
- $f_k(y)$ é a função densidade de probabilidade do componente k.

Considere uma família de variáveis aleatórias discretas independentes $S = \{S_t; t = 1, ..., T\}$, S_t com distribuição multinomial, $S_t \sim mult(1, \boldsymbol{p} = (p_1, ..., P_k))$, com K resultados possíveis e $\sum_{k=1}^{K} p_k = 1$. Considere também uma família de variáveis aleatórias $Y = \{Y_t; t = 1, ..., T\}$, sendo que a distribuição de cada variável aleatória Y_t é controlada pelo valor assumido pela variável aleatória S_t correspondente, ou seja, $Y_t|S_t = k, \theta$ tem distribuição condicional $Pr(Y_t = y_t|S_t = s_t, \theta)$ se Y_t for discreta e por $f(y_t|S_t = s_t, \theta)$ se Y_t for absolutamente

contínua em que $\theta = (\theta_1, ..., \theta_k)$ é o vetor de parâmetros com $\theta \in R^k$, e θ_k é o parâmetro associado a k-ésima componente da mistura para $k = 1, ..., K$.

Complementando a notação, temos que $\boldsymbol{s}$ é a sequência de valores assumidos pela família S, $\boldsymbol{s} = \{s_t; t = 1, ..., T\}$ e $\boldsymbol{y}$ é a sequência de valores assumidos pela família Y, $\boldsymbol{y} = \{y_t; t = 1, ..., T\}$.

A distribuição marginal de Y_t no caso discreto é dada por

$$Pr(Y_t = y_t|\boldsymbol{\theta}) = \sum_{k=1}^{K} Pr(Y_t = y_t|S_t = k, \theta)Pr(S_t = k|p) \tag{3.2}$$

em que $Pr(Y_t = y_t|S_t = k, \theta)$ é a função de probabilidade condicional. A distribuição marginal de Y_t no caso contínuo

$$f_{Yt}(y|\boldsymbol{\theta}) = \sum_{k=1}^{K} f_{Yt}(y|S_t = k, \theta)Pr(S_t = k|p) \tag{3.3}$$

em que $f_{Yt}(y|S_t = k, \theta)$ é a função de densidade de probabilidade condicional. Note que em ambos os casos a distribuição de Y_t é uma mistura de K componentes.

Capítulo 4

Estudo de simulação

Nesta seção realizaremos um estudo de simulação utilizando um simulador de dados (desenvolvido especificamente para este trabalho) para mistura de modelos logísticos para testar o método implementado no software *Stan*.

Avaliaremos duas situações:

- Simulação de dados provenientes de uma única distribuição ($K = 1$ ou não mistura) e ajuste do modelo logístico sem considerar a mistura ($K = 1$) e, posteriormente, ajuste do modelo logístico considerando a mistura ($K = 2$).
- Simulação de dados provenientes de duas distribuições diferentes ($K = 2$ ou mistura com duas componentes) e ajuste do modelo logístico considerando a mistura de duas componentes ($K = 2$).

Dessa forma, o estudo avaliará três modelos logísticos diferentes. Todos os modelos serão ajustados utilizando o método implementado no *Stan* por meio do pacote *brms* do R. O simulador utilizado para gerar todos os dados está disponível no Apêndice A.

4.1 Dados de apenas uma componente ($K = 1$)

Inicialmente, simularemos 1000 valores de uma variável binária S para identificação da componente (componente 1 ou 2) atrelada a cada uma das observações. Neste caso, queremos 1000 observações provenientes de uma única distribuição ($K = 1$), ou seja, a variável S será um vetor com 1000 uns, indicando que todas as 1000 observações são amostradas de uma mesma distribuição, atrelada ao único componente na mistura ($K = 1$ ou não mistura).

A seguir, definiremos os parâmetros para a curva logística. Precisamos definir dois parâmetros: o ponto médio da curva logística, isto é, o valor de x em que $P(Y_i = 1|X = x) = 0.5$; e a declividade da curva logística. Para mais detalhes, ver Wikipedia (2024).

Nós optamos por escolher o par de parâmetros (2, 1.2) como sendo o ponto médio e a declividade da curva, respectivamente. A matriz de parâmetros está disposta na Tabela 4.1.

Tabela 4.1: Parâmetros para a curva logística ($K = 1$).

Ponto médio	Declividade da curva
2	1.2

Para fins de ilustração, a Figura 4.1 mostra o comportamento da curva parametrizada com os valores que definimos na Tabela 4.1.

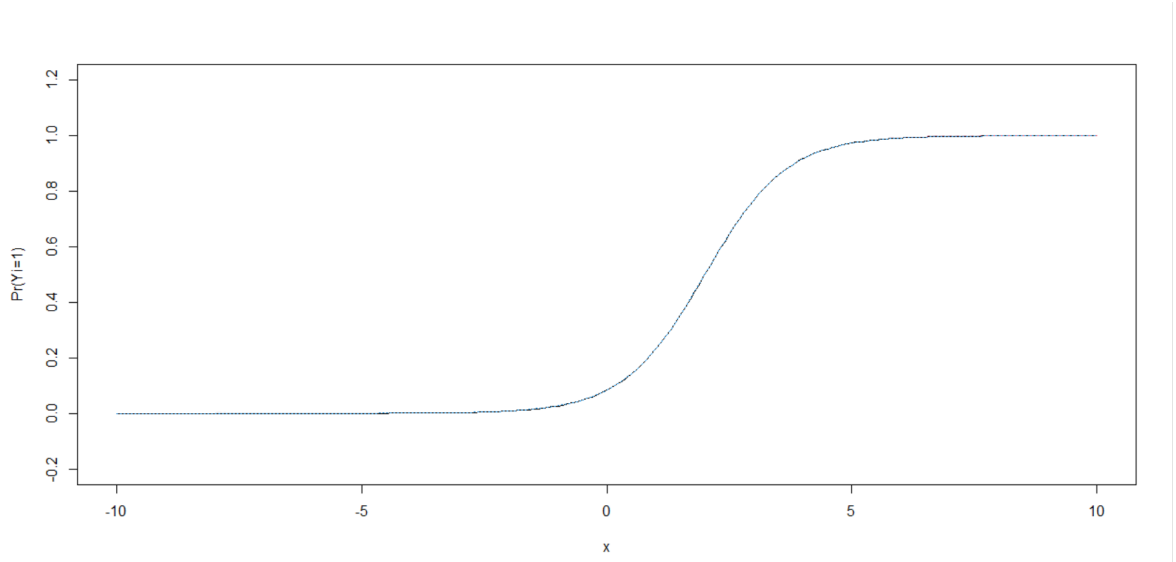

Figura 4.1: Curva logística com ponto médio em $X = 2$ e declividade 1.2.

Considerando a variável S e os parâmetros da curva logística definidos na Tabela 4.1, utilizamos o simulador para gerar o banco de dados e as cinco primeiras linhas do mesmo estão apresentadas na Tabela 4.2.

Tabela 4.2: Cinco primeiras linhas do banco de dados simulado considerando $K = 1$.

S	X	$P(Y_i = 1 \mid X = x)$	Y
1	2.3368347	0.599696720	1
1	2.5377152	0.655941136	1
1	0.3177644	0.117255270	0
1	2.2780161	0.582639766	0
1	-3.5063466	0.001348227	0

Agora, realizaremos o ajuste do modelo logístico pelo *Stan*, considerando apenas uma componente ($K = 1$ ou não mistura). Este modelo pode ser ajustado utilizando o código a seguir.

```
modelo_simulacaok1k1 = brm(formula = y ~ lM,
                           family = bernoulli(),
                           data = data_simulacao,
                           chains = 3,
                           cores = 3,
                           iter = 2000,
                           warmup = 1000,
                           seed = 123)
```

Com o modelo ajustado, avaliaremos a convergência do mesmo por meio da análise gráfica dos *traceplots*.

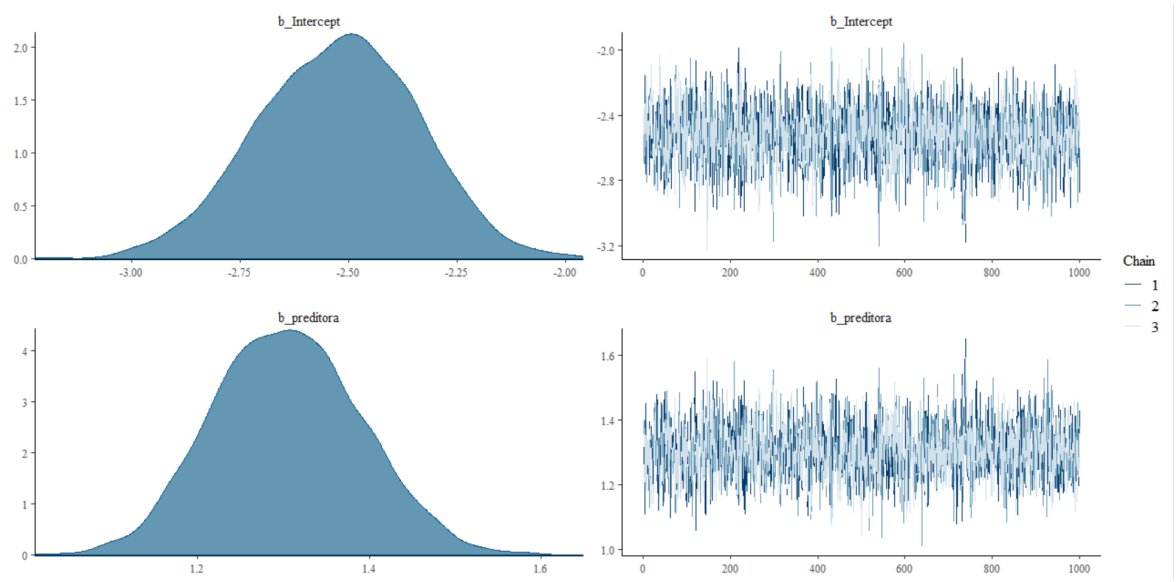

Figura 4.2: Densidades das distribuições a posteriori dos parâmetros e traceplots das Cadeias de Markov para o banco de dados e modelo logístico, ambos considerando $K = 1$.

Analisando os *traceplots* à direita da Figura 4.2, podemos concluir que o modelo que considera apenas uma componente ($K = 1$ ou não mistura) convergiu para o conjunto de dados simulado proveniente de apenas uma componente.

Agora, utilizaremos o mesmo banco de dados para realizar o ajuste do modelo logístico considerando duas componentes na mistura. Podemos fazer isso introduzindo a função *mixture(bernoulli, bernoulli)* do pacote *brms*, no argumento *family* no código utilizado anteriormente. Portanto, o código para o ajuste do modelo de mistura com duas componentes ($K = 2$) é dado por

```
modelo_simulacaok1k2 = brm(formula = y ~ 1M,
                           family = mixture(bernoulli, bernoulli),
                           data = data_simulacao,
                           chains = 3,
                           cores = 3,
                           iter = 2000,
                           warmup = 1000,
                           seed = 123)
```

Com o modelo de mistura de duas componentes ajustado, avaliaremos a convergência do mesmo por meio dos traceplots.

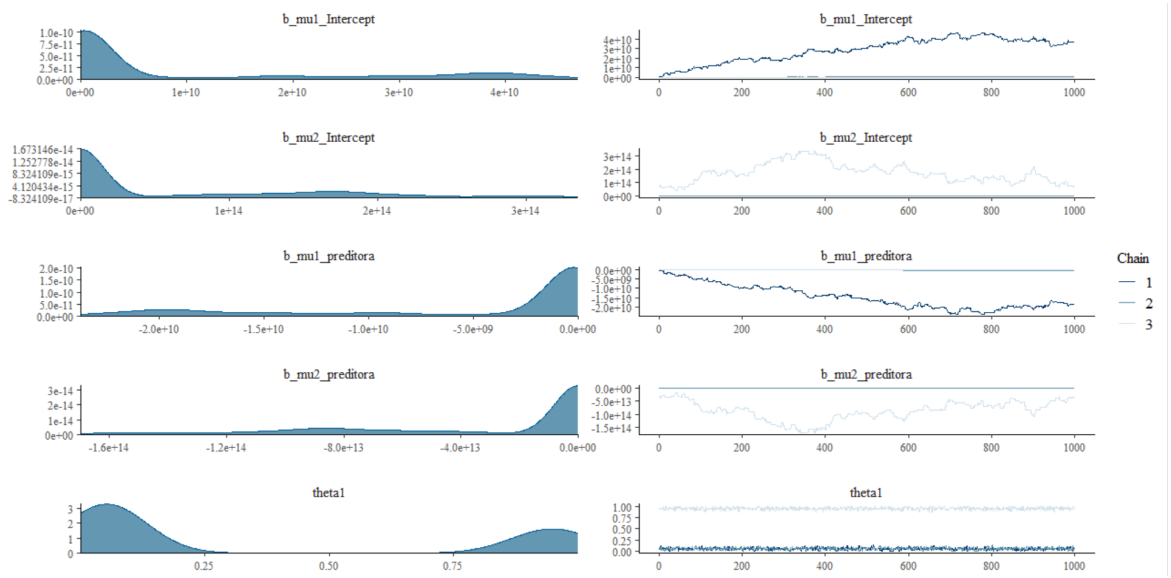

Figura 4.3: Densidades das distribuições a posteriori dos parâmetros e traceplots das Cadeias de Markov para o banco de dados considerando $K = 1$ e o modelo de mistura logístico considerando $K = 2$.

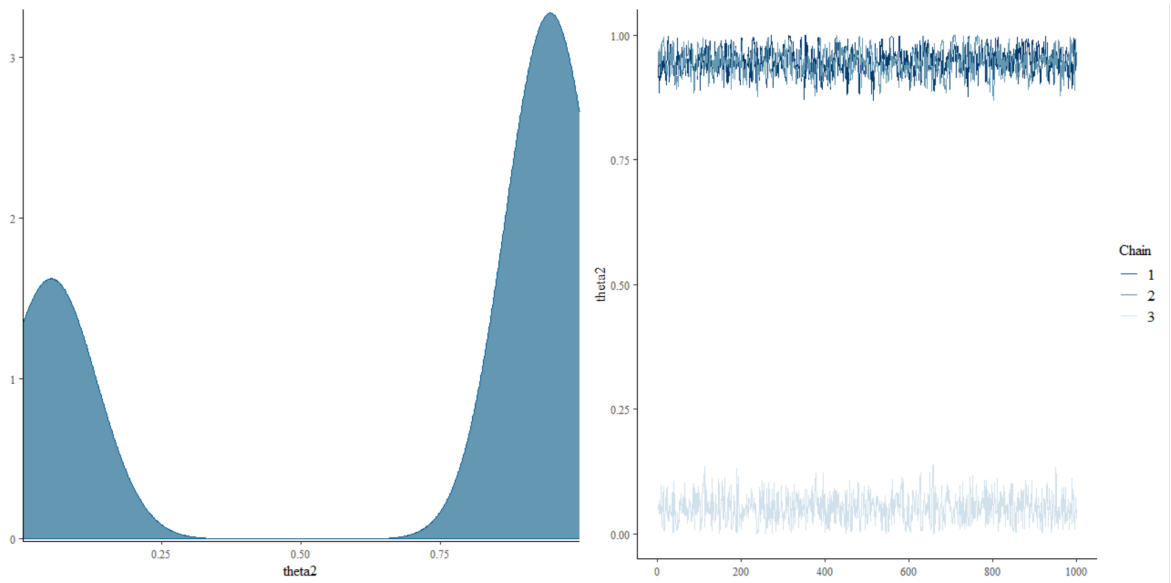

Figura 4.4: Densidades das distribuições a posteriori dos parâmetros e traceplots das Cadeias de Markov para o banco de dados considerando $K = 1$ e o modelo de mistura logístico considerando $K = 2$.

Analisando os *traceplots* à direita das Figuras 4.3 e 4.4, temos clara evidência de não convergência do modelo de mistura de duas componentes considerando o banco de dados simulados provenientes de uma única distribuição. Uma explicação plausível para este resultado é a não identificabilidade dos parâmetros (para mais detalhes, ver Frühwirth-

Schnatter (2010) apud Macerau (2023) p. 24). Como simulamos um banco de dados proveniente de uma única distribuição ($K = 1$), ao tentar realizar o ajuste do modelo de mistura para duas componentes ($K = 2$) ocorre o problema da não identificabilidade. Na prática, significa que o modelo de mistura de duas componentes ($K = 2$) não é adequado para os dados provenientes de apenas uma distribuição, o que causa a não convergência do método MCMC implementado (HMC).

4.2 Dados com mistura de duas componentes ($K = 2$)

Nesta seção, iremos simular um banco de dados com 1000 observações provenientes de duas distribuições, isto é, duas componentes na mistura ($K = 2$).

Analogamente ao que foi feito na Seção 4.1, simularemos 1000 valores de uma variável binária S para identificação da componente (componente 1 ou 2) atrelada a cada uma das observações. Porém, dessa vez, queremos 1000 observações de duas distribuições diferentes ($K = 2$), ou seja, geraremos valores 1 ou 2 para a variável S, indicando à qual componente da mistura de duas distribuições cada uma das observações pertence.

Optamos por utilizar as probabilidades (70%, 30%) para gerar os valores da variável S. Portanto, esses valores são os verdadeiros valores dos pesos w_1 e w_2 do modelo de mistura de duas componentes ($K = 2$) descrito em (3.1).

Tabela 4.3: Número de observações geradas por componente.

	Nº de observações
Componente 1	677
Componente 2	323

A Tabela 4.3 apresenta a proporção obtida de observações provenientes das componentes 1 e 2, respectivamente. Declaramos a proporção (70%, 30%) para gerar os valores de S e obtivemos a proporção de 67.7% e 32.3% no banco de dados.

A seguir, adicionaremos o par de parâmetros (-2, 0.8) referentes ao ponto médio e a declividade da curva logística pertencente à segunda componente, respectivamente, na segunda linha da matriz exibida na Tabela 4.1. Dessa forma, a primeira linha da matriz continua sendo referente à primeira componente e a segunda linha referente à segunda componente. Assim, a matriz de parâmetros utilizados na simulação está disposta na Tabela 4.4.

Tabela 4.4: Parâmetros para a curva logística ($K = 2$).

Ponto médio	Declividade da curva
2	1.2
-2	0.8

Para fins de ilustração, a Figura 4.5 mostra o comportamento das duas curvas logísticas parametrizadas com os valores que definimos na Tabela 4.4 e a curva que representa a mistura das curvas das duas componentes.

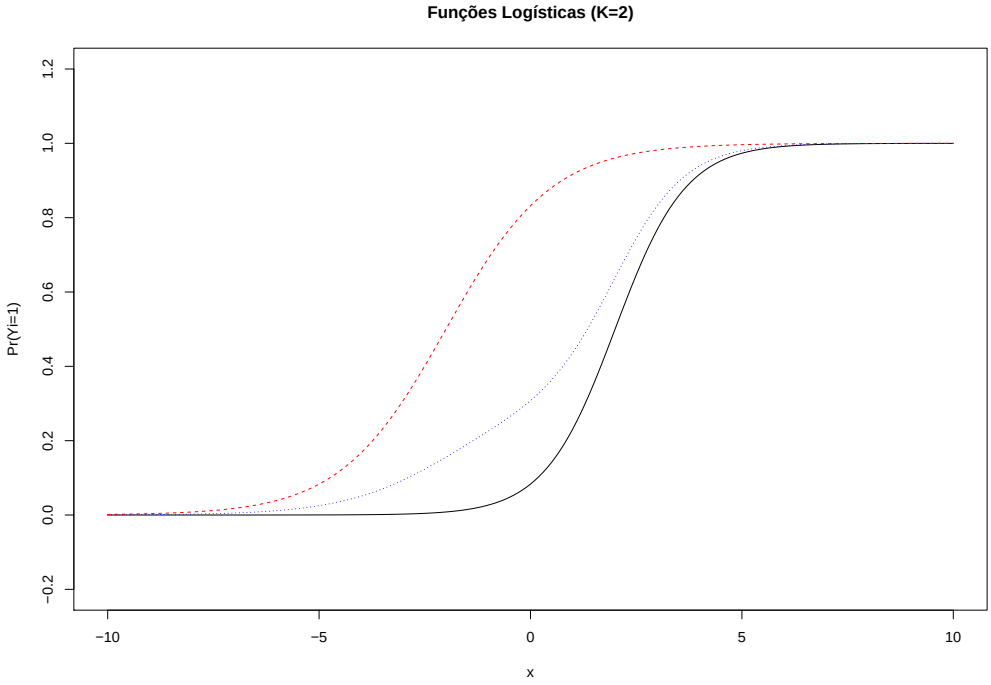

Figura 4.5: Curvas logísticas para as componentes 1 e 2 (utilizando a parametrização da Tabela 4.4) e curva logística que representa a mistura das duas componentes.

Na Figura 4.5, a curva preta representa a curva logística da primeira componente, cujos parâmetros de ponto médio e declividade estão na primeira linha da Tabela 4.4. A curva vermelha representa a curva logística da segunda componente, cujos parâmetros estão na segunda linha da Tabela 4.4 e a curva azul representa a mistura das duas componentes, ou seja, a média ponderada das duas curvas, utilizando os pesos (70%, 30%) que definimos anteriormente para as componentes 1 e 2, respectivamente. Note que a curva preta, que representa a primeira componente, tem uma declividade mais acentuada, isto é, possui um crescimento mais acelerado quando comparada à curva vermelha, que representa a

segunda componente. Este comportamento é definido pelos parâmetros de declividade que definimos anteriormente como 1.2 e 0.8 para as componentes 1 e 2, respectivamente.

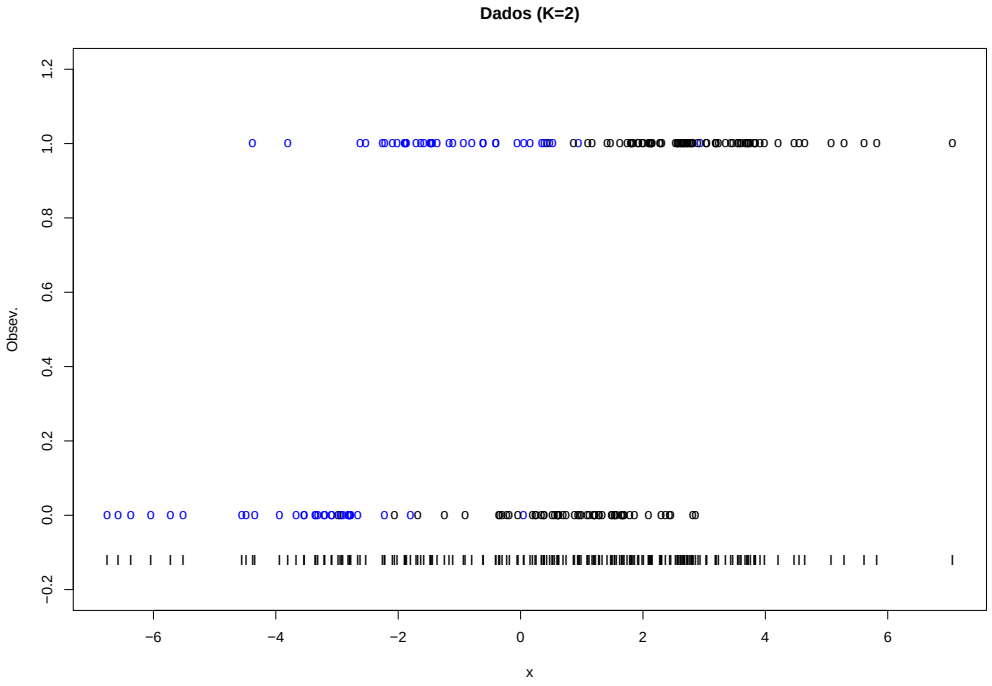

Figura 4.6: Observações simuladas atreladas à sua respectiva componente.

