UNIVERSIDADE FEDERAL DO CEARÁ

CENTRO DE CIÊNCIAS

DEPARTAMENTO DE ESTATÍSTICA E MATEMÁTICA APLICADA

PROGRAMA DE PÓS-GRADUAÇÃO EM MODELAGEM E MÉTODOS QUANTITATIVOS

LÍVIA DE OLIVEIRA ALVES

ABORDAGEM BAYESIANA PARA TRATAMENTO DE DADOS FALTANTES COM APLICAÇÃO EM UM MODELO LOGÍSTICO.

FORTALEZA

2019

LÍVIA DE OLIVEIRA ALVES

ABORDAGEM BAYESIANA PARA TRATAMENTO DE DADOS FALTANTES COM APLICAÇÃO EM UM MODELO LOGÍSTICO.

Dissertação apresentada ao Programa de Pós-graduação em Modelagem e Métodos Quantitativos do do Centro de Ciências da Universidade Federal do Ceará, como requisito parcial à obtenção do título de mestre em Modelagem e Métodos Quantitativos. Área de Concentração: Métodos Quantitativos.

Orientador: Prof. Dr. José Ailton Alencar Andrade

Co-Orientador: Prof. Dr. Leandro Chaves Rêgo

FORTALEZA

2019

Dados Internacionais de Catalogação na Publicação
Universidade Federal do Ceará
Biblioteca Universitária
Gerada automaticamente pelo módulo Catalog, mediante os dados fornecidos pelo(a) autor(a)

A48a Alves, Lívia de Oliveira.
Abordagem bayesiana para tratamento de dados faltantes com aplicação em um modelo logístico. / Lívia de Oliveira Alves. – 2019.
124 f. : il. color.

Dissertação (mestrado) – Universidade Federal do Ceará, Centro de Ciências, Programa de Pós-Graduação em Modelagem e Métodos Quantitativos, Fortaleza, 2019.
Orientação: Prof. Dr. José Ailton Alencar Andrade.
Coorientação: Prof. Dr. Leandro Chaves Rêgo.

1. Dados faltantes. 2. Imputação múltipla. 3. Modelos preditivos. 4. Regressão logística. 5. Métodos bayesianos. I. Título.

CDD 510

LÍVIA DE OLIVEIRA ALVES

ABORDAGEM BAYESIANA PARA TRATAMENTO DE DADOS FALTANTES COM APLICAÇÃO EM UM MODELO LOGÍSTICO.

Dissertação apresentada ao Programa de Pós-graduação em Modelagem e Métodos Quantitativos do do Centro de Ciências da Universidade Federal do Ceará, como requisito parcial à obtenção do título de mestre em Modelagem e Métodos Quantitativos. Área de Concentração: Métodos Quantitativos.

Aprovada em:

BANCA EXAMINADORA

Prof. Dr. José Ailton Alencar Andrade (Orientador)
Universidade Federal do Ceará (UFC)

Prof. Dr. Leandro Chaves Rêgo (Co-Orientador)
Universidade Federal do Ceará (UFC)

Prof. Dr. Anselmo Ramalho Pitombeira
Universidade Federal do Ceará (UFC)

Prof. Dr. Gustavo Leonel Gilardoni Avalle
Universidade de Brasília (UNB)

À minha família, que sempre acreditaram e investiram em mim. Mãe, seu cuidado e dedicação foi que deram, em muitos momentos, a esperança e força para seguir. Pai, sua presença e apoio significou segurança e certeza de que não estou sozinho nessa caminhada.

AGRADECIMENTOS

Meus agradecimentos são a todos os que me apoiaram e me ajudaram durante essa minha jornada acadêmica em especial durante todo este percurso do mestrado:

A Deus, por todas as oportunidades e graças que Ele me forneceu até o presente momento, incluindo a ajuda e força para superar muitos problemas que surgiram durante o mestrado, assim como concluir este trabalho.

Aos meus pais, Maria Antonieta e Francisco Alves pela educação, carinho e suporte para que eu chegasse até esse ponto.

Aos meus familiares, Flávia Franco e Renan Melo pelo apoio.

A Fundação Cearense de Apoio ao Desenvolvimento Científico e Tecnológico (FUNCAP), pelo apoio através do financiamento por meio da bolsa de estudos que a FUNCAP fornece aos alunos de Mestrado.