Na Figura 4.6, temos os valores da variável X, gerada pelo simulador de dados, atrelados à sua respectiva componente da mistura. Representado pelos pontos pretos, temos os valores de X pertencentes à componente 1 da mistura, cuja média é próxima de 2, que é o valor do ponto médio que definimos como parâmetro da curva logística da componente 1. Em azul, temos os valores de X pertencentes à componente 2 da mistura, cuja média é próxima de -2, ponto médio definido anteriormente para esta componente. Para construir este gráfico, utilizamos apenas 100 valores de X, em vez de 1000, por motivos de melhor visualização. Note que os pontos em preto estão mais concentrados em torno da média 2 que os pontos azuis em torno da média -2. Isso se dá pela diferença na declividade das duas curvas ilustradas na Figura 4.5. Como a curva preta, que representa a primeira componente tem um parâmetro de declividade maior que a curva vermelha, que representa a segunda componente, os valores de X da primeira componente ficam mais concentrados em torno da média da curva quando comparados com os valores de X da segunda componente.

Considerando a variável S e os parâmetros das curvas logísticas definidos na Tabela 4.4, utilizamos o simulador para gerar o banco de dados e as cinco primeiras linhas do

mesmo estão apresentadas na Tabela 4.5.

Tabela 4.5: Cinco primeiras linhas do banco de dados simulado considerando $K = 2$.

S	X	$P(Y_i = 1 \mid X = x)$	Y
1	2.8219044	0.599696720	1
2	-2.6616048	0.974173094	1
1	0.4935655	0.117255270	0
1	2.4444802	0.582639766	1
1	2.7781284	0.596366067	0

A seguir, realizaremos o ajuste do modelo logístico pelo *Stan*, considerando duas componentes na mistura ($K = 2$) utilizando o banco de dados simulado proveniente de duas distribuições ($K = 2$ ou duas componentes). Este modelo pode ser ajustado utilizando o código a seguir.

```
modelo_simulacaok2k2 = brm(formula = y ~ lM,
                           family = mixture(bernoulli, bernoulli),
                           data = data_simulacao1,
                           chains = 3,
                           cores = 3,
                           iter = 2000,
                           warmup = 1000,
                           seed = 123)
```

Com o modelo ajustado, avaliaremos a convergência do mesmo por meio da análise gráfica dos *traceplots*.

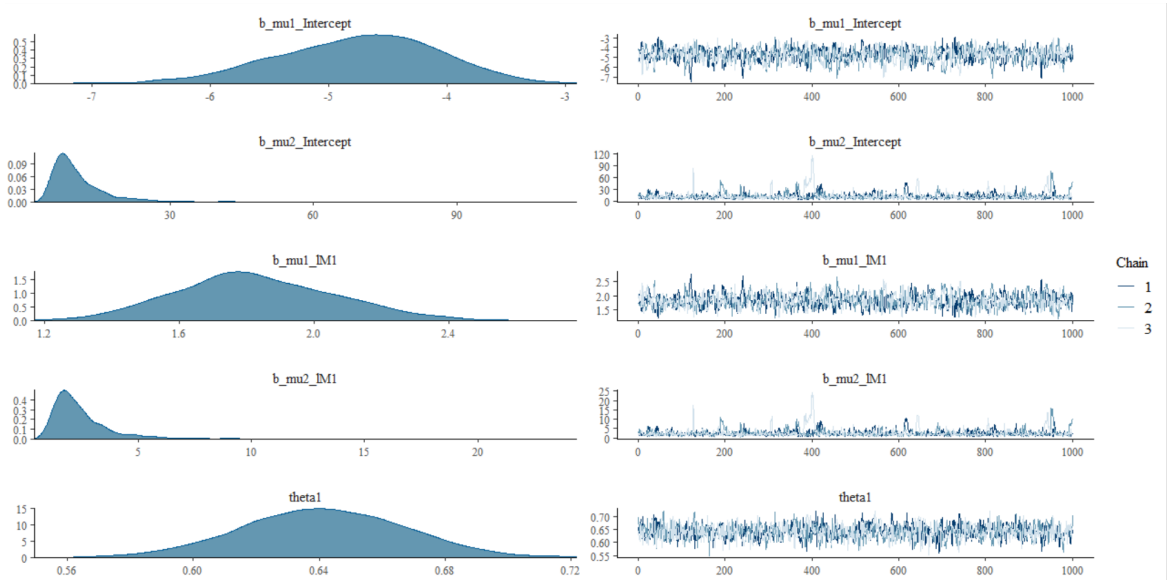

Figura 4.7: Densidades das distribuições a posteriori dos parâmetros e traceplots das Cadeias de Markov para o banco de dados considerando $K = 2$ e o modelo de mistura logístico considerando $K = 2$.

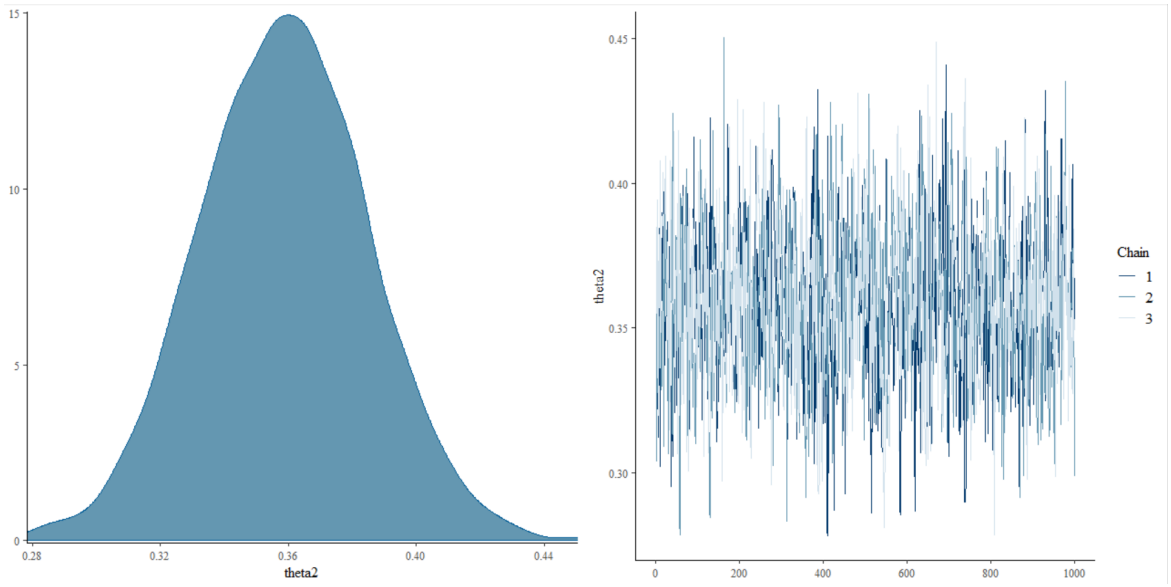

Figura 4.8: Densidades das distribuições a posteriori dos parâmetros e traceplots das Cadeias de Markov para o banco de dados considerando $K = 2$ e o modelo de mistura logístico considerando $K = 2$.

Analisando os *traceplots* à direita das Figuras 4.7 e 4.8, podemos dizer que o modelo de mistura de duas componentes, dessa vez utilizando dados provenientes de duas distribuições ($K = 2$ ou duas componentes), convergiu, diferentemente do modelo de mistura para $K = 2$ que ajustamos na Seção 4.1. Estes dois resultados em conjunto evidenciam

que quando tentamos ajustar um modelo de mistura considerando duas componentes ($K = 2$) utilizando um banco de dados simulados provenientes de apenas uma distribuição ($K = 1$), enfrentamos problemas em relação à convergência do modelo, por conta da não identificabilidade dos parâmetros, conforme citamos na Seção 4.1. Por outro lado, quando ajustamos este mesmo modelo, porém usando dados simulados adequados, provenientes de duas distribuições ($K = 2$), o modelo converge.

Para finalizar o estudo, iremos analisar as estimativas obtidas para os pesos w_1 e w_2 do modelo de mistura para $K = 2$.

Tabela 4.6: Estimativas, erros e intervalos de credibilidade para os pesos w_1 e w_2.

Parâmetro	**Estimativa**	**Erro**	**Lim Inf IC 95%**	**Lim Sup IC 95%**
w_1	0.64	0.03	0.59	0.69
w_2	0.36	0.03	0.31	0.41

Pela Tabela 4.6 temos que os pesos w_1 e w_2 tiveram estimativas de 0.64 e 0.36, respectivamente. A proporção original das observações nas componentes 1 e 2 é de 0.7 e 0.3, respectivamente, como visto na Tabela 4.3. Assim, podemos concluir que o modelo de mistura foi capaz de produzir estimativas relativamente boas dos pesos originais.

Capítulo 5

Banco de Dados

O banco de dados que será utilizado, e que pode ser acessado em UCI Machine Learning Repository, possui informações de uma instituição financeira portuguesa, mais especificamente, os resultados obtidos pela campanha de marketing da instituição na venda de um de seus produtos, o *term deposit* (depósito a prazo), que é um tipo de investimento no qual o investidor pode sacar seu dinheiro apenas após uma data de vencimento estabelecida previamente. A campanha foi baseada em chamadas telefônicas, e em alguns casos mais de uma chamada foi realizada para o mesmo cliente.

Será padronizado o uso de 1 para "Sim" e 0 para "Não" nas variáveis dicotômicas.

Abaixo são apresentadas as variáveis disponíveis no banco de dados, que contém 5.000 observações (a base original possui aproximadamente 45.000 observações, mas por limitações computacionais, foi retirada uma amostra de tamanho 5.000 da base completa para viabilizar o estudo):

- X_1 : Idade (discreta);
- X_2 : Tipo de Trabalho (originalmente 12 categorias);
- X_3 : Estado civil (casado, divorciado, solteiro);
- X_4 : Grau de escolaridade (primário, secundário, terciário, desconhecido);
- X_5 : Variável binária que representa se o cliente é inadimplente (sim ou não);
- X_6 : Saldo anual médio (em euros);
- X_7 : Variável binária que representa se o cliente possui financiamento imobiliário (sim ou não);

- X_8 : Variável binária que representa se o cliente possui empréstimo pessoal (sim ou não);
- X_9 : Tipo de contato utilizado na campanha de marketing atual (desconhecido, telefone, celular);
- X_{10} : Dia do mês do último contato na campanha de marketing atual (1 a 31);
- X_{11} : Mês do último contato na campanha de marketing atual (janeiro a dezembro);
- X_{12} : Duração do último contato na campanha de marketing atual (em segundos);
- X_{13} : Número de chamadas feitas ao cliente na campanha de marketing atual (discreta, 0, 1, ...);
- X_{14} : Número de dias corridos desde o último contato da campanha de marketing anterior (discreta, -1 indica que não houve contato);
- X_{15} : Número de chamadas feitas ao cliente na campanha de marketing anterior (discreta, 0, 1, ...);
- X_{16} : Variável binária que representa se o cliente comprou o produto na campanha de marketing anterior (sim, não, outro ou desconhecido);
- Y : Variável binária que representa se o cliente comprou o produto, na campanha de marketing atual (sim ou não).

5.1 Análise Descritiva

Inicialmente, iremos reduzir a dimensionalidade de algumas variáveis. A variável X_{11}, que representa o mês do último contato na campanha de marketing atual, foi reduzida para trimestres, passando de 12 níveis para apenas 4.

Em seguida, serão feitas modificações na variável X_2, que representa o tipo de trabalho do cliente. Para reduzir a dimensão desta variável, utilizaremos o WOE (*weight of evidence*), um critério muito utilizado em modelos financeiros que é uma forma de medir a distância entre as categorias de Y. Ou seja, o WOE, basicamente, funciona como métrica para separar os clientes que compraram o produto na campanha de marketing atual (bons clientes) dos clientes que não compraram o produto na campanha de marketing atual (maus clientes). O WOE é calculado da seguinte forma:

$$WOE = log\left(\frac{P(\text{categoria}|Y=1)}{P(\text{categoria}|Y=0)}\right).$$

Portanto, neste caso é calculado o logaritmo da razão da proporção amostral dos clientes que compraram o produto em questão ($Y = 1$) pela proporção amostral dos clientes que não compraram o produto em questão ($Y = 0$) para cada categoria da variável X_2 (Tipo de trabalho).

A Figura 5.1 abaixo traz o cálculo do WOE para cada uma das categorias de X_2.

Tipo de Trabalho	0	1	WOE
administrativo	11,37%	11,93%	0,05
colarinho azul	22,60%	13,39%	-0,52
empresário	3,42%	2,33%	-0,38
empregada	2,83%	2,06%	-0,32
administração	20,43%	24,60%	0,19
aposentado	4,38%	9,76%	0,8
autônomo	3,49%	3,54%	0,01
serviços	9,48%	6,98%	-0,31
estudante	1,68%	5,09%	1,11
técnico	16,93%	15,88%	-0,06
desempregado	2,76%	3,82%	0,33
desconhecido	0,64%	0,64%	0,01

Figura 5.1: WOE da variável Tipo de trabalho.

Em seguida, são agrupadas as categorias com valores para o WOE próximos. Para

medir a qualidade do agrupamento, utilizaremos o IV (*information value*), que é a soma dos WOE agrupados e ponderados dado por

$$IV = \sum [P(\text{categoria}|Y = 1) - P(\text{categoria}|Y = 0)] \cdot log\left(\frac{P(\text{categoria}|Y = 1)}{P(\text{categoria}|Y = 0)}\right).$$

Dessa forma, o melhor agrupamento é aquele com o maior valor para o IV.

Pela Figura 5.1 percebe-se que as categorias *colarinho azul, empresário, empregada e serviços* possuem valores próximos, os menores observados, sugerindo uma união dessas categorias em um só grupo. Além disso, as categorias *técnico, administrativo, autônomo e desconhecido* possuem valores próximos de zero, indicando outra união. Por fim, são separadas as categorias *administração e desempregado*, com WOE similar, e *aposentado e estudante* com os maiores WOE, em que predominam $Y = 1$.

Realizando o agrupamento indicado, obtém-se os resultados apresentados na Figura 5.2

	0	1	IV
C0	38,33%	24,75%	0,0594
C1	6,05%	14,84%	0,0788
C2	32,42%	31,99%	0,0001
C3	23,19%	28,42%	0,0106
			0,1489

Figura 5.2: Tentativa 1 de agrupamento da variável X_2.

Posteriormente, foram realizadas novas tentativas de diminuir ainda mais a dimensão da variável X_2, testando a redução das 12 categorias iniciais em apenas 3. Foram testadas duas formas de agrupamento distintas, considerando os WOEs mais próximos.
A tentativa 2 consiste nos seguintes agrupamentos:

- C0 = *colarinho azul, empresário, empregada e serviços;*
- C1 = *técnico, administração, autônomo e desconhecido;*
- C2 = *administração, desempregado, aposentado e estudante.*

A tentativa 3, por sua vez, consiste nos seguintes agrupamentos:

- C0 = *colarinho azul, empresário, empregada e serviços;*
- C1 = *técnico, administrativo, autônomo, desconhecido e administração;*

- C2 = *desempregado, aposentado e estudante.*

	0	1	IV
C0	38,33%	24,75%	0,0594
C1	32,42%	31,99%	0,0001
C2	29,24%	43,26%	0,0549
			0,1144

	0	1	IV
C0	38,33%	24,75%	0,0594
C1	52,85%	56,59%	0,0026
C2	8,81%	18,66%	0,0739
			0,1359

Figura 5.3: Tentativas 2 e 3 de agrupamento da variável X_2.

Pela Figura 5.3, observa-se que nenhum agrupamento supera o IV obtido na Figura 5.2. Portanto, seguiremos com o estudo considerando o primeiro agrupamento realizado, que é descrito por:

- C0 = *colarinho azul, empresário, empregada e serviços;*;
- C1 = *técnico, administrativo, autônomo, desconhecido;*;
- C2 = *administração e desempregado*;
- C3 = *aposentado e estudante.*

Em relação à variável resposta, que possui distribuição de Bernoulli, nota-se que existe forte desbalanceamento nos dados. Pela Tabela 5.1, observamos que 87,88% dos clientes não compraram o produto e, consequentemente, apenas 12,12% dos clientes realizaram a compra.

Não	**Sim**
4394	606
87.88%	**12.12%**

Tabela 5.1: Tabela resumo dos clientes que compraram ou não o produto.

Posteriormente, é feita uma análise gráfica de cada uma das variáveis cruzadas com a variável resposta Y.

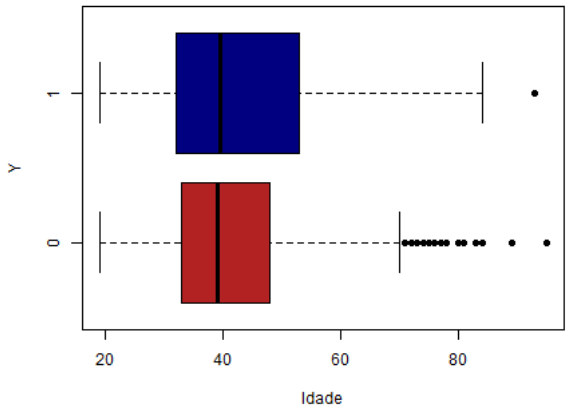

Figura 5.4: Boxplots para a variável idade.

Iniciando pela idade do cliente, percebe-se, na Figura 5.4, que as medianas estão próximas em ambos os casos, por volta de 40 anos de idade. Além disso, existe um maior número de outliers entre aqueles que não compraram o produto, indicando que talvez seja um produto atrativo para pessoas mais velhas, ou seja, a idade pode ser uma variável importante para prever se um cliente comprará ou não o produto.

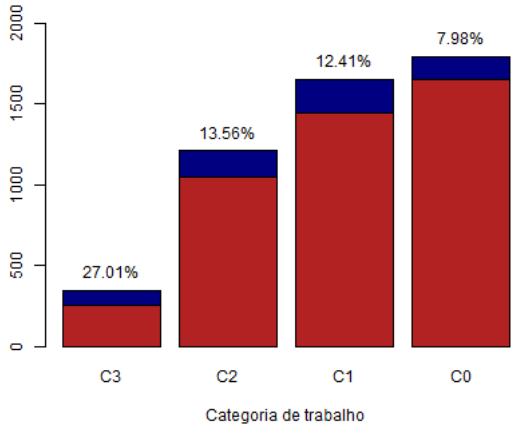

Figura 5.5: Gráfico de barras para a categoria de trabalho.

Em relação à variável tipo de trabalho do cliente, que passou pelo tratamento de agrupamento, observa-se, na Figura 5.5, que a categoria C3 (estudantes e aposentados) é a classe com a maior proporção de clientes que compraram o produto, com cerca de 27% de compradores, seguida pelas categorias C2 (gerentes e desempregados) com cerca de 13,56%, C1 (técnicos, administradores, autônomos e aqueles cuja profissão é desconhecida)

com 12,41% e C0 (profissionais de colarinho azul, empreendedores, donas de casa e pessoas que trabalham com serviços) com quase 8%. Diante disso, esta variável também parece ser importante para discriminar os compradores dos não compradores.

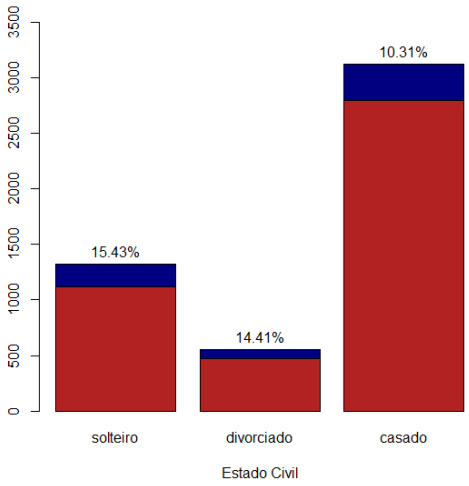

Figura 5.6: Gráfico de barras para o estado civil.

Em relação ao estado civil do cliente, pela Figura 5.6, percebe-se que a proporção de compradores é maior dentre os solteiros quando comparados com divorciados ou casados, porém, de forma geral, não há grande diferença entre as proporções.

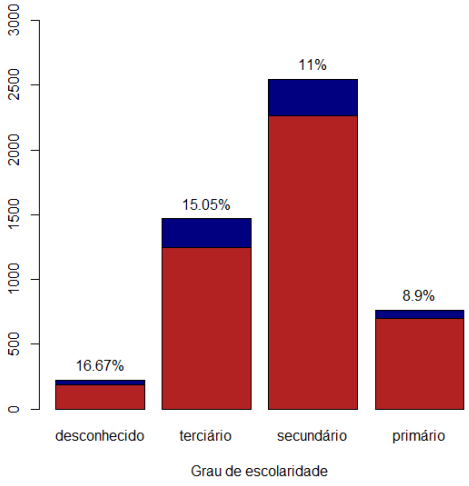

Figura 5.7: Gráfico de barras para o grau de escolaridade.

Em termos do grau de escolaridade do cliente, percebe-se que conforme o grau aumenta, aumenta razoavelmente a proporção daqueles que compraram, embora essa pro-

porção seja maior para aqueles com grau de escolaridade desconhecido, o que não é informativo.

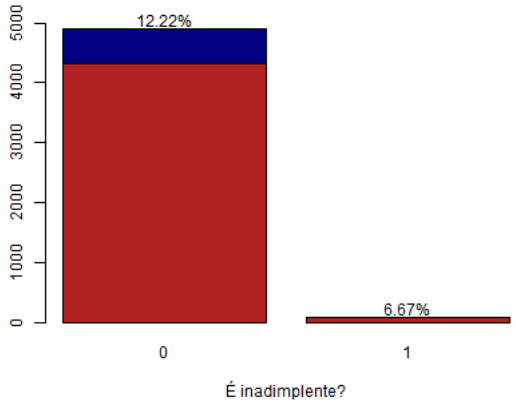

Figura 5.8: Gráfico de barras para a variável que indica se o cliente é inadimplente ou não.

A respeito da variável que denota se o cliente é inadimplente ou não, percebe-se que há um fortíssimo desbalanceamento nos dados, com poucos clientes inadimplentes. Nota-se, também, que a proporção de compradores dentre os adimplentes é quase duas vezes maior em relação à proporção de compradores dentre os inadimplentes.

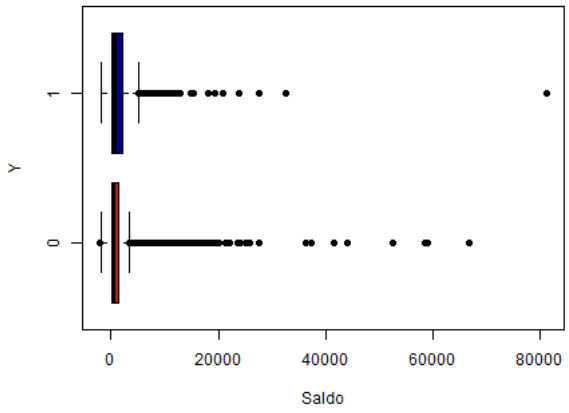

Figura 5.9: Boxplots para a variável saldo anual médio.

Em relação ao saldo anual médio dos clientes, podemos observar que a amplitude da distribuição dos clientes que compraram o produto é levemente maior que a dos não compradores. A mediana dos dois grupos é próxima e existem muitos pontos discrepantes

em ambos, principalmente no grupo de clientes que não compraram o produto. De forma geral, esta variável parece ser pouco informativa.

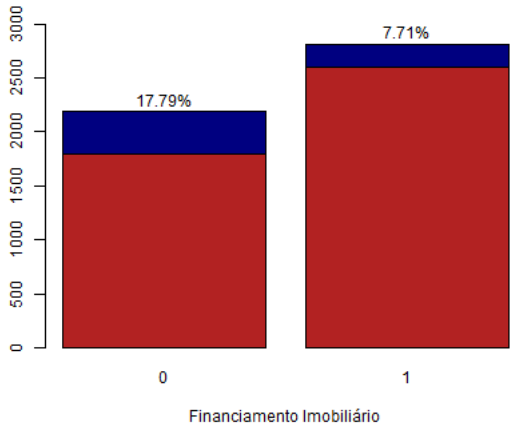

Figura 5.10: Gráfico de barras para a variável financiamento imobiliário.

A partir da Figura 5.10, percebe-se que, proporcionalmente, mais clientes que não possuem financiamento imobiliário compraram o produto em relação aos que possuem esse tipo de financiamento. Tal fato talvez possa ser explicado pelo fato do financiamento imobiliário ser um tipo de despesa fixa que a pessoa tenha que se comprometer a pagar e, portanto, tenha menos dinheiro em caixa para realizar investimentos.

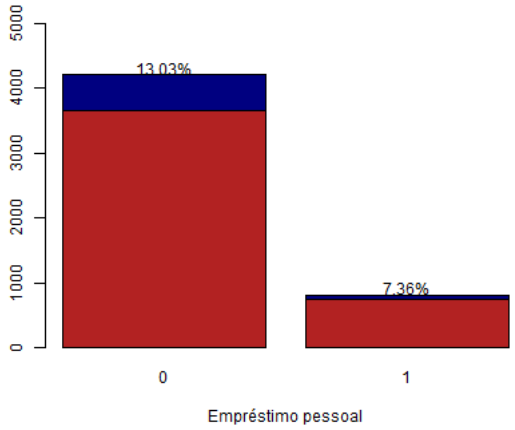

Figura 5.11: Gráfico de barras para a variável empréstimo pessoal.

Pela Figura 5.11, nota-se que a proporção de clientes que compraram o produto é quase duas vezes maior dentre os clientes que não possuem empréstimo pessoal em relação aqueles que possuem. Isso talvez possa ser explicado pelo mesmo motivo citado na análise

da Figura 5.10.

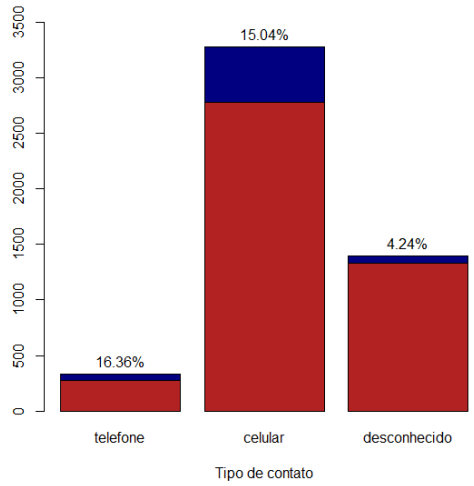

Figura 5.12: Gráfico de barras para a variável tipo de contato.

A Figura 5.12 mostra as proporções dos clientes que compraram o produto em relação ao tipo de contato realizado pela empresa. Nota-se que a proporção de clientes contatados por telefone fixo que compraram o produto é relativamente próxima em relação a proporção dos clientes contatados por celular que realizaram a compra, com 16,36% e 15,04%, respectivamente. Além disso, os clientes contatados por meios desconhecidos tem uma proporção de compra do produto aproximadamente 4 vezes inferior aqueles contatados por telefone fixo ou celular. De forma geral, essa variável não aparenta ter bom poder para discriminar os compradores dos não compradores.

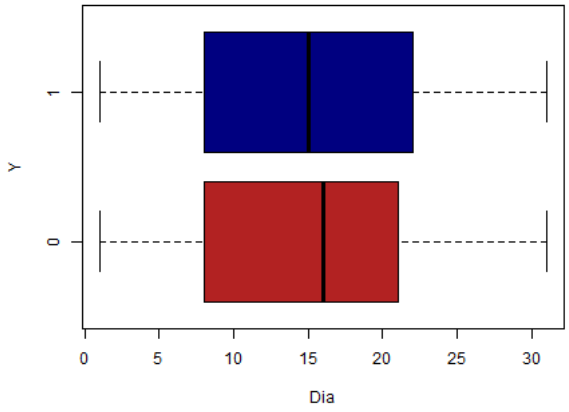

Figura 5.13: Boxplots da variável dia do último contato.

A partir da Figura 5.13, percebe-se que as distribuições da variável dia do último con-

tato em relação a ambos os grupos são muito semelhantes e aproximadamente simétricas. Além disso, metade dos clientes foi contatada pela última vez entre os dias 8 e 22 e a mediana é aproximadamente 15, independentemente do grupo. Essa variável também não aparenta ter bom poder para discriminar os compradores dos não compradores.

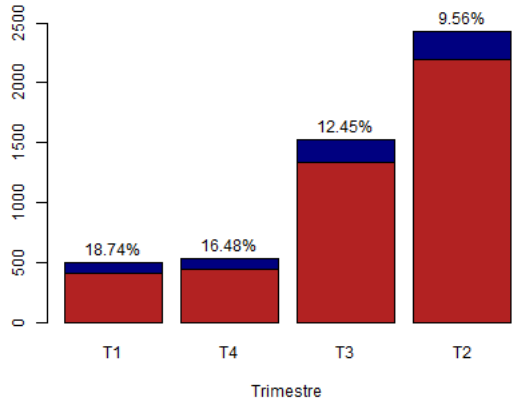

Figura 5.14: Gráfico de barras para a variável mês do último contato.

Pela Figura 5.14, é possível observar que a proporção de clientes que adquiriram o produto é maior nos grupos de clientes que foram contatados pela última vez no primeiro trimestre, seguida pelo quarto, terceiro e segundo trimestre, com 18,74%, 16,48%, 12,45% e 9,56%. De modo geral, a partir da análise gráfica, esta variável aparenta ter bom poder para discriminar os compradores dos não compradores.

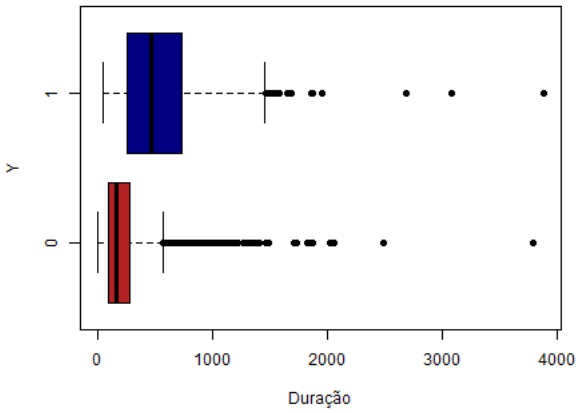

Figura 5.15: Boxplots para a variável duração do último contato.

Em relação à duração do último contato, essa variável também se destaca. Nota-se,

pela Figura 5.15, que os clientes que efetuaram a compra do produto, em geral, ficavam mais tempo na linha em relação aos clientes que não realizaram a compra. 50% dos compradores receberam ligações de duração entre 250 e 750 segundos, aproximadamente, enquanto que 50% dos clientes que não compraram o produto receberam ligações de duração entre 100 e 260 segundos, aproximadamente.

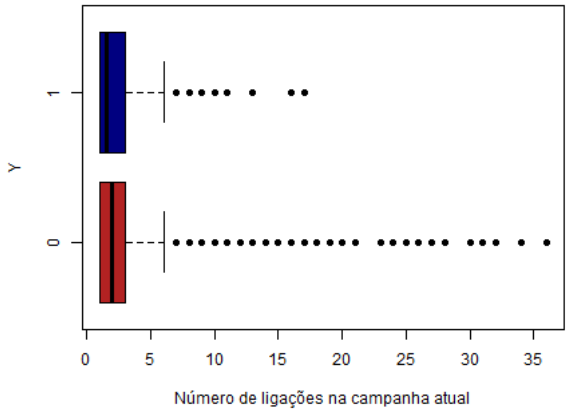

Figura 5.16: Boxplots para a variável número de ligações na campanha atual.

Pela Figura 5.16, é possível notar que 50% dos clientes que compraram o produto, efetuou a compra já na primeira ligação e 75% deles compraram após no máximo 3 ligações. Por outro lado, oito desses clientes receberam uma quantidade discrepante de ligações até efetuarem a compra, eles compraram o produto entre a sétima e a décima sétima ligação. Em relação aos não compradores, 75% deles foram contatados até 3 vezes. Ademais, 26 deles receberam quantidades discrepantes de ligações na campanha atual, esses foram contatados entre 7 e 36 vezes.

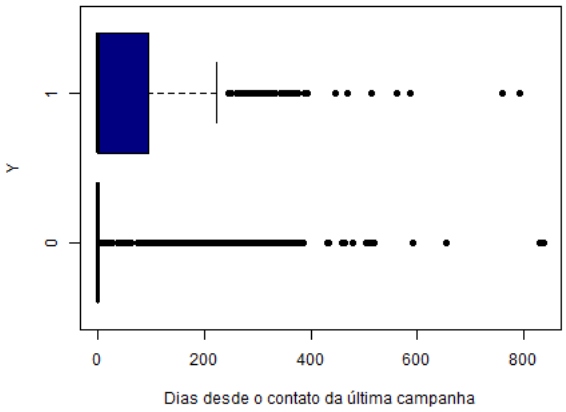

Figura 5.17: Boxplots para a variável dias desde o último contato na última campanha.

Pela Figura 5.17 nota-se que, dentre os compradores, mais dias se passaram desde o último contato na última campanha quando comparados aos não compradores. Além disso, observa-se que existem muitos outliers no grupo de não compradores, ou seja, já se passaram muitos dias desde o último contato com esses clientes e talvez seria uma boa estratégia tentar entrar em contato novamente.

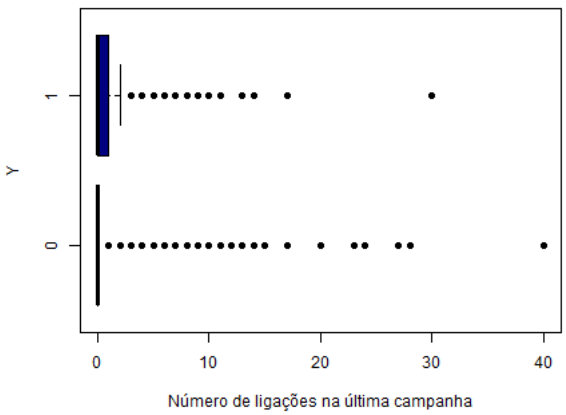

Figura 5.18: Boxplots para a variável número de ligações na última campanha.

Pela Figura 5.18, observa-se que, pelo menos 75% dos clientes que não compraram o produto, não foram contatados nenhuma vez na última campanha. Além disso, nota-se que 75% dos compradores receberam no máximo 2 ligações na última campanha.

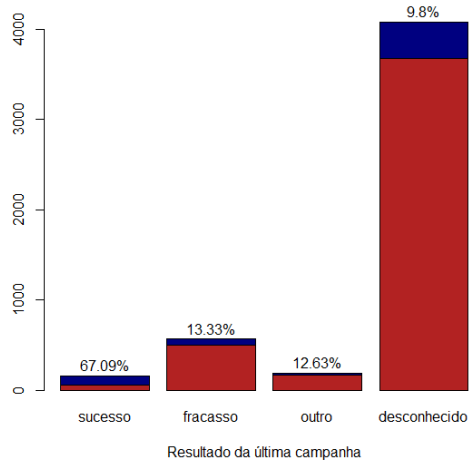

Figura 5.19: Gráfico de barras para a variável que indica se o cliente comprou o produto na última campanha.

A partir da Figura 5.19, observa-se que 67,09% dos clientes que compraram o produto na campanha atual também efetuaram a compra na última campanha e 13,33% deles não adquiriram o produto na campanha passada. Ou seja, 67,09% dos compradores demonstraram ser fiéis ao depósito a prazo disponibilizado pela instituição financeira e a taxa de conversão de vendas considerando a última campanha foi de 13,33%.

Em seguida, verificaremos a existência de correlação entre as variáveis não categóricas, com o objetivo de impedir futuros problemas de multicolinearidade.

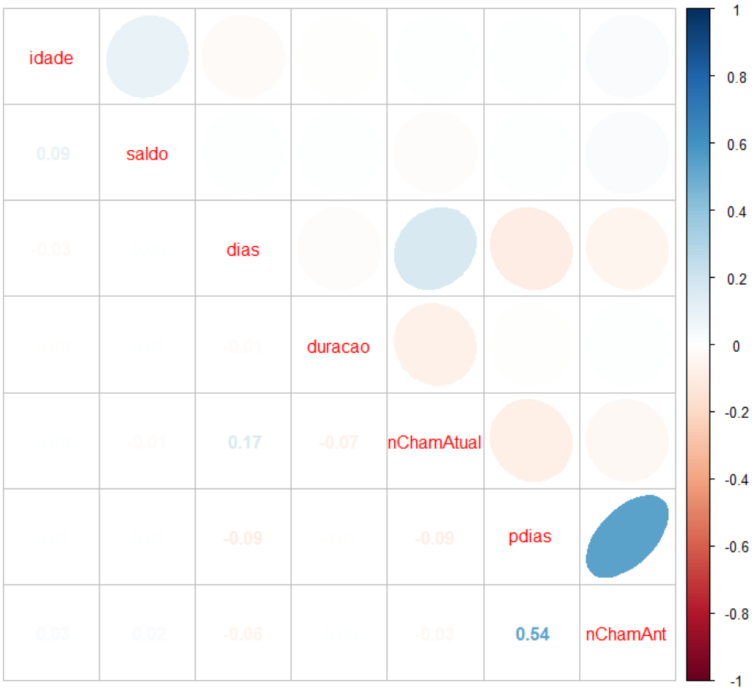

Figura 5.20: Representação da matriz de correlações.

Pela Figura 5.20 observa-se baixíssima correlação entre as variáveis na maioria dos casos. A maior identificada foi entre o número de dias desde a última chamada, e o número de ligações na última campanha, ambas vinculadas a eventos passados, com correlação de 0.54.

Em resumo, com a análise descritiva, foi possível observar algumas variáveis interessantes, como X_2 trabalho/ocupação do cliente (estudantes e aposentados possuem destaque), X_{11} trimestre do contato (o primeiro e último trimestre do ano são os que melhor converteram as investidas), X_{12} duração da chamada (dentre aqueles que compraram, a duração da chamada foi maior), e X_{16} resultado da última campanha (a maioria dos que compraram na última campanha, tomaram a mesma decisão na campanha atual).

Capítulo 6

Resultados

6.1 Estimador de máxima verossimlhança

Nesta Seção serão apresentados todos os procedimentos de modelagem e avaliação do ajuste obtido.

6.1.1 Modelo completo

Como visto pela análise descritiva, X_2, X_{11}, X_{12} e X_{16} tiveram um destaque visual para discriminar os clientes que compraram ou não o produto. Por esse motivo, serão incluídas no modelo completo todas as interações duplas entre essas variáveis, com exceção a X_{12} (duração do último contato na campanha atual) pelo fato de a mesma ser contínua e diminuir a margem de interpretação da interação.

Daqui em diante, para realizar todos os procedimentos da modelagem, foi utilizado o Software *R Studio*.

Primeiramente, utilizando a função *glm()* do *R Studio*, foi ajustado o modelo completo, ou seja, considerando todas as variáveis independentes e todas as interações duplas entre as variáveis X_2 (tipo de trabalho), X_{11} (mês do último contato) e X_{16} (variável que indica se o cliente adquiriu o produto na última campanha ou não), que se destacaram durante a análise descritiva. Portanto, utilizando a função de ligação logit, o modelo completo é dado por

$$ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_{17} X_2 X_{11} + \beta_{18} X_2 X_{16} + \beta_{19} X_{11} X_{16}. \quad (6.1)$$

6.1.2 Seleção de covariáveis

Posteriormente, verificamos, a partir do modelo completo, se existe a necessidade de todas as covariáveis estarem presentes. O método escolhido para essa etapa é o Stepwise AIC, disponível na biblioteca *MASS* do *R Studio* através da função *stepAIC()*.

Após a aplicação do método *Stepwise AIC*, algumas variáveis foram desconsideradas. Das 16 iniciais + 3 interações, restaram 10 + 2 interações, diminuindo a complexidade do modelo. As variáveis retiradas do modelo pelo método *Stepwise AIC* foram a idade, o grau de escolaridade, a indicadora que representa se o cliente é inadimplente ou não, o saldo anual médio, o número de dias corridos desde o último contato na campanha anterior, o número de chamadas realizadas par ao cliente na campanha anterior e a interação dupla entre as variáveis tipo de trabalho e o a indicadora que representa se o cliente comprou o produto na última campanha ou não.

Portanto, o modelo obtido é dado por

$$ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_{11} X_2 X_{11} + \beta_{12} X_{11} X_{16}. \tag{6.2}$$

6.1.3 Estimação dos parâmetros

Os coeficientes do modelo final foram estimados pelo *R Studio* e estão apresentados na Tabela 6.1.

Tabela 6.1: Coeficientes estimados do modelo.

Parâmetro	**Coef. estimado**
Intercepto	-3.13
tipoTrabalhoC1	0.42
tipoTrabalhoC2	0.37
tipoTrabalhoC3	1.61
estadoCivilcasado	-0.34
estadoCivilsolteiro	-0.08
financImobiliário	-0.90
emprPessoal	-0.50
tipoContatotelefone	-0.05
tipoContatodesconhecido	-1.79

trimestreT2	0.02
trimestreT3	1.71
trimestreT4	0.46
duração	0.00
numChamAtual	-0.06
resultUltCampoutro	-0.60
resultUltCampsucesso	2.85
resultUltCampdesconhecido	0.23
tipoTrabalhoC1:trimestreT2	0.15
tipoTrabalhoC2:trimestreT2	0.02
tipoTrabalhoC3:trimestreT2	-0.76
tipoTrabalhoC1:trimestreT3	-0.45
tipoTrabalhoC2:trimestreT3	-0.24
tipoTrabalhoC3:trimestreT3	-1.95
tipoTrabalhoC1:trimestreT4	0.11
tipoTrabalhoC2:trimestreT4	-0.97
tipoTrabalhoC3:trimestreT4	-0.04
trimestreT2:resultUltCampoutro	0.29
trimestreT3:resultUltCampoutro	1.35
trimestreT4:resultUltCampoutro	1.08
trimestreT2:resultUltCampsucesso	0.07
trimestreT3:resultUltCampsucesso	-1.46
trimestreT4:resultUltCampsucesso	-0.61
trimestreT2:resultUltCampdesconhecido	0.28
trimestreT3:resultUltCampdesconhecido	-1.95
trimestreT4:resultUltCampdesconhecido	-0.47

6.1.4 Métricas de avaliação do ajuste

Para o problema em questão, um bom modelo deveria ser capaz de separar os clientes que compraram o produto daqueles que não compraram, para que em campanhas futuras os esforços sejam direcionados naqueles com perfil adequado ao produto. Nessa linha de raciocínio, após inserir as informações de determinado cliente no modelo, é retornado um

valor que varia entre 0 e 1. Será que existe alguma concentração de clientes que compraram o produto em algum intervalo? Dessa forma, algumas técnicas que medem o quão bem os clientes são separados a partir do valor de saída do modelo logito são a curva ROC (Receiver Operating Characteristic) e a estatística de Kolmogorov-Smirnov (KS).

Curva ROC

A partir da matriz de confusão (verdadeiros e falsos positivos/negativos) é construída a curva variando o ponto de corte do modelo (a partir de qual valor $\in [0, 1]$ classificar como 1) e suas respectivas sensibilidades (probabilidade do cliente ser classificado como comprador dado que ele realmente comprou) e especificidade (probabilidade do cliente ser classificado como não comprador dado que ele realmente não comprou) do modelo.

Uma medida derivada da curva ROC, é a AUC (area under the curve), que nada mais é do que a integral da curva. Quanto maior o seu valor, melhor a qualidade do ajuste. O valor da AUC do modelo final é apresentado na Tabela , lembrando que $0 \leq AUC \leq 1$.

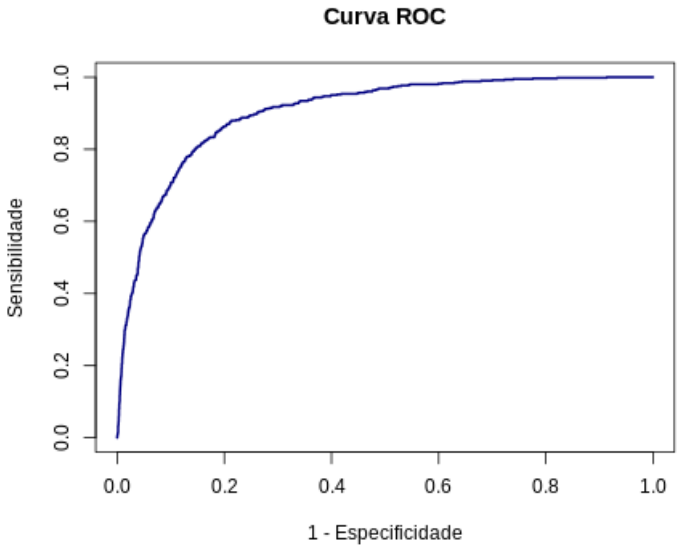

Figura 6.1: Curva ROC do modelo final.