A Universidade Federal do Ceará (UFC) por fornecer suporte durante o percurso da graduação e do mestrado.

Ao meu orientador e coorientador Aílton Andrade e Leandro Rêgo pela orientação e paciência durante todo o mestrado.

Aos professores que tive a oportunidade de conhecer na graduação e mestrado que foram muito importantes para minha formação, Ana Maria Araújo, André Jalles, Jacqueline Batista, Juvêncio Nobre, Maurício Mota, Sílvia Freitas, Rafael Farias, Ronald Targino e Rosa Maria.

Ao João Brainer Clares de Andrade por obter e compartilhar os dados de modo que tornou possível a aplicação dessa pesquisa.

Aos amigos adquiridos durante a graduação e mestrado pelos momentos de estudo em grupo, Alice Ximenes, Amanda Merian, Armando Dauer, Chagas Junior, Cristina Guedes, Eliene Monteiro, Francílio Araújo, Henrique Sena, Jamila Fernandes, Kennedy Araújo, Raquel Lima, Raul Furtado, Renato Gil, Robert Plant, Rossana del Valle e Wasley Correia.

Especialmente gostaria de agradecer aos meus amigos, Débora Ferreira, Vinícius Osterne e Wendel pela inestimável ajuda durante todo o mestrado de modo que muito provavelmente sem tal ajuda não tivesse conseguido chegar até aqui.

E por fim mas não menos importante, aos meus amigos, Renato Barros e Diego Rafael por ter sido grandes incentivadores.

"Para tudo há um tempo determinado... Tempo para plantar e tempo para arrancar o que se plantou."

(ECLESIASTES 3:1,2.)

RESUMO

Dados faltantes surgem frequentemente em aplicações práticas e podem ocasionar muitos problemas. O impacto dos dados ausentes na modelagem e em inferências estatísticas é iminentemente importante, principalmente em casos em que os sujeitos com dados faltantes possuem padrões de respostas que diferem muito daqueles de dados completos. O tratamento inadequado ou o não tratamento dos dados faltantes também pode afetar os resultados gerais da análise. Existem várias abordagens para enfrentar o problema de informações omissas. Dado este cenário, neste trabalho serão discutidas metodologias de tratamento de dados faltantes em modelos preditivos através de uma aplicação do problema. Para tal desenvolvimento será utilizada a técnica de regressão logística para elaboração de ferramenta preditiva do risco de transformação hemorrágica em pacientes com Acidente Vascular Cerebral isquêmico em uma unidade hospitalar pública de referência em Fortaleza, Ceará, na qual dentre suas covariáveis, algumas delas possuem uma quantidade representativa de dados omissos. Assim, o objetivo principal do estudo é aplicar técnicas diferentes de tratamentos de dados faltantes para cada variável de acordo com sua natureza e ajustar um modelo preditivo e posteriormente comparar com uma base de dados mais completa obtida em outro momento da pesquisa.

Palavras-chave: Dados faltantes. Imputação múltipla. Modelos preditivos. Regressão logística. Métodos bayesianos.

ABSTRACT

Missing data often comes up in practical applications and may cause many problems. The impact of missing data on modeling and statistical inferences is eminently important, especially in the face of subjects with missing data who have response patterns that differ greatly from those with complete data. Inadequate treatment or non-treatment of missing data may also affect the overall results of the analysis. There are several approaches of addressing the missing information problem. In this work, methodologies for missing data treatment in predictive models through an application of the problem are discussed. For this, the logistic regression technique is used to develop a predictive tool for the risk of hemorrhagic transformation in patients with ischemic stroke in a public hospital in Fortaleza, Brazil, in which, among their covariates, some of them have a representative amount of missing data. The main objective of this study is to apply different techniques of missing data treatment for each variable according to its nature and to adjust a predictive model, and then compare such approaches with a more complete database obtained at another point of this research.

Keywords: Missing data. Multiple Imputation. Predictive models. Logistic Regression. Bayesian methods.