Tabela 6.2: Valor da AUC do modelo final.

AUC
0,9042739

Como notado acima, obtém-se um valor de aproximadamente 0,9 para a AUC do modelo final, que é consideravelmente alto, dando indícios de que o modelo se ajustou

consideravelmente bem aos dados.

Estatística Kolmogorov-Smirnov

A estatística de Kolmogorov-Smirnov tem origem no teste não-paramétrico de Kolmogorov-Smirnov, que verifica se duas amostras possuem distribuições iguais ou não. No nosso caso, queremos que o modelo ajustado seja capaz de discriminar os compradores dos não compradores considerando o conjunto de covariáveis, ou seja, desejamos que a distribuição dos clientes que compraram o produto seja diferente da distribuição dos clientes que não realizaram a compra. Quanto maior o valor dessa estatística, melhor a capacidade do modelo de satisfazer o objetivo do trabalho, isto é, discriminar os compradores dos não compradores.

Pela Figura 6.2, percebe-se que o modelo realiza uma boa separação entre os clientes, sendo a curva azul a distribuição acumulada dos clientes que compraram o produto e a curva amarelo a distribuição acumulada dos não compradores.

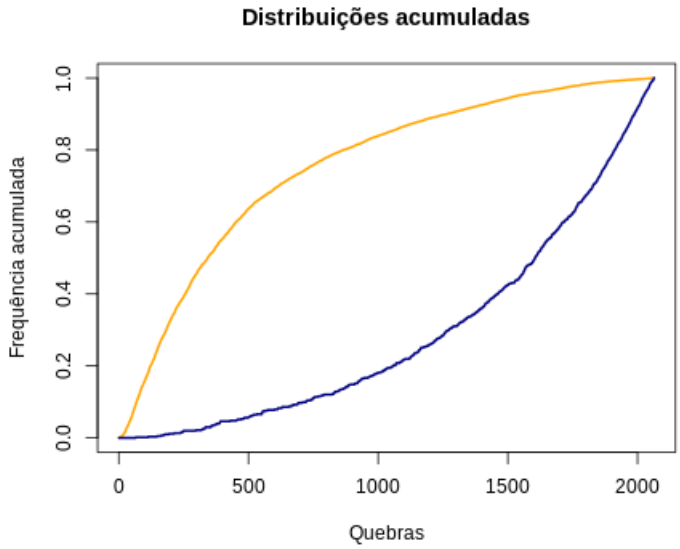

Figura 6.2: Distribuição acumulada dos clientes que compraram o produto (azul) e dos que não compraram (amarelo).

Além disso, o valor obtido na estatística (0.6657) é razoavelmente alto, sendo mais um indício de que o modelo tem boa capacidade de discriminar os compradores dos não compradores.

6.1.5 Interpretação dos coeficientes estimados

Uma medida amplamente utilizada em modelos logito é o *odds ratio*, que nada mais é que a exponencial dos parâmetros estimados, denotado por $\Psi(\beta_i)$. Interpreta-se a medida como a propensão que o cliente possui em comprar o produto, dado que ele pertence à categoria do parâmetro.

Tabela 6.3: Odds ratio dos coeficientes estimados do modelo.

Parâmetro	**Odds ratio**
Intercepto	0.04
tipoTrabalhoC1	1.53
tipoTrabalhoC2	1.44
tipoTrabalhoC3	5.00
estadoCivilcasado	0.71
estadoCivilsolteiro	0.92
financImobiliário	0.41
emprPessoal	0.60
tipoContatotelefone	0.95
tipoContatodesconhecido	0.17
trimestreT2	1.02
trimestreT3	5.50
trimestreT4	1.58
duração	1.00
numChamAtual	0.94
resultUltCampoutro	0.55
resultUltCampsucesso	17.37
resultUltCampdesconhecido	1.26
tipoTrabalhoC1:trimestreT2	1.16
tipoTrabalhoC2:trimestreT2	1.02
tipoTrabalhoC3:trimestreT2	0.47
tipoTrabalhoC1:trimestreT3	0.64
tipoTrabalhoC2:trimestreT3	0.79

tipoTrabalhoC3:trimestreT3	0.14
tipoTrabalhoC1:trimestreT4	1.11
tipoTrabalhoC2:trimestreT4	0.38
tipoTrabalhoC3:trimestreT4	0.96
trimestreT2:resultUltCampoutro	1.33
trimestreT3:resultUltCampoutro	3.85
trimestreT4:resultUltCampoutro	2.96
trimestreT2:resultUltCampsucesso	1.07
trimestreT3:resultUltCampsucesso	0.23
trimestreT4:resultUltCampsucesso	0.55
trimestreT2:resultUltCampdesconhecido	1.32
trimestreT3:resultUltCampdesconhecido	0.14
trimestreT4:resultUltCampdesconhecido	0.62

Observando a Tabela 6.3, percebe-se algumas variáveis com grande impacto. Por exemplo, tipoTrabalhoC3 que foi recategorizada no início do estudo, possui $\hat{\mathbf{\Psi}} \approx 5$, isto é, clientes que fazem parte dessa categoria (estudantes e aposentados) possuem 5 vezes mais chances de comprar o produto quando comparados aqueles que não fazem parte da mesma. Outra variável de bastante impacto é trimestreT3, com $\hat{\mathbf{\Psi}} \approx 5,50$, ou seja, os clientes que foram contatados pela última vez no terceiro trimestre possuem 5,5 vezes mais chances de comprar o produto em relação aos clientes que foram contatados pela última vez no primeiro trimestre do ano (categoria de referência trimestreT1). Mas, sem dúvidas, o que mais chama a atenção é resultUltCampsucesso, com $\hat{\mathbf{\Psi}} \approx 17.37$, isto é, clientes que compraram o produto na última campanha, possuem 17 vezes mais chances de comprar o produto novamente, comparado aqueles que não o compraram, o que confirma a hipótese levantada na análise gráfica realizada na Seção 5.1.

6.2 Estimador bayesiano - Modelo de mistura com K = 1

Inicialmente, iremos ajustar o modelo bayesiano de regressão logística para $K = 1$ (apenas uma componente ou não mistura). Utilizaremos as dez covariáveis numéricas consideradas no modelo descrito na Seção 6.1.2 mais as interações duplas X_2:X_{11} (tipo de trabalho e trimestre do último contato) e X_{11}:X_{16} (trimestre do último contato).

Para realizar o ajuste do modelo, utilizaremos a função *brm* do pacote *brms* do R. Para isso, usaremos o seguinte código:

```
modelok1 = brm(formula = y ~ job + marital + housing + loan + contact +
month + duration + campaign + poutcome + job:month + month:poutcome,
family = bernoulli(),
data = TB_BANK_SF,
prior = set_prior("normal(0,50)",
class = c("b","Intercept")),
chains = 3,
cores = 3,
iter = 2000,
warmup = 1000,
seed = 123)
```

No argumento *formula*, temos a variável resposta Y sendo predita pelas dez covariáveis numéricas mais as interações duplas entre as variáveis X_2, X_{11}, e X_{16}. No argumento *family*, declaramos a distribuição da variável resposta, no caso, a distribuição de Bernoulli, que é o que caracteriza o modelo logístico. No argumento *data*, temos o banco de dados utilizado, em formato de *dataframe do R*. *prior* é o argumento que determina a distribuição a priori dos parâmetros. Neste caso, optamos por utilizar distribuições a priori vagas, isto é, distribuições não informativas, haja vista que não temos conhecimento prévio ou conhecimento específico sobre cada um dos parâmetros. Dessa forma, foi definida uma distribuição Normal com parâmetros $\mu = 0$ e $\sigma = 50$ para todos os parâmetros do modelo. *chains* define o número de cadeias de Markov utilizadas no processo, *iter* define o número de iterações em cada cadeia de Markov e *warmup* define o número de iterações na fase de warmup (burn-in). Todos esses argumentos podem ser definidos de acordo com a

necessidade do usuário para obter um melhor ajuste do modelo. Por exemplo, caso não seja possível atingir a convergência para um determinado modelo, aumentar o número de cadeias de Markov ou o número de iterações em cada cadeia pode ser uma boa tentativa para tentar obter a convergência.

6.2.1 Diagnóstico de convergência do ajuste

A função *brm* do pacote *brms* utilizado disponibiliza algumas métricas e gráficos para avaliação da convergência do modelo ajustado. A seguir, essas métricas serão avaliadas para verificar se a convergência foi obtida e, portanto, se o modelo é confiável nesse quesito.

Na Tabela 6.4, estão dispostos os valores das métricas de alguns parâmetros para fins de demonstrar que a convergência foi atingida.

Tabela 6.4: Métricas para avaliar a convergência do ajuste.

Parâmetro	**R-hat**	**Bulk ESS**	**Tail ESS**
Intercepto	1.00	1388	2444
tipoDeTrabalhoC1	1.00	1325	2168
tipoDeTrabalhoC2	1.00	1278	2212
tipoDeTrabalhoC3	1.00	1402	2373
estadoCivilcasado	1.00	4448	3005
estadoCivilsolteiro	1.00	4179	3144
financImobiliário	1.00	7171	3315
emprPessoal	1.00	6757	3049
tipoContatotelefone	1.00	7006	3284

Pela Tabela 6.4, nota-se que foram obtidos valores de *R-hat* iguais a 1 para os coeficientes estimados. Além disso, foram obtidos valores grandes para *Bulk ESS* e *Tail ESS*. Segundo o manual do Stan (2024b), para podermos concluir, a partir destas métricas, que o modelo convergiu, devemos ter valores de *R-hat* menores que 1.01 e valores para *Bulk ESS* e *Tail ESS* maiores que 100 vezes o número de cadeias de Markov utilizadas no ajuste. Neste caso, 100 * 3 cadeias = 300. Dessa forma, concluímos que os valores obtidos para estas métricas indica que o modelo convergiu.

Análise Gráfica

A partir da Figura 6.3 até a Figura 6.10, nos gráficos à direita, observa-se os *traceplots* obtidos no ajuste do modelo. O comportamento dos gráficos evidencia que as cadeias de Markov se misturaram bem e, assim, nos dá indícios de que o ajuste convergiu. No gráfico à esquerda, temos a distribuição a posteriori dos coeficientes estimados, cujas médias são as estimativas pontuais desses coeficientes, que estão apresentadas na Tabela 6.5.

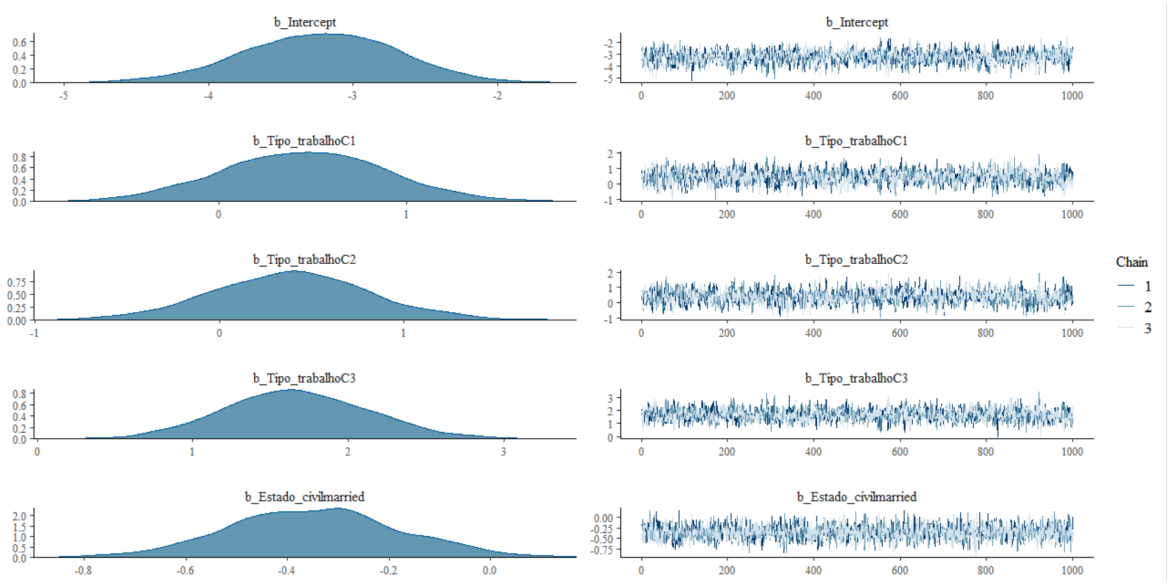

Figura 6.3: Densidades das distribuições a posteriori dos parâmetros e traceplots das Cadeias de Markov.

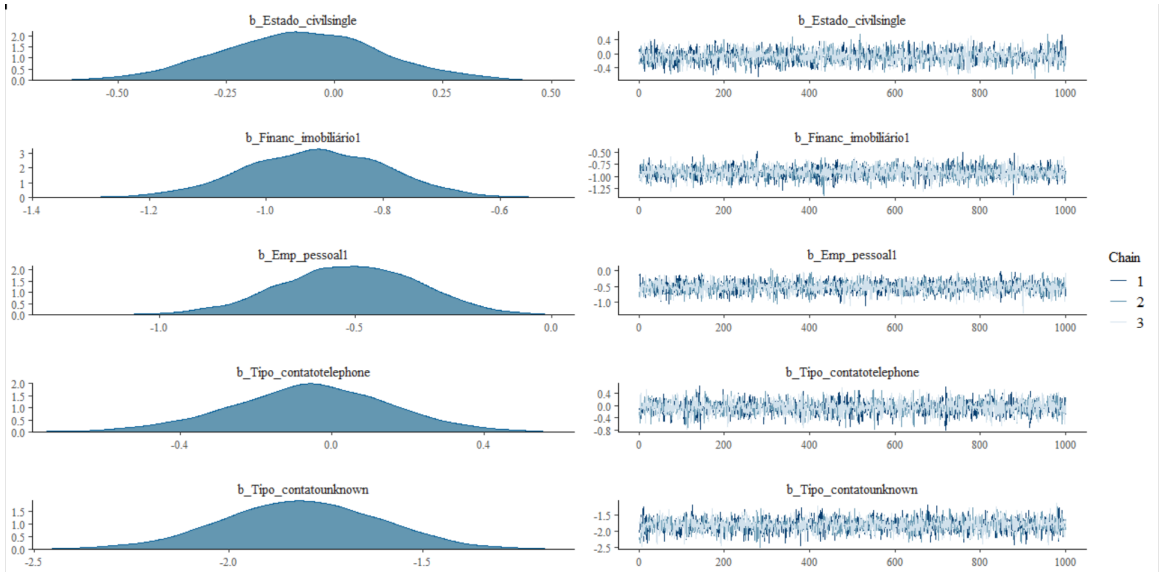

Figura 6.4: Densidades das distribuições a posteriori dos parâmetros e traceplots das Cadeias de Markov.

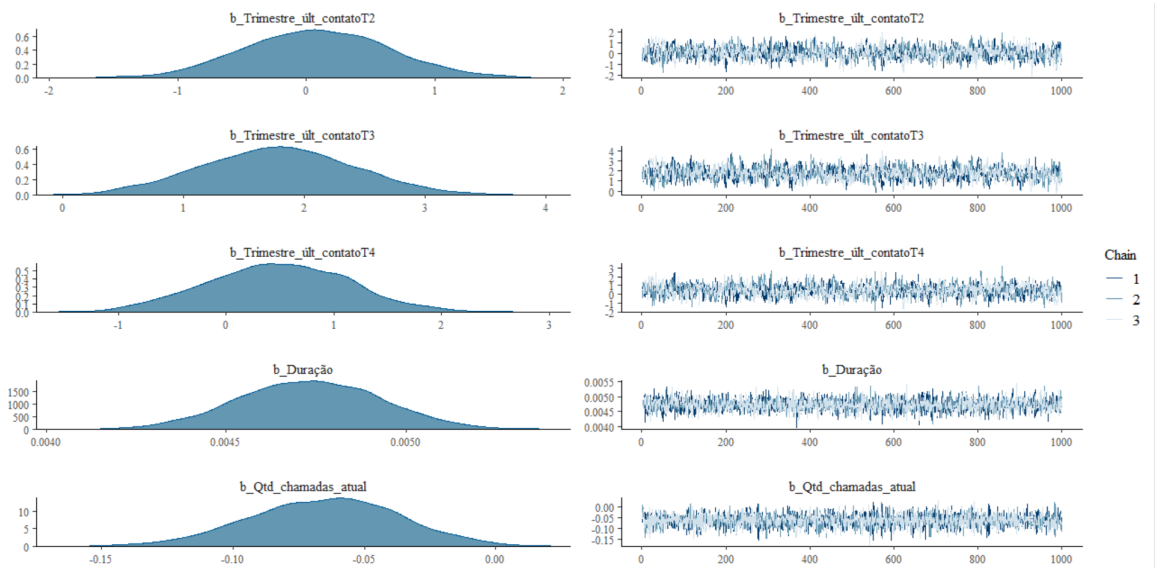

Figura 6.5: Densidades das dsitribuições a posteriori dos parâmetros e traceplots das Cadeias de Markov.

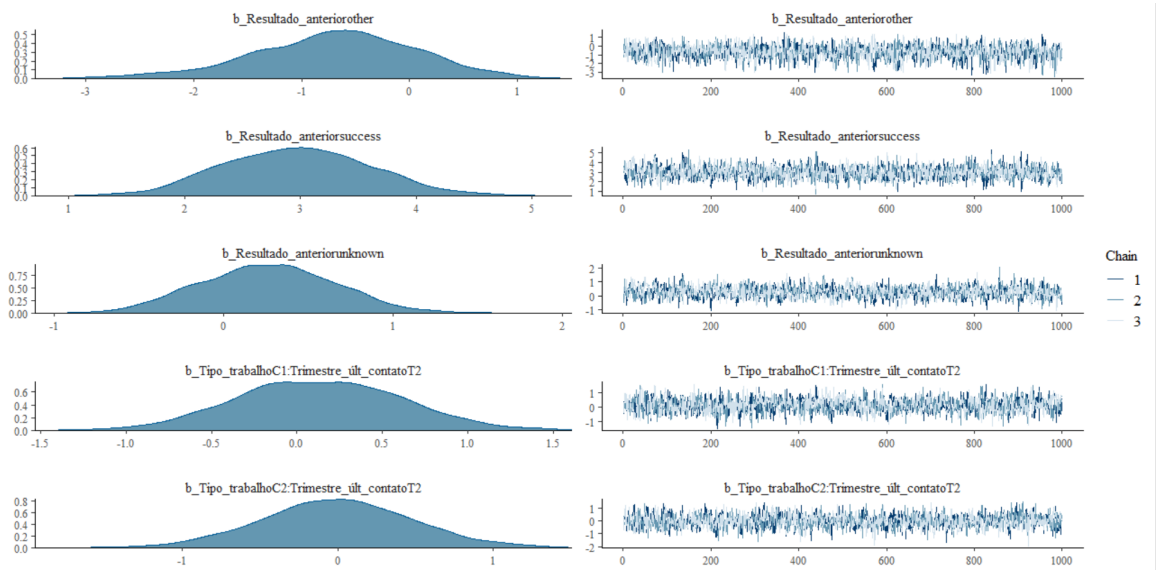

Figura 6.6: Densidades das dsitribuições a posteriori dos parâmetros e traceplots das Cadeias de Markov.

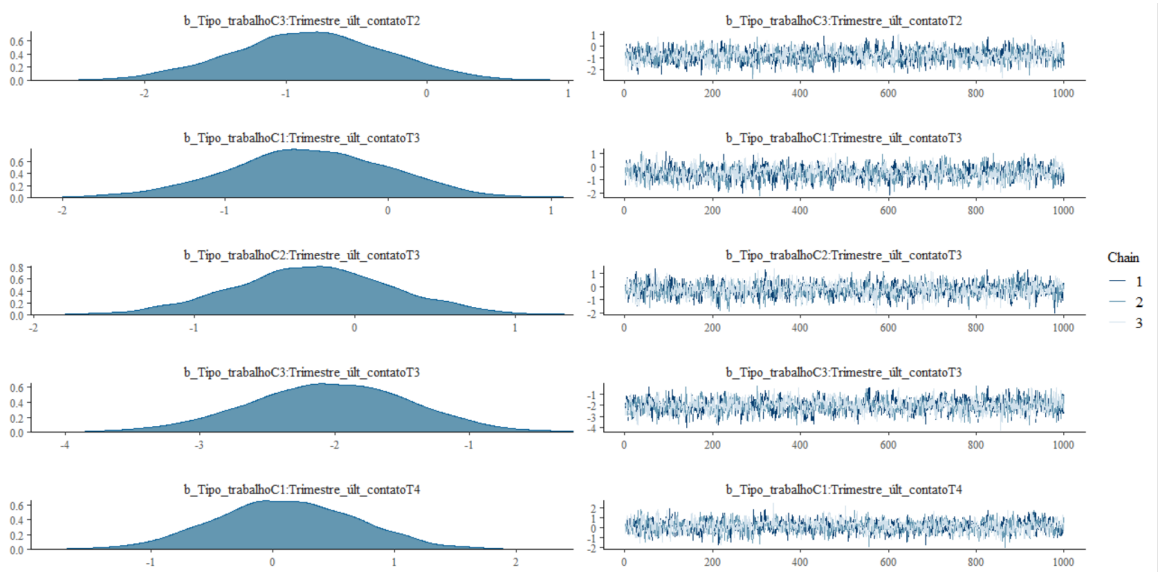

Figura 6.7: Densidades das dsitribuições a posteriori dos parâmetros e traceplots das Cadeias de Markov.

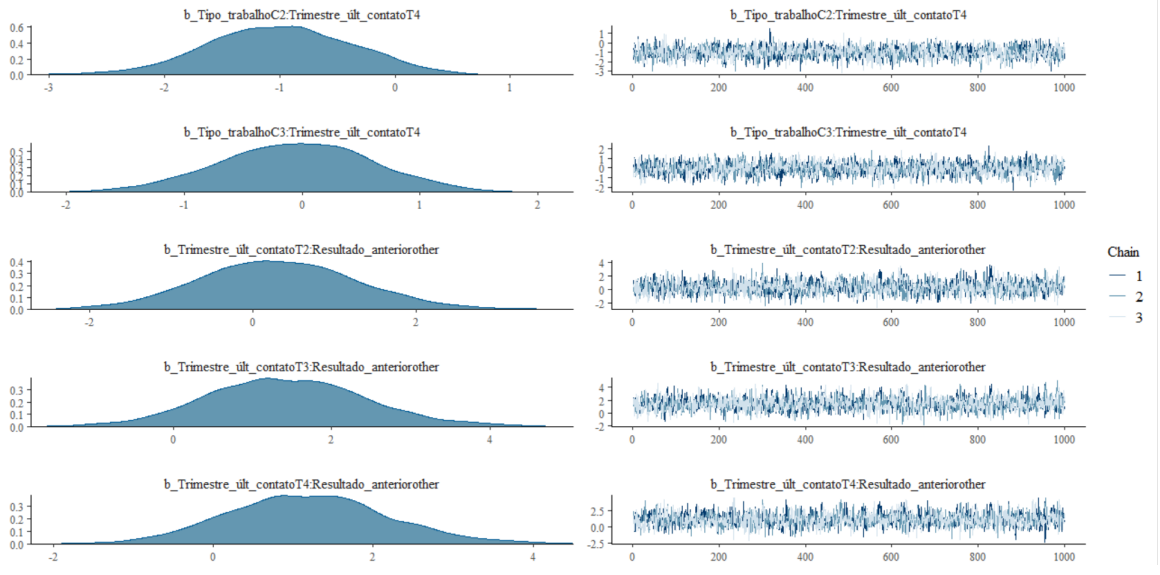

Figura 6.8: Densidades das dsitribuições a posteriori dos parâmetros e traceplots das Cadeias de Markov.

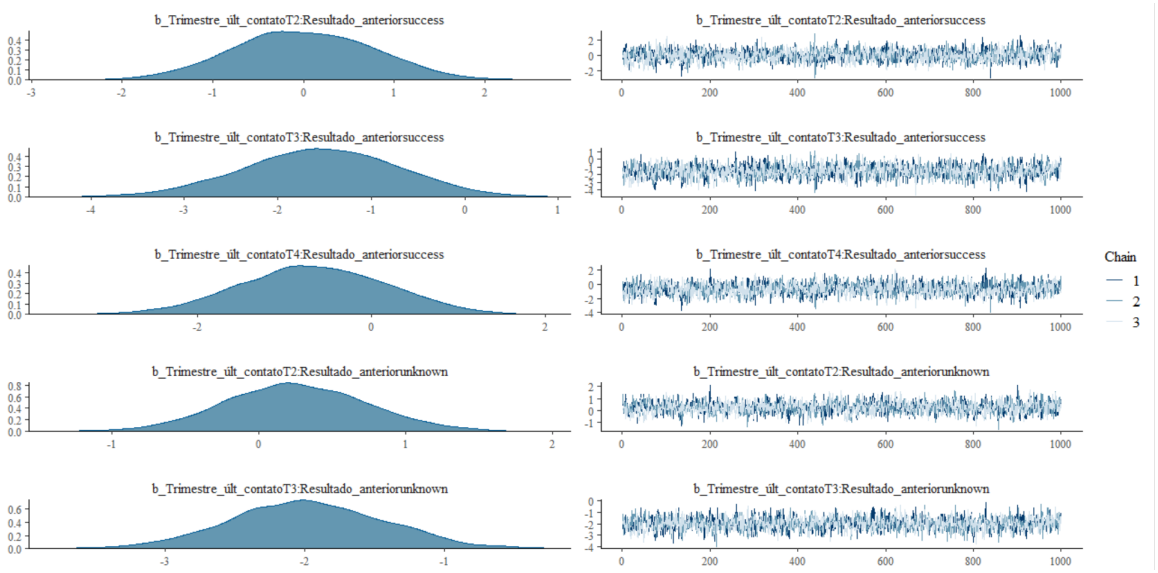

Figura 6.9: Densidades das dsitribuições a posteriori dos parâmetros e traceplots das Cadeias de Markov.

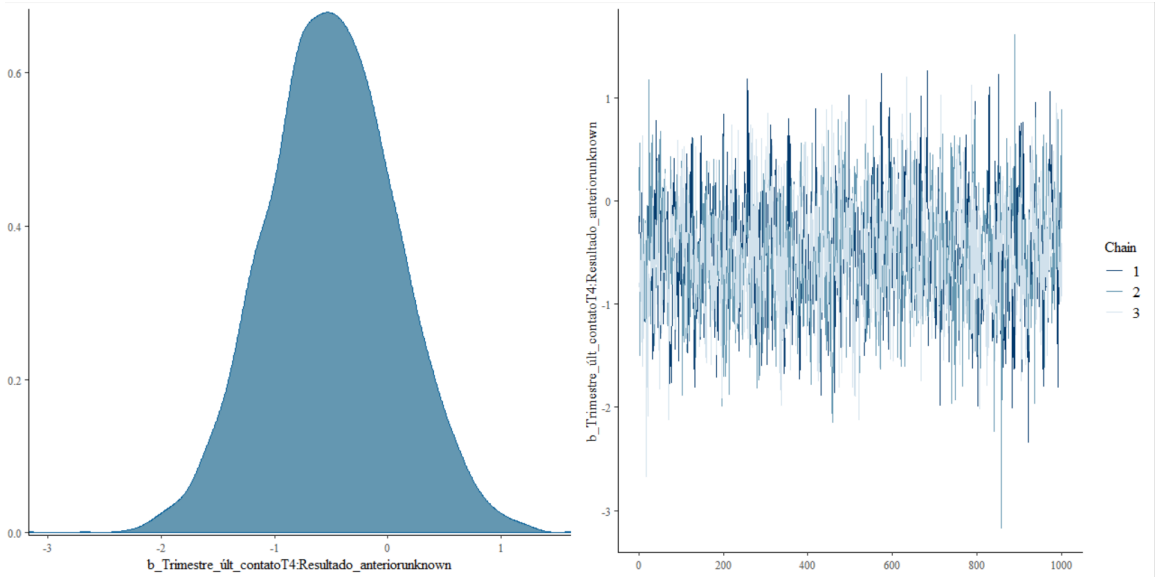

Figura 6.10: Densidades das dsitribuições a posteriori dos parâmetros e traceplots das Cadeias de Markov.

6.2.2 Coeficientes estimados do modelo

Na Tabela 6.5, temos os coeficientes estimados do ajuste e seus respectivos erros médios, além dos intervalos de credibilidade de 95%.

Tabela 6.5: Coeficientes estimados do modelo, seus respectivos erros e intervalos de credibilidade de 95%.

Parâmetro	**Coef. estimado**	**Erro**	**LI IC 95%**	**LS IC 95%**
Intercepto	-3.23	0.53	-4.31	-2.23
tipoTrabalhoC1	0.45	0.43	-0.39	1.29
tipoTrabalhoC2	0.39	0.43	-0.47	1.26
tipoTrabalhoC3	1.67	0.46	0.78	2.60
estadoCivilmarried	-0.35	0.17	-0.67	-0.01
estadoCivilsolteiro	-0.08	0.18	-0.44	0.28
financImobiliário	-0.91	0.12	-1.16	-0.68
emprPessoal	-0.51	0.18	-0.87	-0.18
tipoContatotelefone	-0.06	0.21	-0.50	0.35
tipoContatodesconhecido	-1.81	0.21	-2.22	-1.42
trimestreT2	0.08	0.57	-0.99	1.19
trimestreT3	1.78	0.63	0.55	3.02
trimestreT4	0.48	0.67	-0.81	1.81
duração	0.0047	0.0002	0.0043	0.0051
numChamAtual	-0.07	0.03	-0.12	-0.01
resultUltCampoutro	-0.70	0.77	-2.39	0.75
resultUltCampsucesso	2.97	0.65	1.77	4.30
resultUltCampdesconhecido	0.26	0.41	-0.50	1.05
tipoTrabalhoC1:trimestreT2	0.13	0.49	-0.82	1.09
tipoTrabalhoC2:trimestreT2	-0.01	0.49	-0.97	0.99
tipoTrabalhoC3:trimestreT2	-0.83	0.55	-1.90	0.24
tipoTrabalhoC1:trimestreT3	-0.49	0.50	-1.50	0.46
tipoTrabalhoC2:trimestreT3	-0.26	0.51	-1.24	0.72
tipoTrabalhoC3:trimestreT3	-2.03	0.60	-3.23	-0.89
tipoTrabalhoC1:trimestreT4	0.11	0.59	-1.03	1.24
tipoTrabalhoC2:trimestreT4	-0.99	0.63	-2.23	0.20
tipoTrabalhoC3:trimestreT4	-0.05	0.64	-1.32	1.17
trimestreT2:resultUltCampoutro	0.31	0.96	-1.54	2.26
trimestreT3:resultUltCampoutro	1.44	1.02	-0.49	3.60

trimestreT4:resultUltCampoutro	1.19	1.02	-0.75	3.30
trimestreT2:resultUltCampsucesso	0.00	0.77	-1.49	1.49
trimestreT3:resultUltCampsucesso	-1.54	0.83	-3.18	0.05
trimestreT4:resultUltCampsucesso	-0.67	0.84	-2.30	0.93
trimestreT2:resultUltCampdesconhecido	0.26	0.48	-0.68	1.20
trimestreT3:resultUltCampdesconhecido	-1.99	0.55	-3.05	-0.91
trimestreT4:resultUltCampdesconhecido	-0.50	0.57	-1.61	0.63

Pela Tabela 6.5 observa-se que, de modo geral, obtivemos coeficientes estimados semelhantes aqueles apresentados na Tabela 6.1, com os mesmos parâmetros se destacando. Dessa forma, a interpretação dos coeficientes estimados é feita de forma análoga ao que foi apresentado na Seção 6.1.5. Além disso, nota-se que existem alguns parâmetros cujo intervalo de credibilidade de 95% contém o zero, como os parâmetros referentes às categorias C1 e C2 da variável tipo de trabalho, os parâmetros referentes à variável estado civil, o tipo de contato telefone, os parâmetros referentes aos trimestres 2 e 4 relacionados ao último contato com o cliente, entre outros. Isto é, a probabilidade a posteriori desses parâmetros serem iguais a 0 é de 95%, indicando que os mesmos não têm grande impacto na probabilidade de compra do produto depósito a prazo quando comparados à categoria de referência de cada variável. Dessa forma, um processo de seleção de variáveis poderia ser aplicado ao modelo para testar se essas variáveis poderiam ser retiradas do mesmo, porém, nós optamos por não fazer isso, pois não faz parte dos objetivos deste trabalho.

Efeitos condicionais das variáveis preditoras

A seguir, exploraremos os efeitos condicionais das variáveis preditoras na variável resposta por meio das Figuras 6.11 até 6.21.

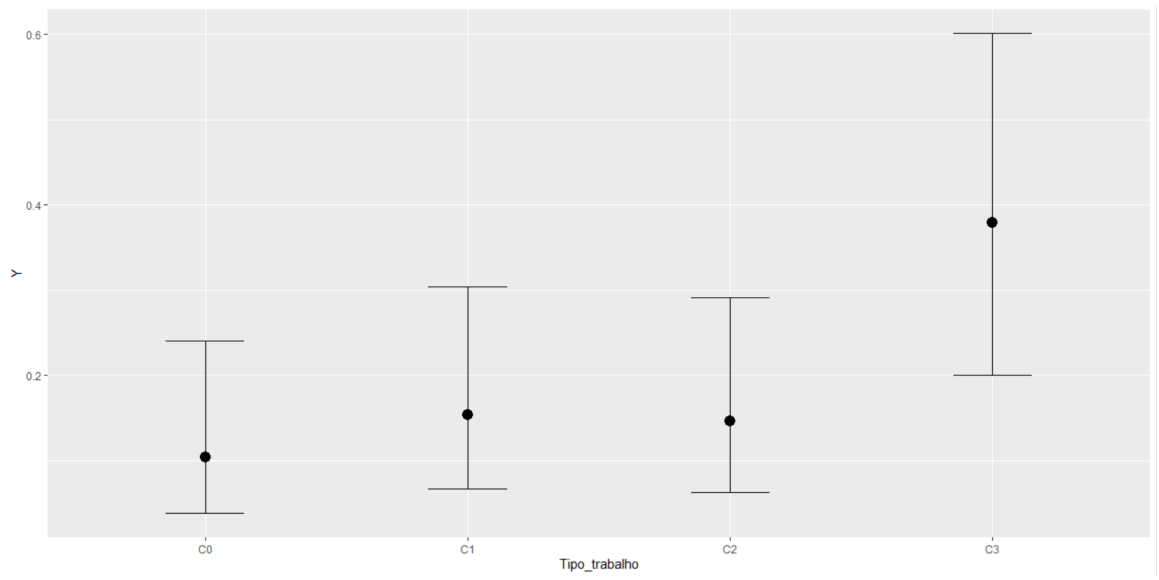

Figura 6.11: Efeito condicional da variável tipo de trabalho.

Pela Figura 6.11, nota-se que a categoria de trabalho C3 (estudantes e aposentados) se destaca. Podemos dizer que estudantes e aposentados têm maior probabilidade de realizar a compra do produto depósito a prazo quando comparados às pessoas que possuem as outras profissões agrupadas em C0, C1 e C2, que possuem estimativas pontuais e intervalos de credibilidade muito semelhantes. Pode-se dizer que este resultado era esperado, haja vista que, na Figura 5.5, verificou-se que os estudantes e aposentados agrupados em C3 têm a maior proporção de compra do produto comparado às outras profissões. Além disso, pela Tabela 6.5, nota-se um maior coeficiente estimado para o parâmetro atrelado à categoria C3 (1.59) quando comparado aos coeficientes estimados das outras categorias C1 e C2 (0.45 e 0.38), indicando a categoria C3 (estudantes e aposentados) impacta mais na probabilidade de compra do produto depósito a prazo que as outras categorias.

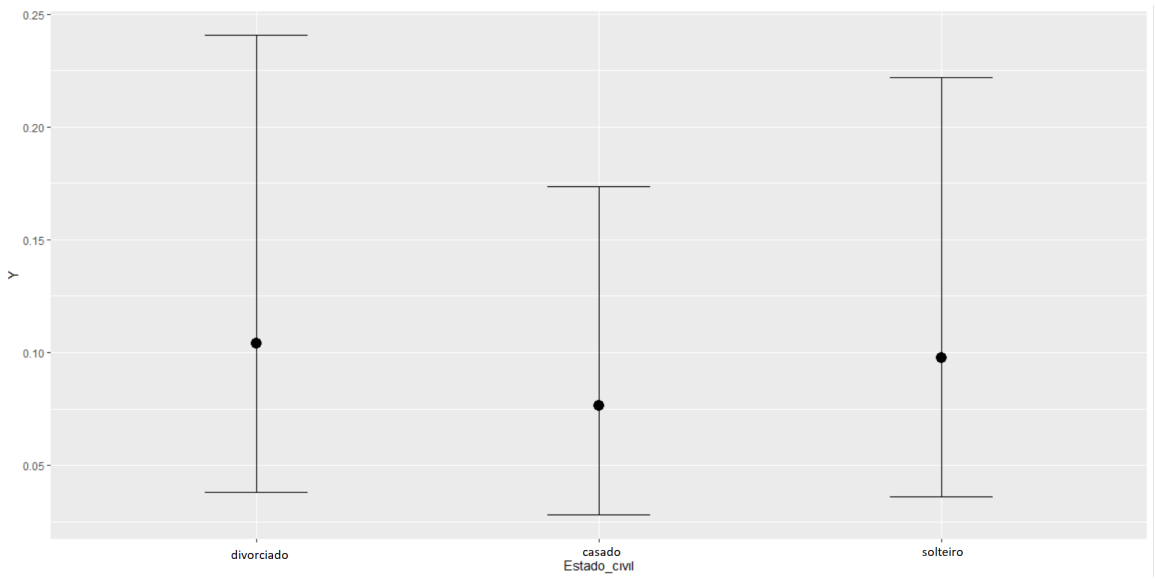

Figura 6.12: Efeito condicional da variável estado civil.

Em relação à variável estado civil, nota-se, pela Figura 6.12, que as três categorias (divorciados, casados e solteiros) não possuem diferença significativa, haja vista que as estimativas pontuais estão muito próximas e os intervalos de credibilidade têm grande intersecção entre si. Isso, aliado com os baixos coeficientes estimados para os parâmetros atrelados às categorias desta variável exibidos na Tabela 6.5 e a presença do zero nos intervalos de credibilidade, indica que esta variável não tem grande significância para alterar a probabilidade de compra do produto depósito a prazo.

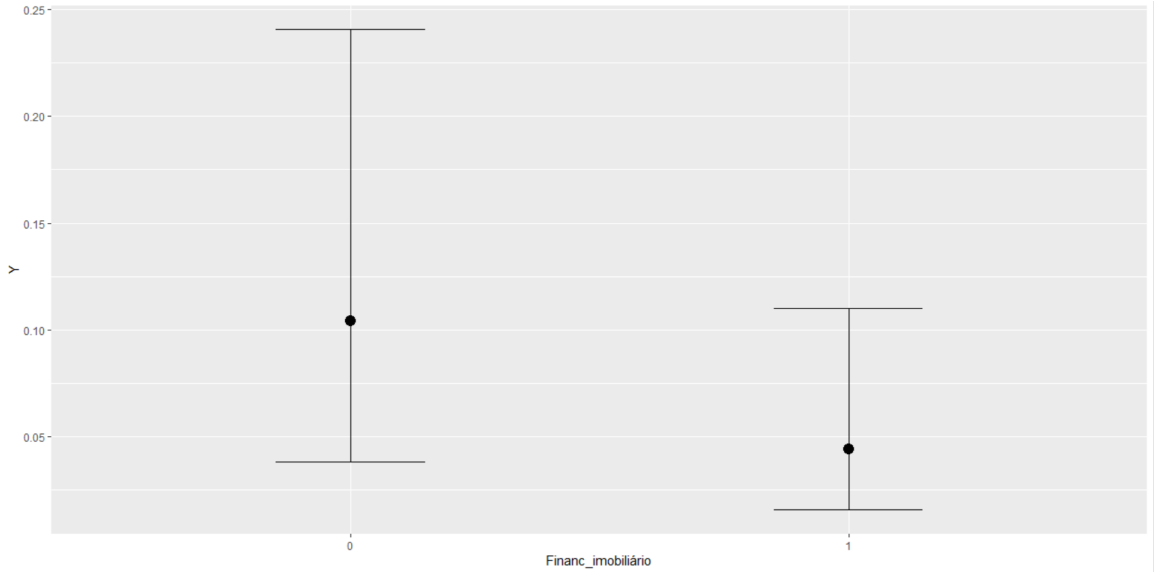

Figura 6.13: Efeito condicional da variável que indica se o cliente possui financiamento imobiliário.

Em relação à variável que indica se o cliente possui financiamento imobiliário, pela Figura 6.13 observamos que, os clientes que possuem esse tipo de financiamento, têm menos probabilidade de realizar a compra do produto depósito a prazo, comparados aqueles clientes que não possuem, ainda que o intervalo de 95% de credibilidade deste último seja muito amplo.

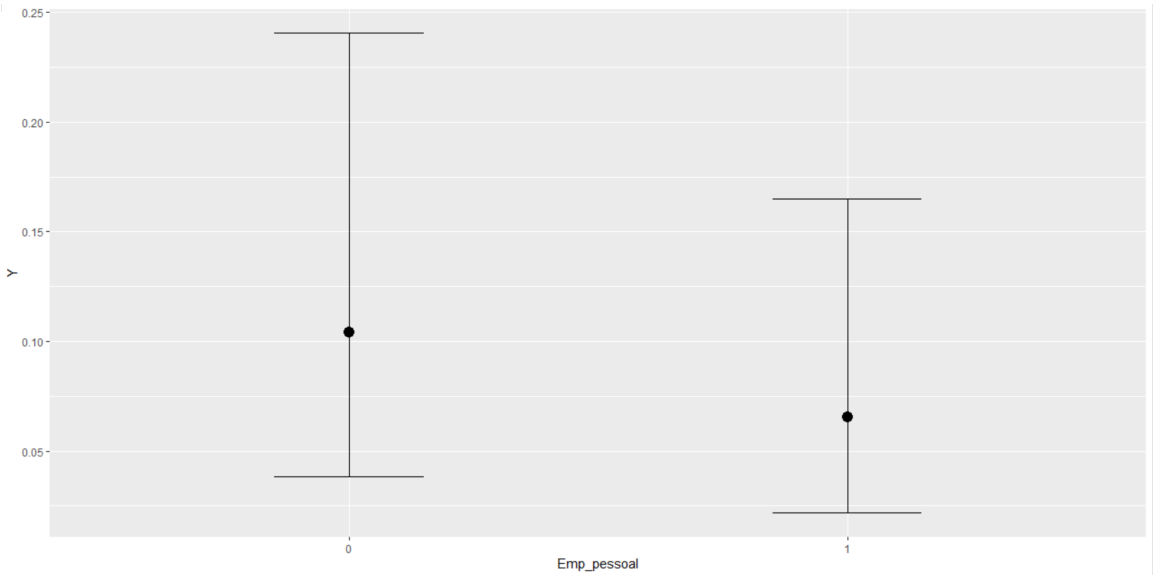

Figura 6.14: Efeito condicional da variável que indica se o cliente possui empréstimo pessoal.

Para a variável que indica se o cliente possui empréstimo pessoal, pela Figura 6.14, observamos um comportamento muito semelhante ao comportamento da variável analisada na Figura 6.13. Aqui, os clientes que possuem empréstimo pessoal têm menos probabilidade de adquirir o produto de depósito a prazo quando comparados aqueles que não possuem empréstimo pessoal. Uma explicação para isso pode ser que os clientes que já possuem algum tipo de despeza fixa proveniente de um financiamento imobiliário ou empréstimo pessoal sejam menos interessados em realizar um investimento do tipo depósito a prazo.

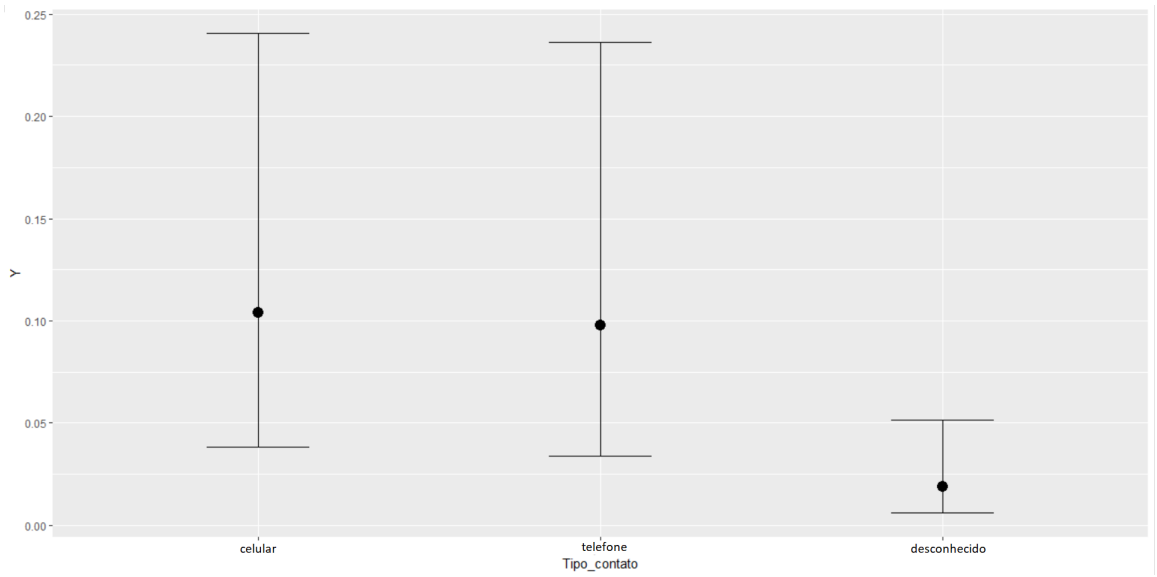

Figura 6.15: Efeito condicional da variável tipo de contato.