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

Em todas as áreas aplicadas da estatística tem-se problemas com dados faltantes, uma vez que os métodos estatísticos convencionais presumem que todas as variáveis em um determinado modelo são medidas para todos os casos (LITTLE e RUBIN, 2002). O impacto dos dados faltantes na modelagem e em inferências estatísticas é importante, principalmente em casos em que os indivíduos com dados faltantes possuem padrões de respostas que diferem muito daqueles de dados completos. Estimativas coerentes e inferências válidas requerem tratamento adequado dos dados faltantes (*missings*), de modo que, simplesmente descartar os dados perdidos pode levar a resultados tendenciosos tendo em vista que ocorre perda de informações.

As causas dos dados faltantes são diversas. Por exemplo, alguns indivíduos do estudo podem se negar a responder uma pergunta específica em um questionário, medidas das variáveis explicativas podem não estar disponível, problemas no armazenamento dos dados, defeitos em equipamentos, falhas humanas na manipulação dos equipamentos de coleta dos dados, entre outros casos.

Alguns tipos de problemas que estão associados aos valores faltantes são: perda de eficiência; complicações na manipulação e análise dos dados; e até mesmo viés, consequente das discrepâncias entre os valores atribuídos aos dados faltantes e os valores reais desconhecidos (FARHANGFAR, 2007). O tratamento inadequado ou o não tratamento dos dados faltantes também podem afetar os resultados gerais da análise já que algumas informações não estão sendo levadas em consideração na análise final (MCKNIGHT, 2007).

Em paralelo, os modelos preditivos são amplamente utilizados em ciências médicas. Estes modelos equivalem a uma função (ou regra) na qual faz uso de informações históricas ou atuais para prever um determinado evento a respeito de um indivíduo. Ao desenvolver estes modelos, é habitual se deparar com variáveis que possuem dados faltantes e em alguns casos, uma ou mais variáveis com dados omissos podem ser fortemente correlacionadas com o evento futuro que o modelo prediz, podendo assim, interferir consideravelmente no poder preditivo do modelo. Então, a escolha da forma de tratamento dos *missings* será fundamental para um bom ajuste do modelo.

Existem várias abordagens para enfrentar o problema de informações omissas. O método padrão para quase todos os *softwares* estatísticos é simplesmente excluir casos em que se tem qualquer falta de dados sobre as variáveis de interesse, um método conhecido como análise de caso completo ou eliminação de lista. No entanto, a principal desvantagem da análise de

caso completo é que ocasiona uma perda de poder estatístico, ou seja, perda de informação importante para a análise que a torna mais próxima de retratar a realidade, de modo que grande parte dos pesquisadores são relutantes em descartar qualquer tipo de dados do estudo. Outro procedimento semelhante é não incluir no modelo as variáveis que possuem dados faltantes. Neste caso, mesmo não havendo problema de viés da base de construção, pode resultar em um modelo com um poder preditivo inferior ao que seria obtido com todas as variáveis.

Com o intuito de resolver problemas relacionados a dados faltantes, alguns métodos mais elaborados têm sido desenvolvidos. O objetivo desses métodos é preencher os dados faltantes, tornando possível a realização da análise com a base de dados completa, isto é, com todos os indivíduos e variáveis. Nos métodos mais simples é feito a substituição por alguma medida resumo como a média ou mediana dos dados válidos (dados presentes) da variável em questão. Mas, por atribuir o mesmo valor para cada um dos dados em falta, este método interfere diretamente na variância da variável em questão, subestimando a variabilidade da população.

Os processos de imputação única procedem substituindo o dado faltante por valores previstos utilizando as informações das demais variáveis. Para realizar tal estimativa essas técnicas fazem uso de substituição por constantes, regressão linear, algoritmos EM (*expectation-maximization*), regressão multinomial, entre outras.

Uma alternativa mais sofisticada para tratamento de informações faltantes é a imputação múltipla (RUBIN, 1996), criada para levar em consideração o erro gerado pelo processo de estimação por imputação única. Consiste, basicamente, em repetir algum destes processos de imputação várias vezes, produzindo diversos bancos de dados imputados, de modo que a análise estatística desejada é então realizada em cada um destes bancos, produzindo vários resultados. Posteriormente, estes resultados são combinados produzindo um resultado final.

Segundo Little e Rubin (2007) o uso de imputação múltipla detém as seguintes vantagens:

a) Após os dados faltantes serem preenchidos, os métodos padrões de análise de dados completos podem ser usados.
b) Há uma facilidade associada a interpretação dos resultados da análise e permite calcular resumos estatísticos de interesse.
c) Na maioria dos casos, a imputação pode ser gerada apenas uma vez pelo coletor de dados que geralmente tem melhor conhecimento e compreensão sobre o mecanismo que ocasionou o dado omisso em relação ao usuário comum.

d) A imputação adequada produz inferências válidas levando a estimadores com boas propriedades.