Para a variável tipo de contato, nota-se pela Figura 6.15 grande semelhança entre as categorias celular e telefone, que possuem estimativas pontuais e intervalos de credibilidade praticamente idênticos, indicando que não existe grande diferença no impacto na probabilidade de compra do produto depósito a prazo por clientes contatados por esses dois dispositivos, já que trata-se de chamadas de voz nos dois casos. O tipo de contato desconhecido presente no banco de dados se destaca com uma estimativa pontual menor e um intervalo de credibilidade muito menos amplo em relação às categorias celular e telefone, evidenciando a diferença entre o efeito do contato por chamada de voz (celular e telefone) e outro tipo de contato desconhecido, que não é explícito na descrição do banco de dados.

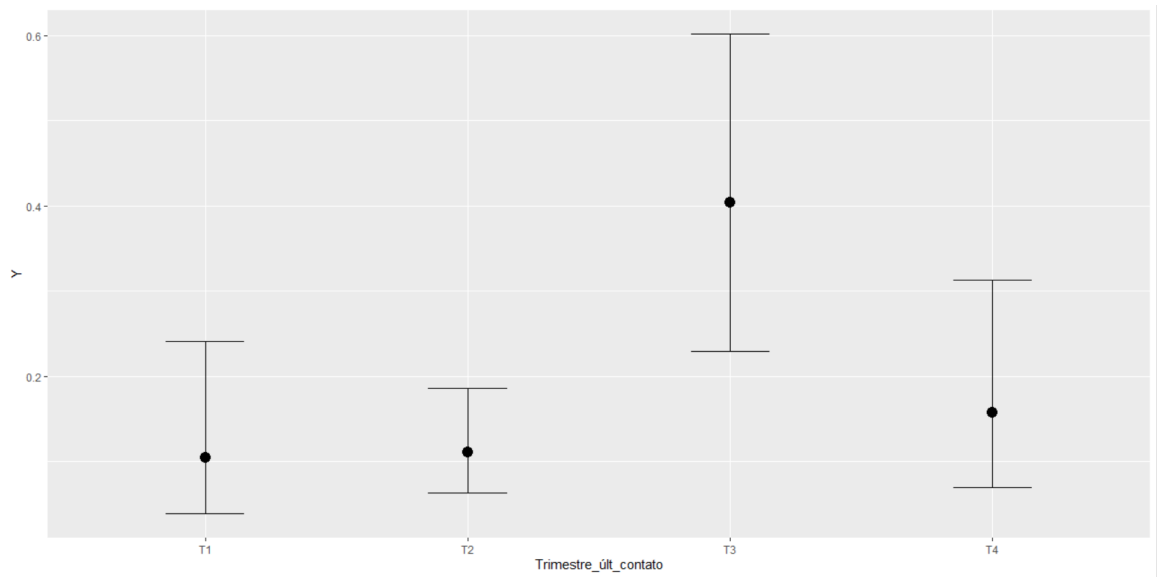

Figura 6.16: Efeito condicional da variável trimestre do último contato.

Para a variável que indica o trimestre do último contato, Pela Figura 6.16, nota-se certa semelhança entre os trimestres 1, 2 e 4, este último com um intervalo de credibilidade mais amplo que os trimestres 1 e 2. As estimativas pontuais também são muito próximas, com probabilidades inferiores a 20%. Já o terceiro trimestre se destaca perante os outros, com uma estimativa pontual bem maior (cerca de 40%) e um intervalo de credibilidade também bem mais amplo e uma certa intersecção com um intervalo do quarto trimestre. Assim, podemos dizer, juntamente com o coeficiente estimado maior em relação aos outros trimestres, que os clientes contatados pela última vez no terceiro trimestre do ano têm mais probabilidade de adquirir o produto depósito a prazo que os clientes contatados nos trimestres 1, 2 e 4.

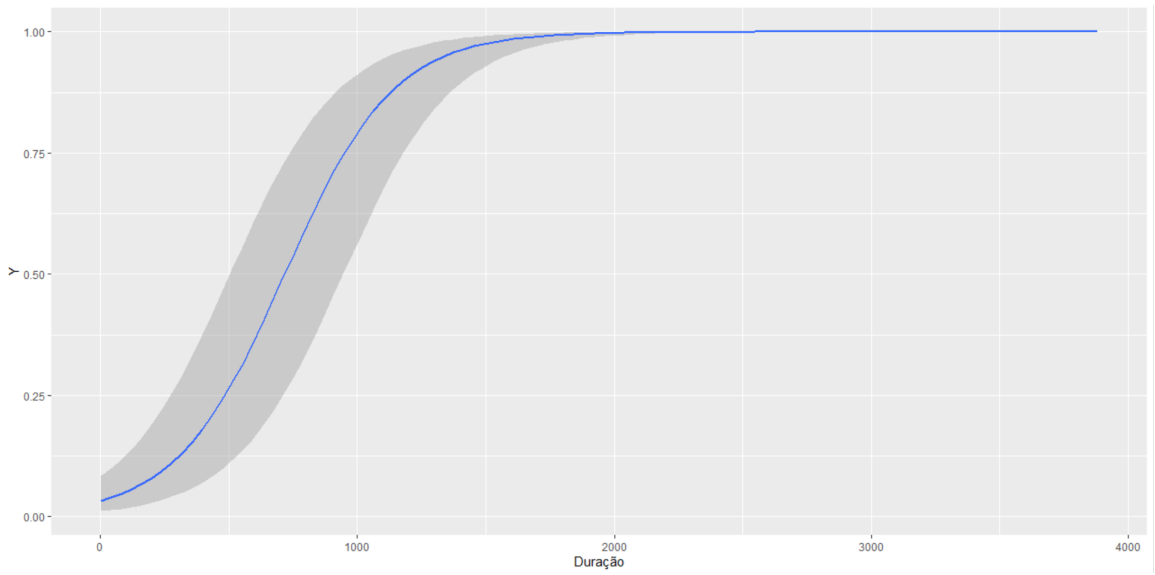

Figura 6.17: Efeito condicional da variável duração do último contato.

Em relação a variável duração do último contato em segundos, pela Figura 6.17, nota-se que quanto maior a duração do último contato, maior a probabilidade de o cliente adquirir o produto depósito a prazo. A probabilidade de um cliente que permaneceu na última ligação por aproximadamente 500 segundos (aproximadamente 8 minutos) é de 25%. Essa probabilidade aumenta para 50% para clientes que permaneceram na última ligação por aproximadamente 700 segundos (aproximadamente 12 minutos) e para 75% para clientes que permaneceram na última ligação por aproximadamente 950 segundos (aproximadamente 15 minutos). A partir de aproximadamente 1750 segundos (aproximadamente 29 minutos) de duração da última ligação, a probabilidade de compra do depósito a prazo atinge 100% e se mantém constante.

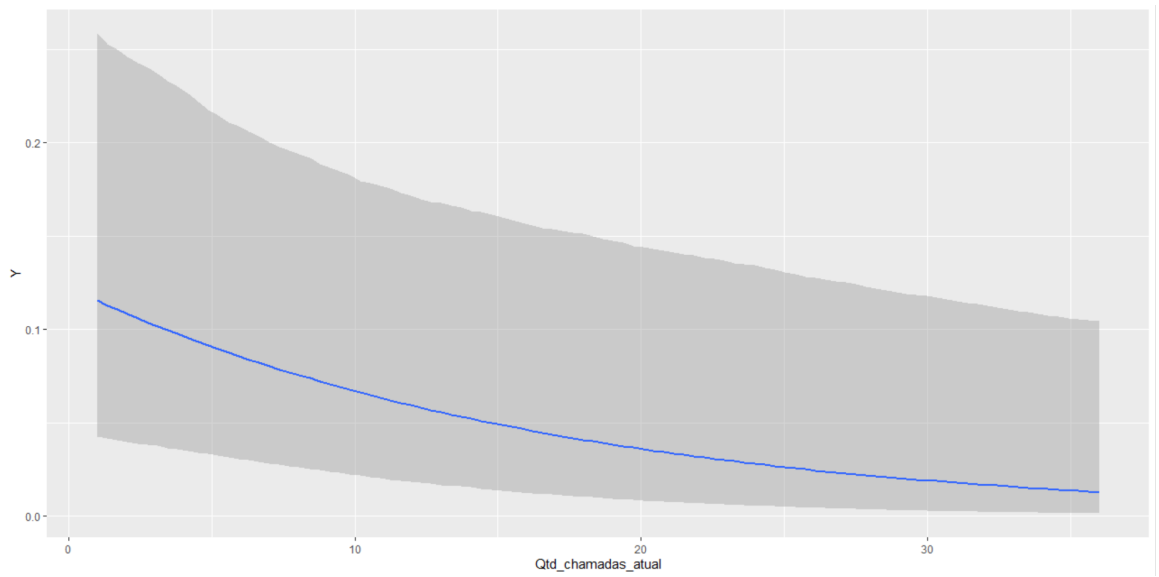

Figura 6.18: Efeito condicional da variável número de ligações na campanha de marketing atual.

Para a variável número de ligações na campanha de marketing atual nota-se, pela Figura 6.18, que a probabilidade de compra do depósito a prazo diminui à medida que o número de ligações nesta campanha atual aumenta, indicando que insistir nos clientes que já receberam muitas ligações não fará com que os mesmos mudem de ideia e adquiram o produto. Talvez fosse interessante considerar a definição de um ponto de corte para esta variável, determinando um número de ligações limite para o mesmo cliente que, se atingido, a empresa poderia considerar parar de realizar outras ligações oferecendo o produto para esse cliente. Porém, deve ressaltar-se que, pelo baixo coeficiente estimado para esta variável, exibido na Tabela 6.5, esta variável não seja tão significativa no modelo, de forma geral.

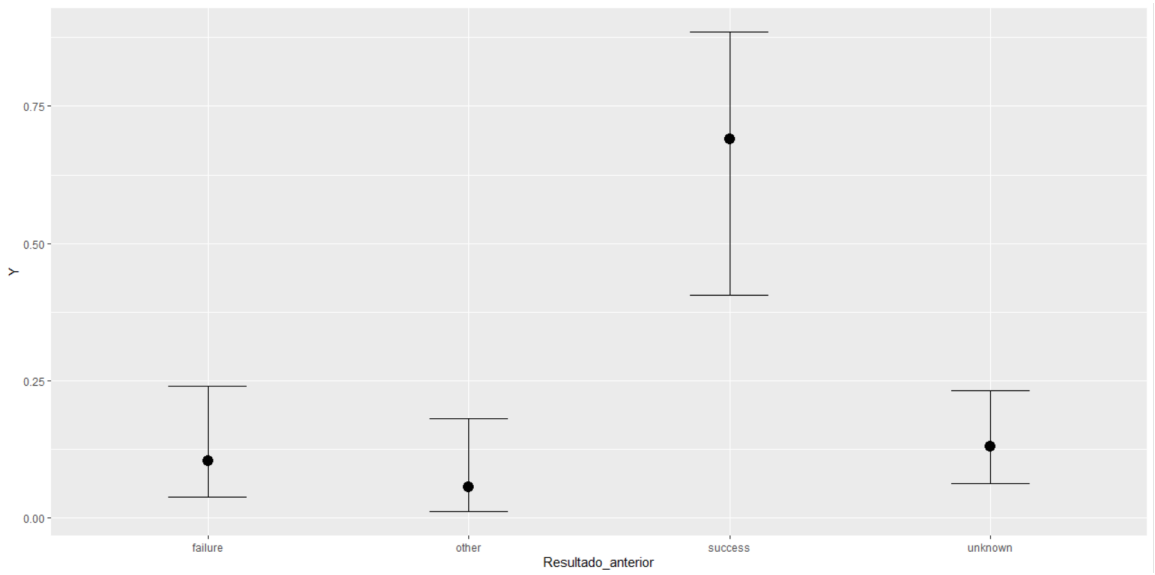

Figura 6.19: Efeito condicional da variável que indica se o cliente comprou o produto depósito a prazo na última campanha.

Para a variável que indica se o cliente realizou a compra do produto depósito a prazo na última campanha de marketing, pela Figura 6.19, nota-se grande semelhança para as categorias que indicam que o cliente não adquiriu o produto ou que o resultado da última campanha foi desconhecido, com estimativas e intervalos de credibilidade muito próximos. Por outro lado, há grande destaque para a categoria que indica que o cliente adquiriu o produto na campanha anterior, com uma estimativa pontual bem maior, porém com um intervalo de credibilidade também mais amplo, ainda que este não possua intersecção com os intervalos das outras categorias. Isso, aliado com o coeficiente estimado bem maior que aqueles das outras categorias, exibidos na Tabela 6.5, indica que os clientes que compraram o produto na última campanha têm mais probabilidade de realizar a compra novamente nesta campanha, algo que também havia sido observado na Seção 5.1.

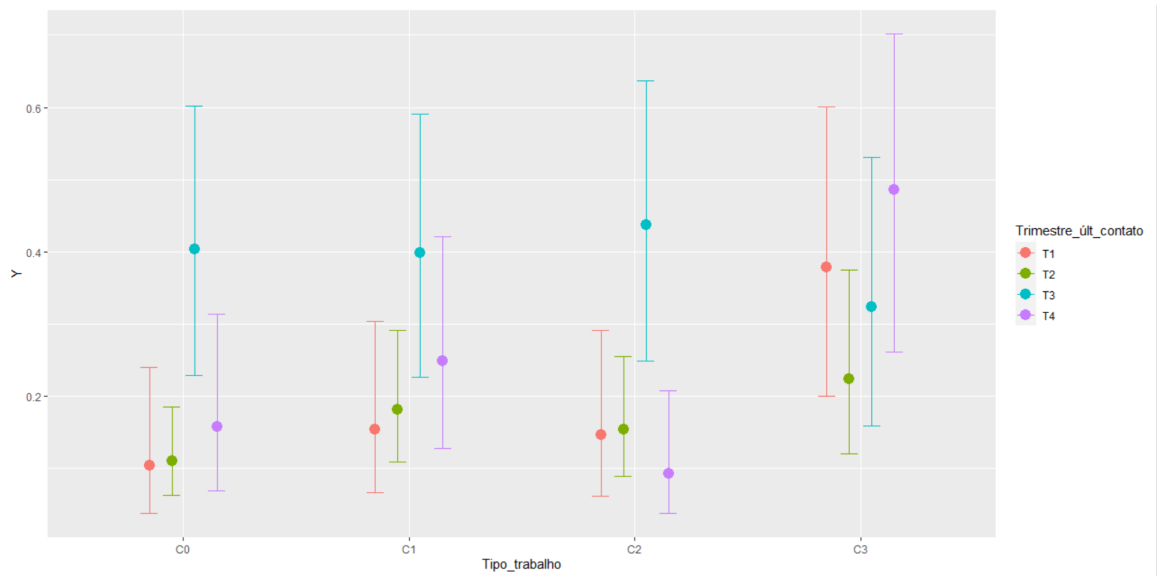

Figura 6.20: Efeito condicional da interação entre as variáveis tipo de trabalho e trimestre do último contato.

Pela Figura 6.20, podemos analisar o efeito da interação entre as variáveis tipo de trabalho e o trimestre do último contato com o cliente. Podemos observar que o terceiro trimestre do ano se destaca em todas as categorias de trabalho, com maiores probabilidades de aquisição do produto de depósito a prazo em relação aos outros trimestres. Na categoria de trabalho C3, a probabilidade de aquisição do depósito a prazo é maior para todos os trimestres do ano, comprados às outras categorias.

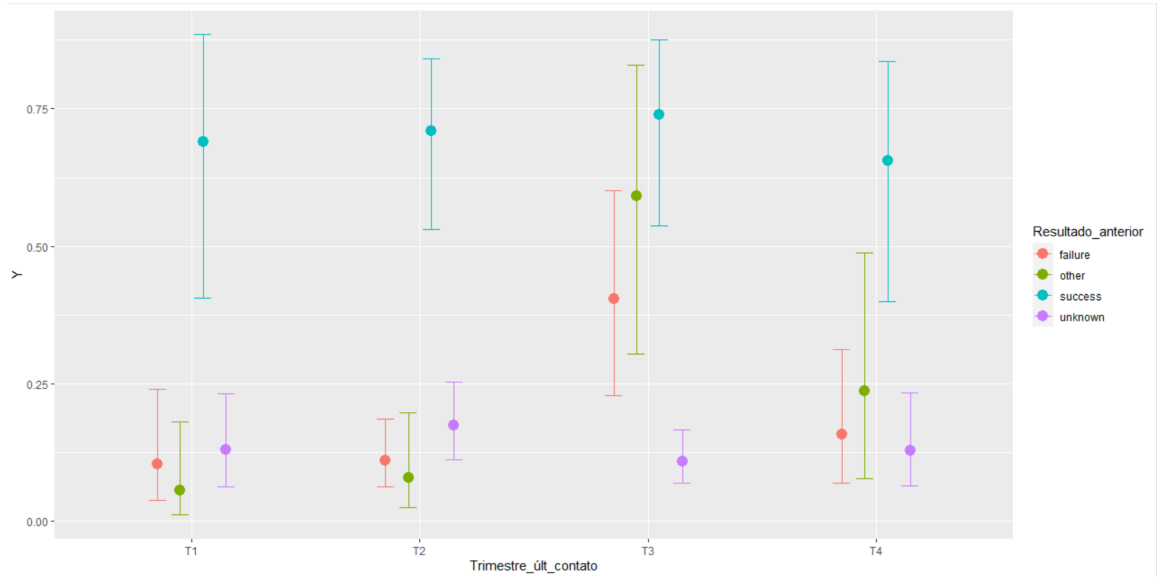

Figura 6.21: Efeito condicional da interação entre a variável trimestre do último contato e a variável que indica se o cliente comprou o produto depósito a prazo na última campanha.

Pela Figura 6.21, podemos analisar o efeito da interação entre as variáveis trimestre do último contato com o cliente e o resultado da última campanha de marketing. Aqui, em relação ao trimestre do último contato, notamos grande destaque para o terceiro trimestre, com maiores probabilidades de aquisição do depósito a prazo independentemente dos resultados da última campanha. Além disso, podemos notar que os clientes que realizaram a compra do produto na campanha anterior têm probabilidade de realizar a compra do produto novamente, independentemente do trimestre do último contato.

6.3 Estimador bayesiano - Modelo de mistura com K = 2

Podemos introduzir a mistura de distribuições no modelo ajustado na Seção 6.2 por meio da função *mixture()* do pacote *brms*. Para isso, basta definir o argumento

```
family = mixture(bernoulli, bernoulli)
```

para considerar que a variável resposta Y é proveniente de duas distribuições de Bernoulli diferentes no modelo. Assim, o código utilizado para ajustar o modelo de mistura com $K = 2$ é dado por

```
fit_k2 <- brm(y ~ age + job + marital + housing + loan + contact +
            month + duration + campaign + poutcome, job:month +
            month:poutcome,
          family = mixture(bernoulli, bernoulli),
          data = TB_BANK_SF,
          chains = 4,
          cores = 4,
          iter = 4000,
          warmup = 2000,
          seed = 123)
```

É importante ressaltar que o modelo de mistura implementado no *Stan*, devido à sua maior complexidade, exige mais esforço computacional que o modelo apresentado em 6.2, que não considera a mistura de distribuições. Diante disso, optamos por utilizar mais

cadeias de Markov e mais iterações em cada uma das cadeias e, consequentemente, o tempo necessário para ajustar o modelo foi muito maior.

Ao tentar realizar o ajuste considerando a mistura de duas distribuições de Bernoulli, utilizamos 4 cadeias de Markov, 4000 iterações, sendo 2000 na fase de *warmup (burn-in)*, o tempo total para executar o código foi de mais de 8 horas e mesmo assim não foi possível obter a convergência.

Na Seção seguinte, analisaremos os traceplots para tentar verificar o(s) motivo(s) da não convergência.

6.3.1 Análise gráfica da não convergência

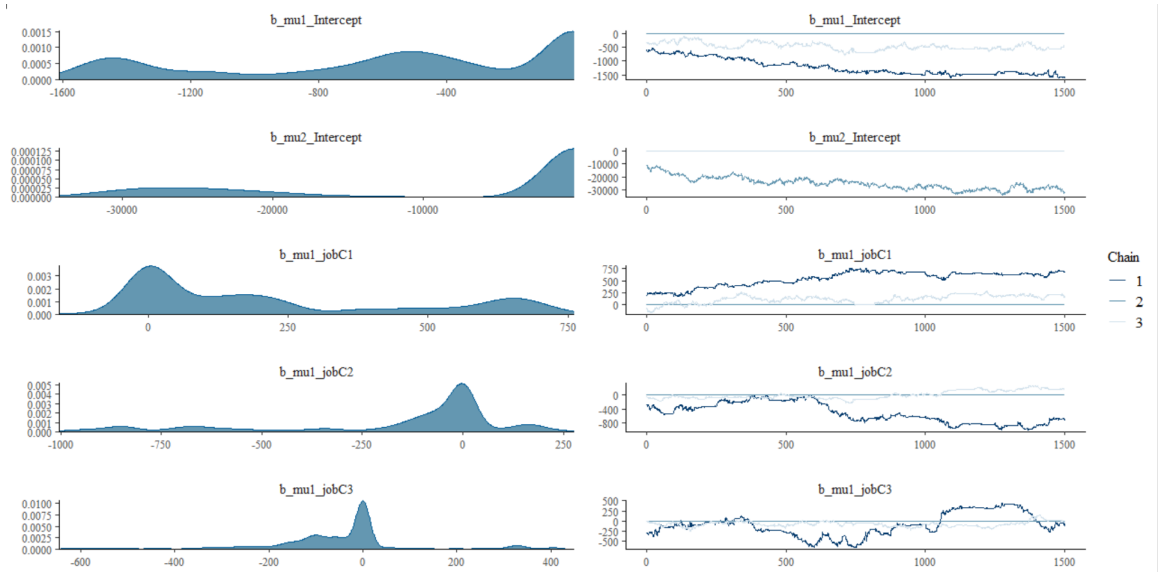

Figura 6.22: Densidades das distribuições a posteriori dos parâmetros e traceplots das Cadeias de Markov para o modelo de mistura aplicado nos dados reais com $K = 2$.

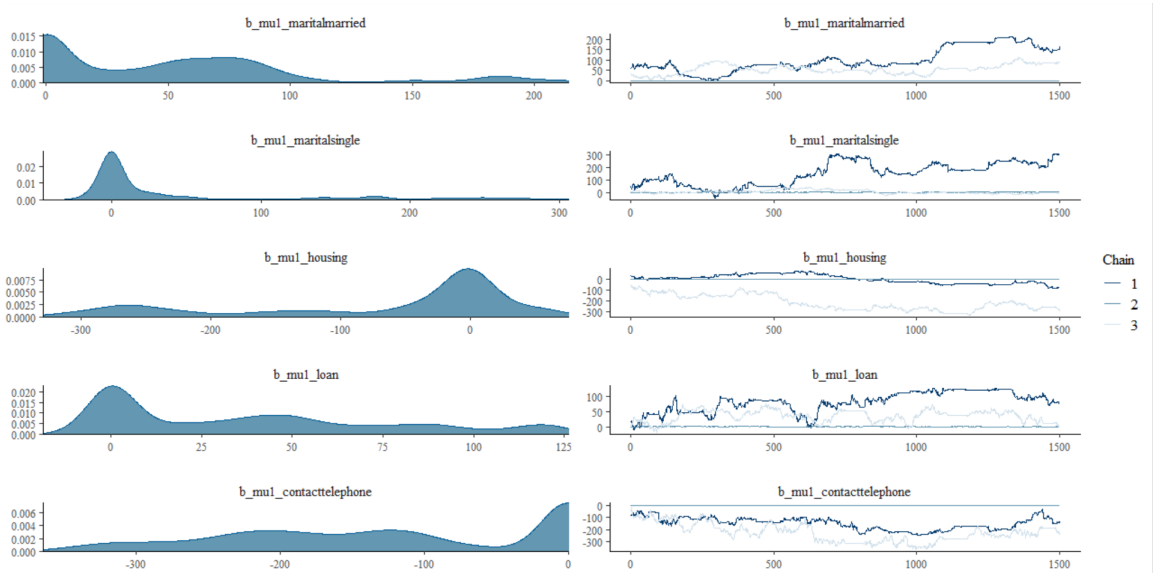

Figura 6.23: Densidades das distribuições a posteriori dos parâmetros e traceplots das Cadeias de Markov para o modelo de mistura aplicado nos dados reais com $K = 2$.

Analisando os traceplots de alguns dos parâmetros (o restante dos gráficos pode ser visualizado no Apêndice B), à direita nas Figuras B.1 e B.2, podemos ver que o método MCMC considerando duas componentes na mistura, não convergiu. Uma explicação plausível para este resultado é a ocorrência da não identificabilidade dos parâmetros, citada na Seção 4.1. Assim, podemos concluir que o modelo de mistura com $K = 2$ não é adequado para o banco de dados utilizado, ou seja, o verdadeiro valor de K é 1, isto é, os dados são provenientes de apenas uma distribuição de Bernoulli, e não duas. Dessa forma, até o presente momento da análise, concluímos que o modelo mais adequado para este banco de dados é o modelo que considera apenas uma componente ($K = 1$ ou não mistura), explorado na Seção 6.2.

Capítulo 7

Conclusões

Em resumo, primeiramente, na Seção 5.1, as variáveis tipo de trabalho e mês do último contato foram remodeladas com o intuito de diminuir a dimensão e facilitar a interpretação das mesmas. Além disso, foram destacadas algumas variáveis que poderiam ser úteis durante o processo de modelagem, como a categoria de trabalho do cliente, resultado da última campanha, e mês do último contato. Destas, todas realmente mostraram-se úteis no modelo final, com alta significância e *odds ratio*.

Na Seção 6.1, foi ajustado o modelo utilizando a função de ligação logit considerando todas as variáveis independentes e todas as interações duplas entre as variáveis que se destacaram durante a Análise Descritiva. Utilizando o método Stepwise AIC para a seleção de covariáveis do modelo, descartamos parte das covariáveis do modelo inicial que demonstraram não ser significativas para discriminar os compradores do produto daqueles que não realizaram a compra, diminuindo, assim, a complexidade do modelo sem prejudicar a qualidade do mesmo.

A distribuição de Bernoulli sugerida ao estudo e, provavelmente a única que se encaixa no problema, combinada com a função de ligação logit apresentou bom desempenho e produzindo valores consideravelmente altos para a AUC e a Estatística KS, métricas utilizadas para avaliar a qualidade do ajuste.

Na Seção 6.1.5, foi utilizada a *odds ratio* para interpretar os coeficientes obtidos com o ajuste do modelo, evidenciando categorias de algumas variáveis que obtiveram altos valores para essa métrica e, dessa forma, foi possível identificar características significativas no que diz respeito a probabilidade dos clientes realizarem a aquisição do depósito a prazo disponibilizado pela instituição financeira em relação aos clientes que não possuem essas características. Esses resultados podem ser considerados para aprimorar as investidas que

venham a ser realizadas em uma campanha de marketing futura e, assim, aumentar a taxa de vendas por clientes prospectados.

Na Seção 6.2, foi realizado o ajuste do modelo logístico utilizando o estimador bayesiano por meio do método MCMC (HMC - Monte Carlo Hamiltoniano) implementado no software *Stan*. O resultado desta implementação foi a convergência do método para $K = 1$ (uma componente na mistura ou não mistura). Já para a mistura de duas componentes ($K = 2$), o método não convergiu. Isso motivou a busca por uma explicação plausível para essa não convergência.

A explicação para a não convergência foi feita através de um procedimento de simulação em que as duas variações da mistura ($K = 1$ e $K = 2$) foram simuladas através de um simulador de dados para a mistura de modelos logísticos para testar o ajuste fornecido pelo software *Stan*. Testamos essas duas condições simulando dados com 1000 observações em cada condição e ajustamos o modelo logístico para $K = 1$ e o modelo de mistura logística para $K = 2$. Obtivemos convergência em ambos os casos.

Quando submetemos o conjunto de dados para $K = 1$ ao ajuste com $K = 2$, o mesmo não convergiu. A explicação para a não convergência é a não identificabilidade devido ao sobre ajuste, para mais detalhes ver Frühwirth-Schnatter (2010) apud Macerau (2023).

Embora no caso dos dados simulados sabemos do sobre ajuste devido à natureza simulada dos dados, convém lembrar que o sobre ajuste não é a única causa para a não convergência do modelo.

Na aplicação realizada nos dados reais, os métodos tratados tiveram bom desempenho para apenas uma componente ($K = 1$). Numa comparação rápida, as estimativas em ambos os métodos (máxima verossimilhança e bayesiano) foram bem próximas. Já para $K = 2$, não obtivemos convergência para o método HMC. Uma explicação plausível para a não convergência é a não identificabilidade devido ao sobre ajuste, citada na Seção 4.1. Isso nos levou à conclusão que o modelo mais adequado para este conjunto de dados até o presente momento da análise é o modelo para uma componente ($K = 1$).

Referências Bibliográficas

Agresti, A. (2002). *Categorical Data Analysis*. John Wiley & Sons.

Frühwirth-Schnatter, S.; Pyne, S. (2010). *Bayesian inference for finite mixtures of univariate and multivariate skew-normal and skew-t distributions*. Tese de doutorado.

Gonzalez, L. d. A. (2002). Regressão logística e suas aplicações. página 5.

Macerau, W. M. O. (2023). *Métodos Bayesianos para seleção de modelos de mistura de distribuições normais e t de Student assimétricas*. Tese de doutorado.

Radford, N. (2011). *MCMC using Hamiltonian dynamics*. In Handbook of Markov Chain Monte Carlo, edited by Steve Brooks, Andrew Gelman, Galin L. Jones, and Xiao-Li Meng, 116–62. Chapman; Hall/CRC.

Reis, E.A., R. I. (2002). Análise descritiva de dados. *Departamento de Estatística da UFMG*, página 5.

Stan (2024a). Stan. Disponível em: <https://mc-stan.org/>. Acessado em: 17 de janeiro de 2024.

Stan (2024b). Stan reference manual. Disponível em: <https://mc-stan.org/docs/reference-manual/index.html>. Acessado em: 17 de janeiro de 2024.

Wikipedia (2024). Função logística. Disponível em: <https://pt.wikipedia.org/wiki/Fun%C3%A7%C3%A3o_log%C3%ADstica>. Acessado em: 05 de janeiro de 2024.

Wikipédia (2024). Stan (software). Disponível em: <https://pt.wikipedia.org/wiki/Stan_(software)>. Acessado em: 17 de janeiro de 2024.

Meira, Silvana Aparecida (2014). Modelo de Mistura com Dependência Markoviana de Primeira Ordem. Tese de doutorado.

Saraiva, Erlandson Ferreira (2009). Modelo de mistura com número de componentes desconhecido: estimação via método split-merge. Tese de doutorado.

Didier Chauveau, Bernard Garel, Sabine Mercier (2018). Testing for univariate two-component Gaussian mixture in practice.

Apêndice A

Simulador de dados para mistura de modelos logísticos.

```
#
#   simula dados:   K=2    #
#
#       rm(list=ls())
#
#     pc: Prob. das componentes
#     pa: Parâmetros das componentes
#     ns: Número de valores simulados
#
Simula_Mistura_Logistica2Comp=function( ns, pc=c(.6,.4), pa)  {
   lg=function(x,b) 1/(1+exp(-b[2]*(x-b[1])))
   K=length(pc)
   #
   x= seq(-10,10,.1)
   lx=numeric(0)
   for (s in 1:K) lx=cbind(lx,lg(x,b=pa[s,]))
   lx=cbind(lx,lx%*%pc)
   matplot(x,lx,type='l',ylab='Pr(Yi=1)',main='Funções Logísticas (K=2)',
           col=c(1,2,4,1),ylim=c(-.2,1.2))
   #
   # passo 1: simular variável que identifica componente, 1 ou 2 (binomial).
```

```
   s = sample(c(1,2), size = ns, replace = TRUE, prob = pc)
   #
   rlgM=function(ns,s,pa) log(1/runif(ns)-1)/pa[s,2] + pa[s,1]
   lgC=function(lM,s,pa) 1/(1+exp(-pa[s,2]*(lM-pa[s,1])))
   #
   lM=rlgM(ns,s,pa)
   pC=lgC(lM,s,pa)
   y = rbinom(ns,1,prob=pC);
   plot(lM,y,pch='o',ylab='Observ.',main='Dados (K=2)',
        col=c(1,4)[s],ylim=c(-.2,1.2), xlab='x')
   points(lM,rep(-.12,ns),pch='I')
   # points(lM,pC,pch='x')
   cbind(s,lM,pC,y)
}
#
#
# Parâmetros para as curvas logísticas
pa = matrix(c(2,1.2,-2,.8),nrow = 2, ncol = 2, byrow = TRUE);  pa
ns = 200
set.seed(86)
Simula_Mistura_Logistica2Comp(ns, pc=c(.7,.3), pa)
```

Apêndice B

Traceplots do modelo de mistura para $K = 2$.

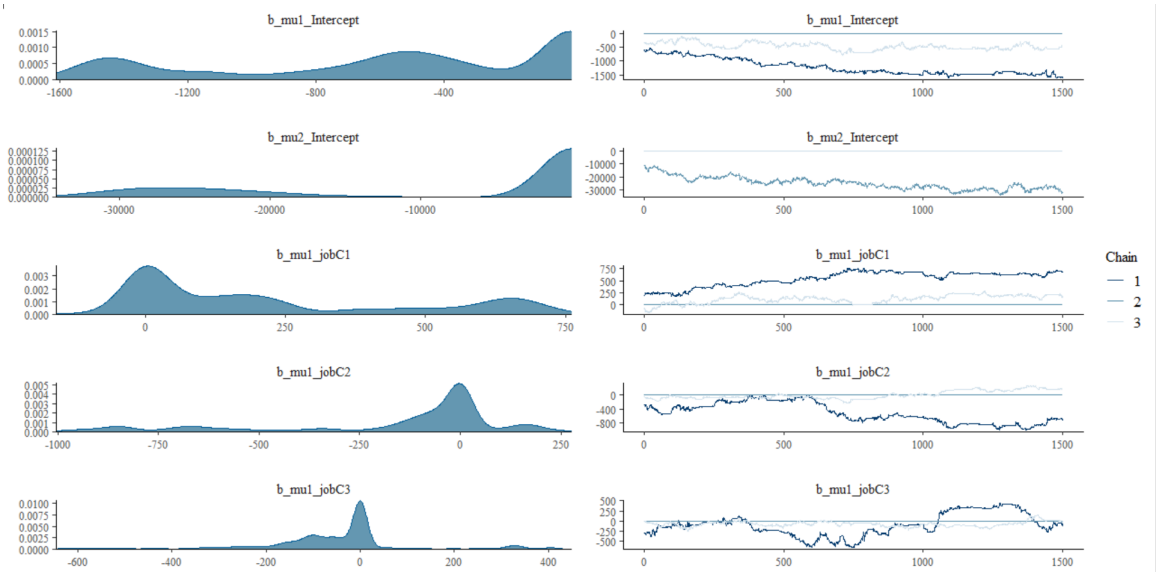

Figura B.1: Densidades das distribuições a posteriori dos parâmetros e traceplots das Cadeias de Markov para o modelo de mistura aplicado nos dados reais com $K = 2$.

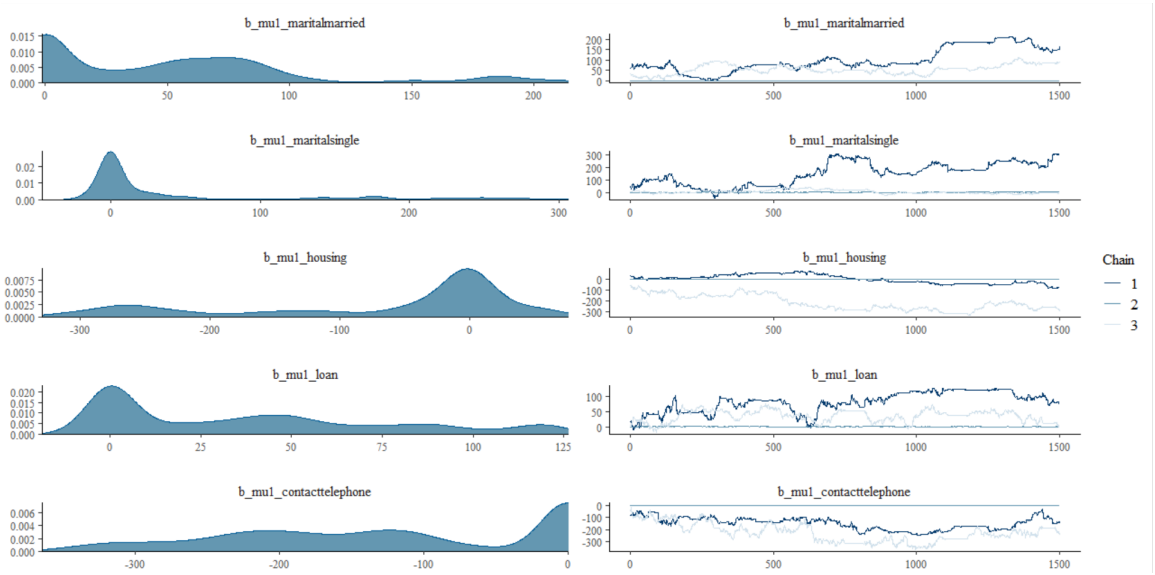

Figura B.2: Densidades das distribuições a posteriori dos parâmetros e traceplots das Cadeias de Markov para o modelo de mistura aplicado nos dados reais com $K = 2$.

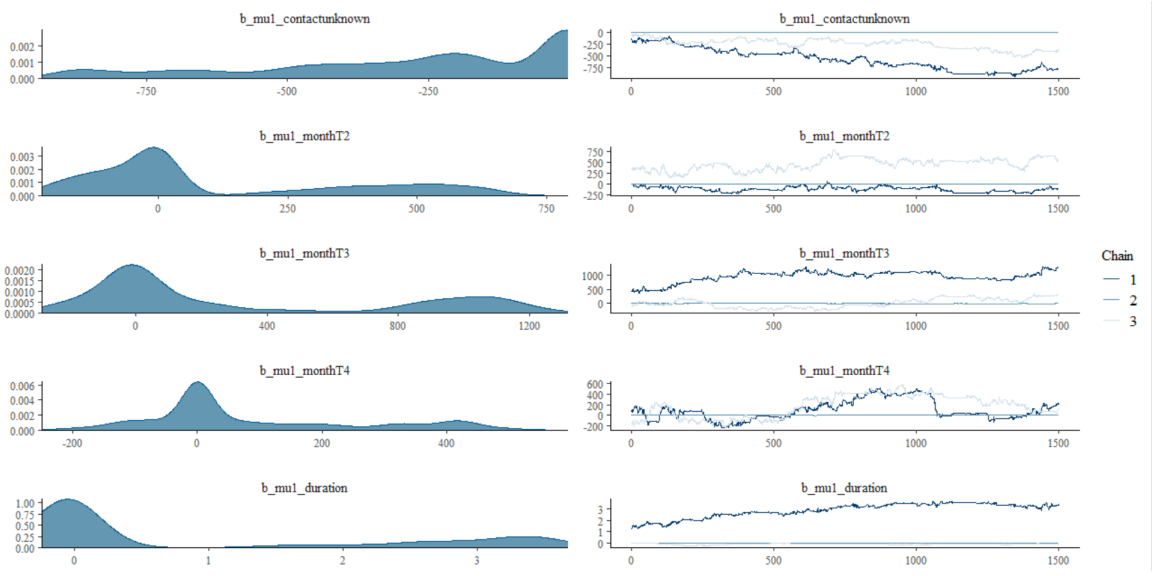

Figura B.3: Densidades das distribuições a posteriori dos parâmetros e traceplots das Cadeias de Markov para o modelo de mistura aplicado nos dados reais com $K = 2$.

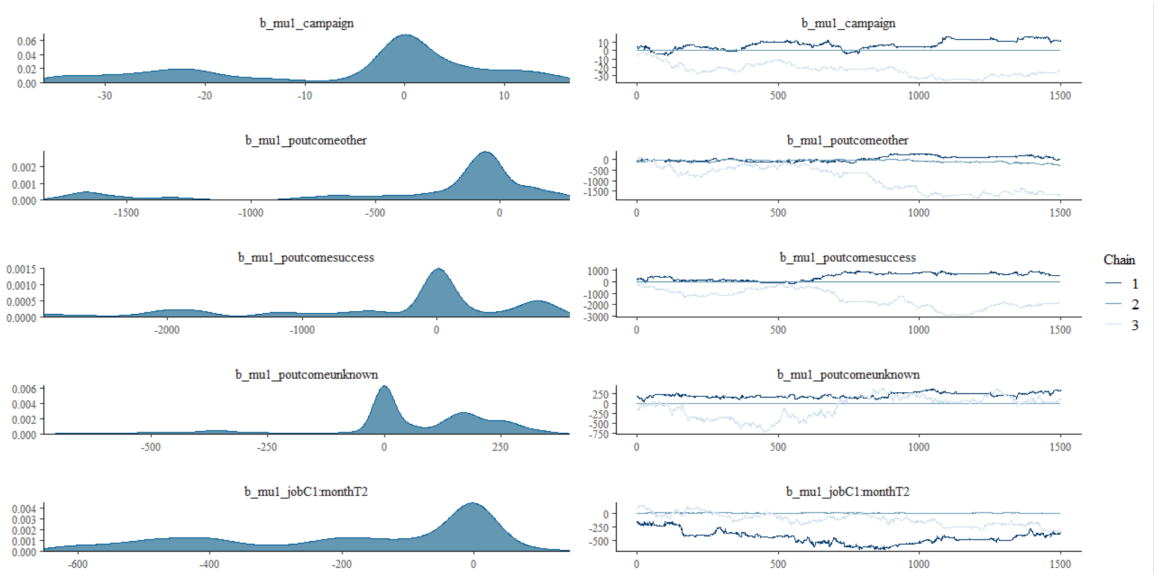

Figura B.4: Densidades das distribuições a posteriori dos parâmetros e traceplots das Cadeias de Markov para o modelo de mistura aplicado nos dados reais com $K = 2$.

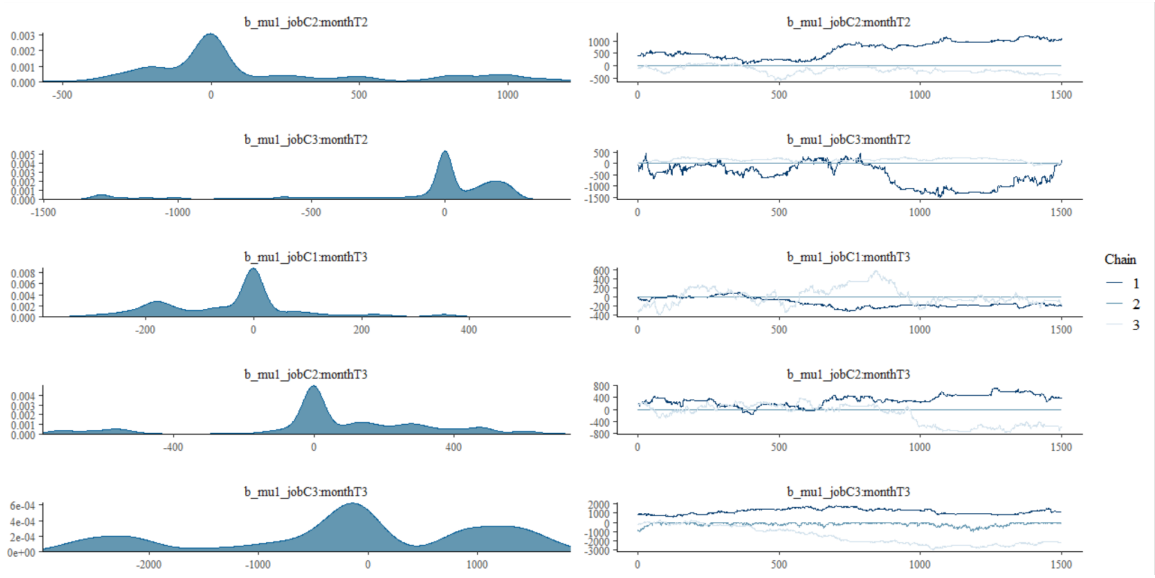

Figura B.5: Densidades das distribuições a posteriori dos parâmetros e traceplots das Cadeias de Markov para o modelo de mistura aplicado nos dados reais com $K = 2$.