Os métodos de simulação produzem estimativas mais precisas, de forma que pode-se obter as distribuições à posteriori dos dados incompletos dado a observação dos dados (por exemplo, para fins de predição). Além disso, a abordagem bayesiana considera a incerteza sobre os valores faltantes e permite estimar as distribuições marginais à posteriori dos parâmetros de interesse condicionais aos dados observados.

1.1 Justificativa

Diante das complicações decorrentes da ausência de dados na base de construção de um modelo preditivo, torna-se necessária alguma técnica para tratá-los, de modo que seja possível aproveitar as informações existentes e não perder o poder de predição além de tentar minimizar os possíveis erros associados à imputações. Portanto, estudos que tratam de técnicas para tratamento de informações omissas são muito importantes em vários campos de conhecimento, como nas ciências médicas em que muitas vezes há presença de *missing data*.

Este estudo irá comparar o poder preditivo de modelos ajustados em uma base de dados real com dados faltantes em que serão utilizadas técnicas de imputação múltipla para o tratamento de dados faltantes com modelo preditivo ajustado na base de dados completa. Vale ressaltar que na aplicação dispomos da base de dados em dois momentos distintos, sendo um com dados faltantes e outro com a base quase completa, os dados foram recuperados por Andrade (2017), assim será possível a comparação do poder preditivo dos modelos ajustados para o desfecho de transformação hemorrágica em pacientes com Acidente Vascular Cerebral (AVC) em um determinado hospital de referência em neurologia de Fortaleza/CE.

1.2 Objetivos

Este trabalho tem como principal objetivo avaliar o quão próximo o poder preditivo do modelo ajustado após a aplicação das técnicas de tratamento de dados faltantes estará do poder preditivo do modelo ajustado com a base de dados mais completa.

Dessa forma, através de um banco de dados real, o qual foi utilizado para construção de um modelo preditivo, serão desenvolvidos alguns modelos utilizando métodos de tratamento para dados faltantes como imputação múltipla, e, posteriormente, seus resultados serão com-

parados por meio de métodos usuais de avaliação de modelos preditivos, como Curva Roc e Coeficiente de Gini.

São objetivos específicos:

i) Utilizar algumas técnicas de imputação de acordo com a natureza de cada variável para obtenção de uma base completa, tais como imputação múltipla.

ii) Comparar o poder preditivo do modelo ajustado obtido com as técnicas de tratamento de dados faltantes aplicadas com o modelo ajustado com base de dados mais completa, obtida em um segundo momento da pesquisa.

1.3 Trabalhos Relacionados

Nunes (2007) fez comparações de ganhos de precisão de seus resultados utilizando técnicas de imputações múltiplas em relação a exclusão de casos com dados faltantes com base de dados reais de epidemiologia.

Assunção (2012) trabalhou com algumas abordagens de tratamento de dados omissos para o desenvolvimento de modelos preditivos de *credit score* fazendo uso de métodos de avaliação de ajustes do modelo para obter o mais adequado e por sua consequência identificar o tratamento de dados faltantes mais indicado para seu objetivo de estudo.

Chen (2013) trabalhou com uma abordagem bayesiana para tratamento de dados faltantes e como utilizar o SAS como ferramenta de análise.

Nenhum dos casos citados comparou o poder preditivo dos seus modelos ajustados com a base em outro momento em que estaria mais completa, isto é, sem informações faltantes.

1.4 Organização do texto

O presente trabalho está dividido em capítulos e organizado da seguinte forma: o capítulo 2 abordará a fundamentação teórica em que será tratado Inferência Bayesiana, descrevendo brevemente sobre teorema de Bayes, distribuição a posteriori, função de máxima verossimilhança e distribuições a priori além de descrever o processo iterativo do método de MCMC (Monte Carlo Markov Chain - Métodos de Monte Carlo via Cadeia de Markov). No capítulo 3, será abordado os diferentes tipos de dados faltantes e como proceder em diferentes tratamentos. O capítulo 4 abordará regressão logística, métodos para seleção de variáveis, e alguns critérios de informação foram explicados para a seleção e escolha do modelo adequado. No capítulo 5 será

apresentado o problema e o desenvolvimento de técnicas computacionais nas análises e avaliação dos dados, mostrando um problema real acerca dos fatores que influenciam no desenvolvimento da transformação hemorrágica em paciente com AVC. O capítulo 6 tratá das considerações finais do trabalho desenvolvido.