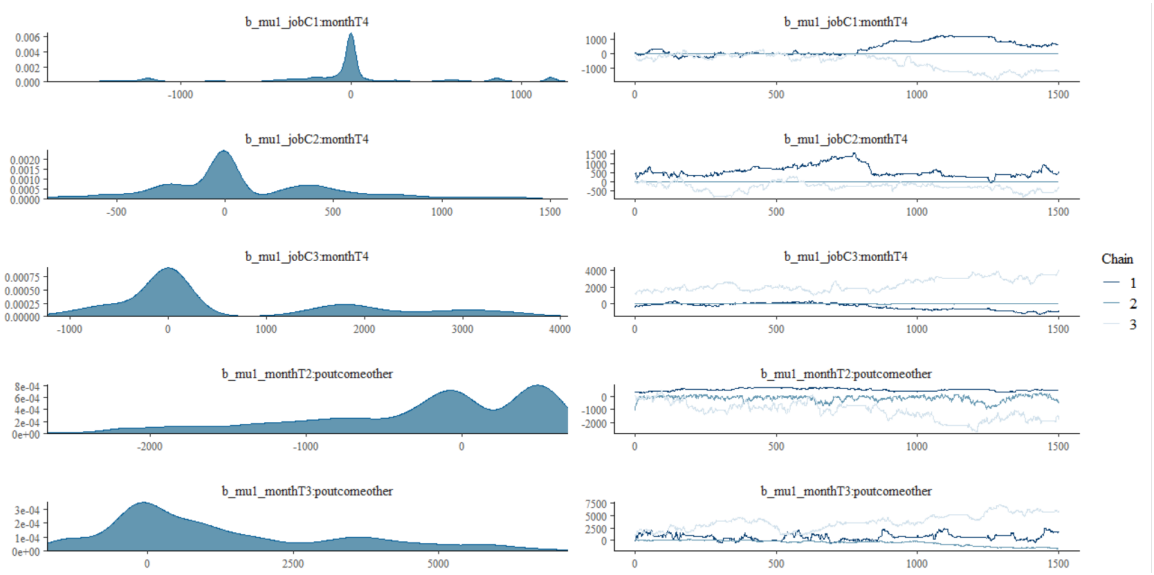

Figura B.6: Densidades das distribuições a posteriori dos parâmetros e traceplots das Cadeias de Markov para o modelo de mistura aplicado nos dados reais com $K = 2$.

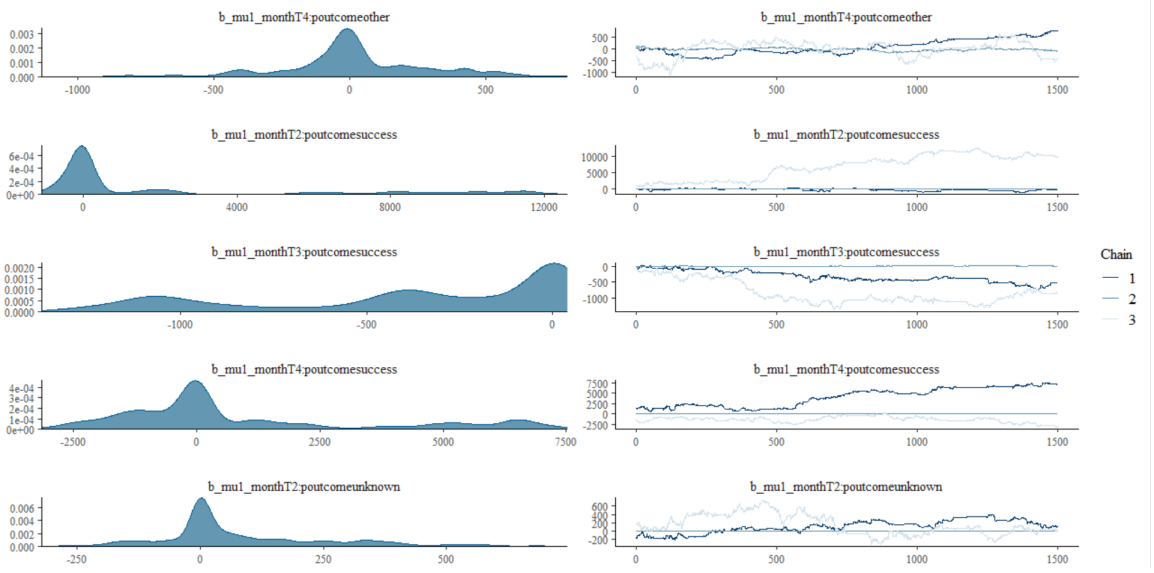

Figura B.7: Densidades das distribuições a posteriori dos parâmetros e traceplots das Cadeias de Markov para o modelo de mistura aplicado nos dados reais com $K = 2$.

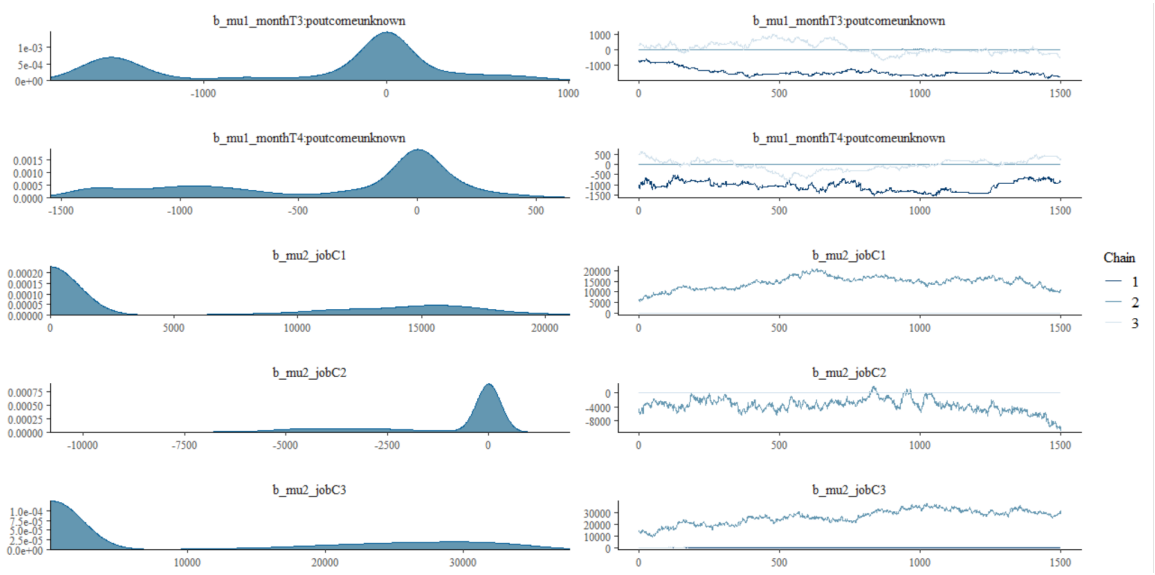

Figura B.8: Densidades das distribuições a posteriori dos parâmetros e traceplots das Cadeias de Markov para o modelo de mistura aplicado nos dados reais com $K = 2$.

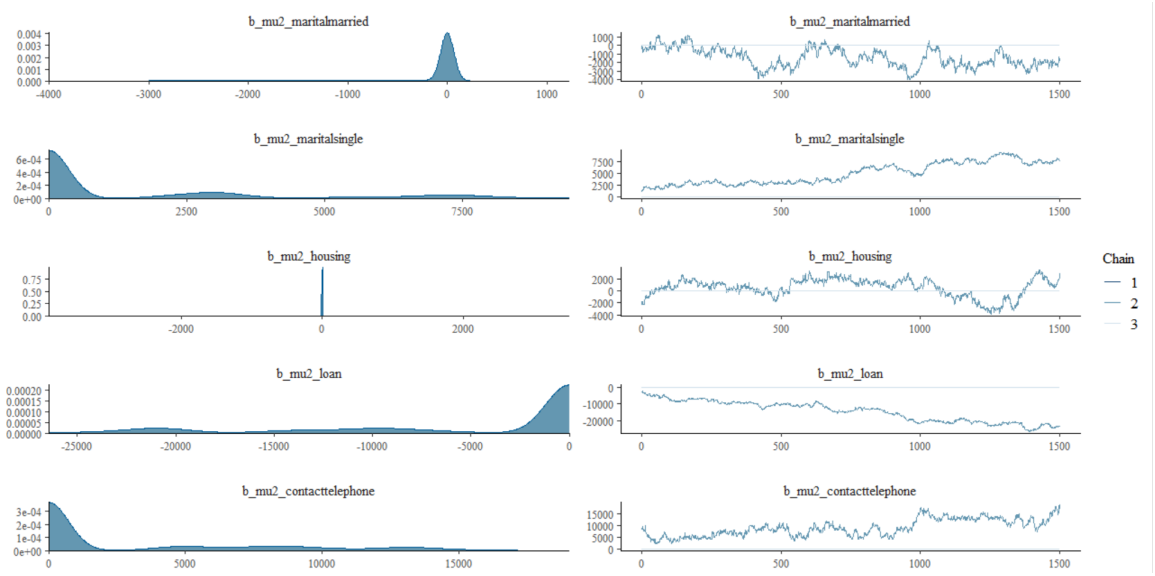

Figura B.9: Densidades das distribuições a posteriori dos parâmetros e traceplots das Cadeias de Markov para o modelo de mistura aplicado nos dados reais com $K = 2$.

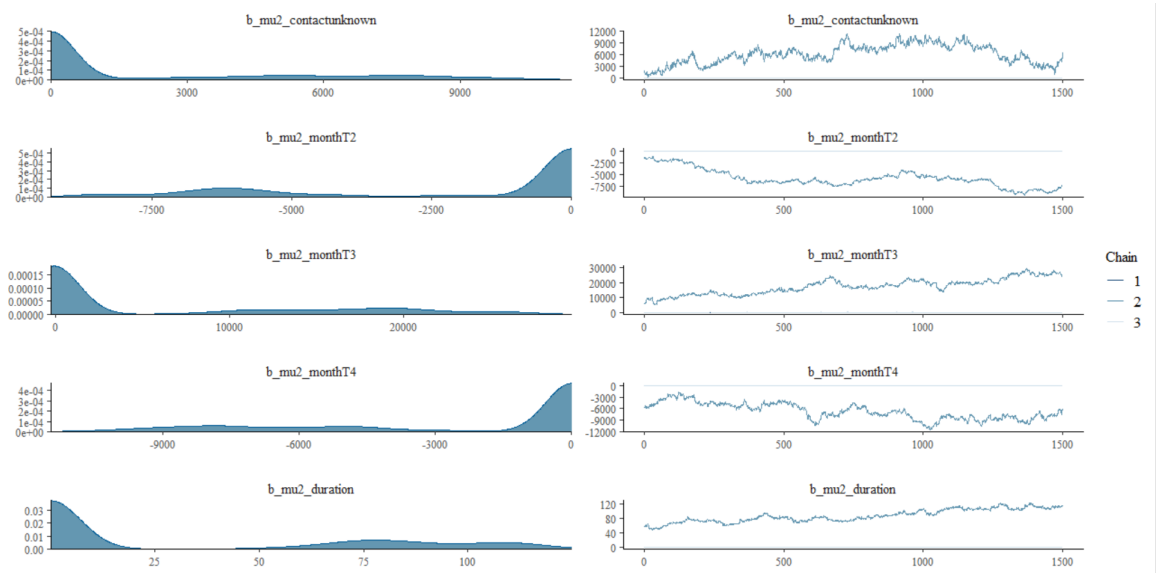

Figura B.10: Densidades das distribuições a posteriori dos parâmetros e traceplots das Cadeias de Markov para o modelo de mistura aplicado nos dados reais com $K = 2$.

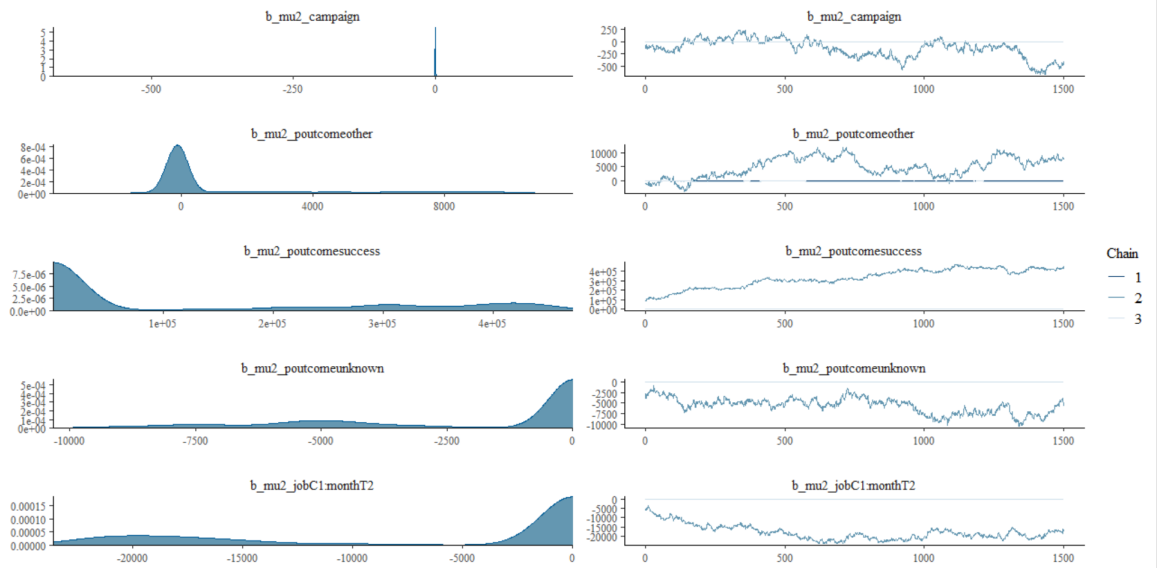

Figura B.11: Densidades das distribuições a posteriori dos parâmetros e traceplots das Cadeias de Markov para o modelo de mistura aplicado nos dados reais com $K = 2$.

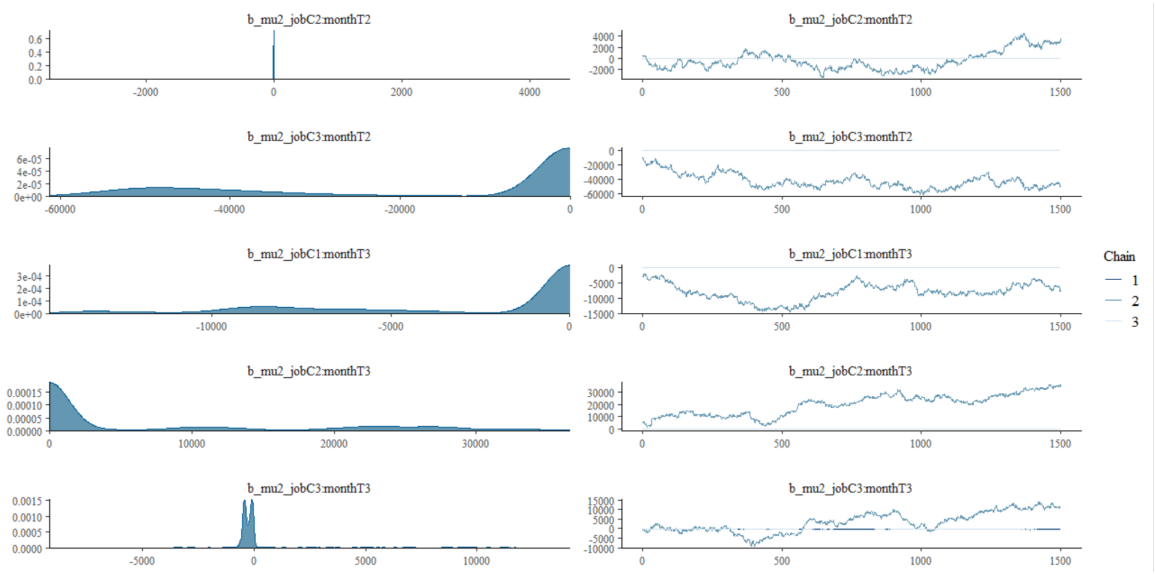

Figura B.12: Densidades das distribuições a posteriori dos parâmetros e traceplots das Cadeias de Markov para o modelo de mistura aplicado nos dados reais com $K = 2$.

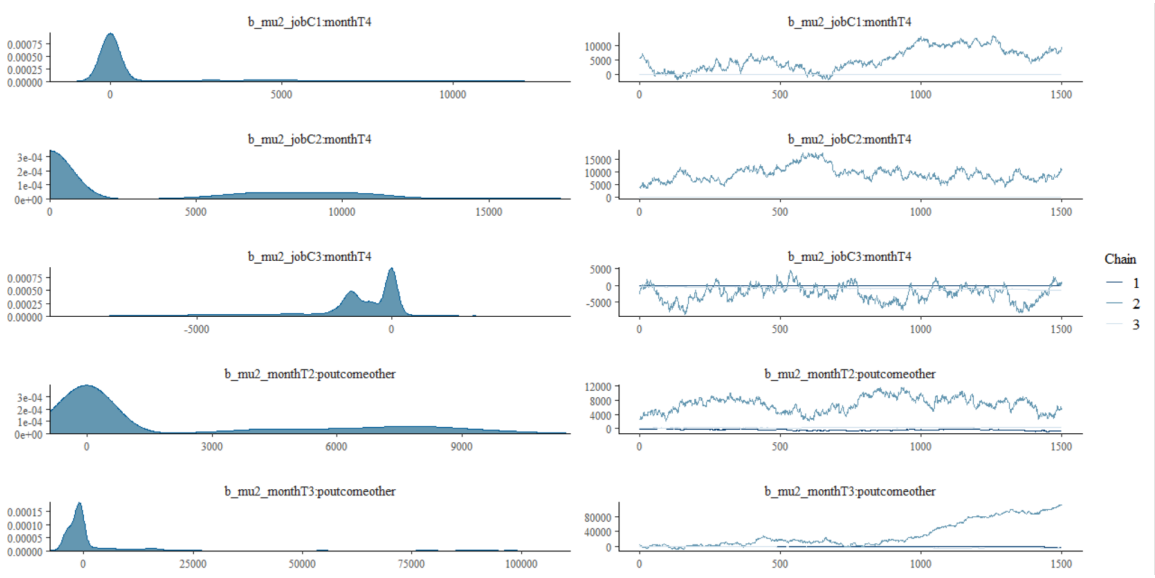

Figura B.13: Densidades das distribuições a posteriori dos parâmetros e traceplots das Cadeias de Markov para o modelo de mistura aplicado nos dados reais com $K = 2$.

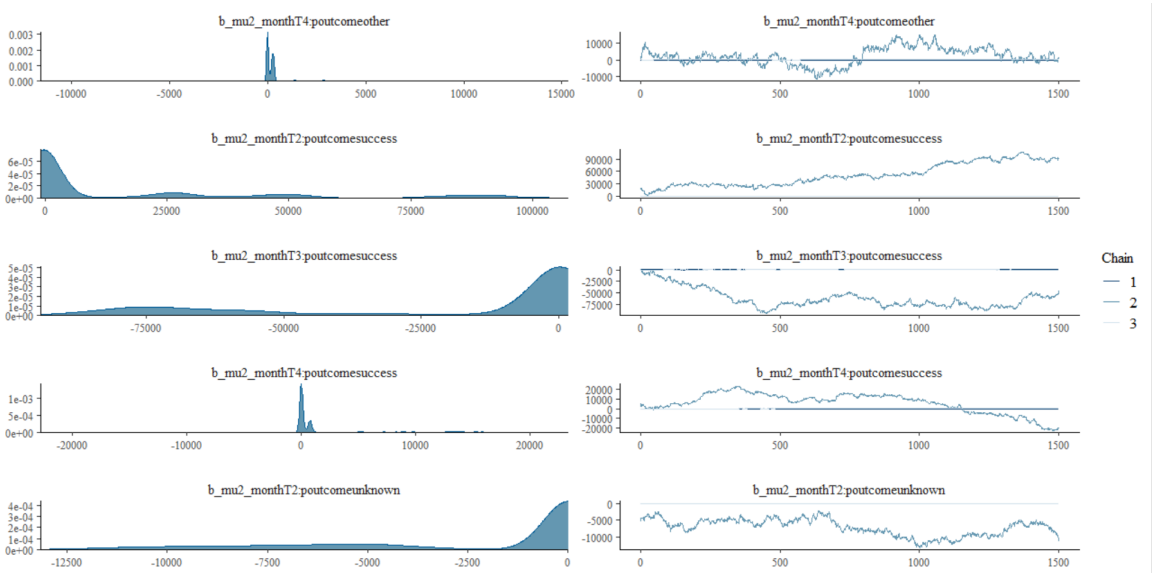

Figura B.14: Densidades das distribuições a posteriori dos parâmetros e traceplots das Cadeias de Markov para o modelo de mistura aplicado nos dados reais com $K = 2$.

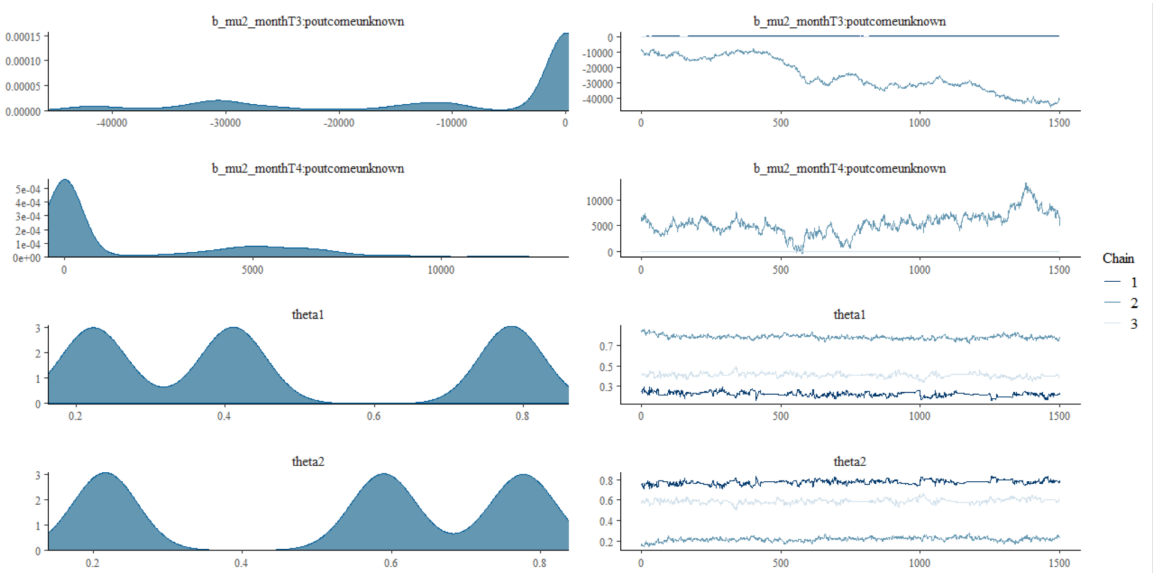

Figura B.15: Densidades das distribuições a posteriori dos parâmetros e traceplots das Cadeias de Markov para o modelo de mistura aplicado nos dados reais com $K = 2$.