2 FUNDAMENTAÇÃO TEÓRICA

Ao analisar um conjunto de dados espera-se que as conclusões obtidas através das estimativas sejam as mais precisas possíveis. Embora não sendo possível eliminar totalmente o viés quando o conjunto de dados possui informações faltantes, deseja-se que esse erro seja reduzido ao máximo. Assim torna-se interessante que haja um conhecimento prévio acerca das informações faltantes e o motivo que levou essa omissão, de modo a identificar algum comportamento padrão de dados faltantes, caso exista, para um direcionamento adequado ao tratamento de dados faltantes.

Os dados faltantes podem ocorrer na variável resposta, nas variáveis explicativas ou em ambas. Este capítulo abordará como se classifica o mecanismo de dados faltantes além do padrão de dados faltantes, assim como também algumas ferramentas estatísticas de análises e alguns métodos de tratamentos encontrados na litereatura para tratamento de dados faltantes.

2.1 Inferência Bayesiana

Quando tem-se interesse em uma característica específica da população, normalmente extrai-se uma amostra aleatória daquela população, ou seja, um subconjunto dos indíviduos que possuem a característica e faz-se inferências acerca do parâmetro, isto é, da característica de interesse da população. Todas as conclusões feitas com base em uma amostra acompanha um grau de incerteza. Na análise inferencial, o desejável é reduzir a incerteza acerca do parâmetro θ desconhecido com base na amostra utilizada.

O princípio básico da inferência bayesiana é que todas as inferências devem ser extraídas da distribuição à posteriori, sendo esta obtida através da combinação da distribuição a priori com a função de verossimilhança.

2.1.1 Teorema de Bayes

O Teorema de Bayes calcula a probabilidade de um evento ocorrer, dado as condições prévias relacionadas a tal evento. Este teorema é utilizado para mensurar o aumento da informação sobre parâmetro desconhecido θ, combinando toda a informação subjetiva disponível com uma quantidade aleatória Y observável. A distribuição da amostra $p(Y|\theta)$ define a relação

entre a variável aleatória e o parâmetro desconhecido. A fórmula de Bayes é definido como:

$$p(\theta|y) = \frac{p(y,\theta)}{p(y)} = \frac{p(y|\theta)p(\theta)}{\int p(y,\theta)d\theta}, \tag{2.1}$$

sendo θ uma variável contínua.

Para θ discreta, a fórmula de Bayes é definida por:

$$p(\theta|y) = \frac{p(y,\theta)}{p(y)} = \frac{p(y|\theta)p(\theta)}{\sum p(y,\theta)}, \tag{2.2}$$

A constante normalizadora é dada por $\frac{1}{p(y)}$, pois não depende do parâmetro desconhecido θ. Para y fixo, a função $p(y|\theta)$ fornece a função de verossimilhança dos possíveis valores de θ, $p(y,\theta)$ a função é a distribuição conjunta e a $p(\theta)$ é chamada de distribuição a priori de θ. A combinação das informações a priori e verossimilhança levam à distribuição a posteriori $p(\theta|y)$ de θ. Portanto, o teorema de Bayes pode ser escrito da seguinte forma:

$$p(\theta|y) \propto p(y|\theta)p(\theta), \tag{2.3}$$

em que $\propto$ denota proporcionalidade. Quando usa $\propto$ para encontrar o núcleo da distribuição a posteriori não é considerado a constante normalizadora, portanto, para recuperar essa constante basta reescrever a distribuição a posteriori da seguinte forma:

$$p(\theta|y) = kp(y|\theta)p(\theta), \tag{2.4}$$

em que k representa $\frac{1}{p(y)}$ que é a constante normalizadora, determinada de modo que:
$k = \int_\Theta p(y|\theta)p(\theta)d\theta = E_\theta[p(y|\theta)]$, caso contínuo;
e
$k = \sum_\Theta p(y|\theta)p(\theta) = E_\theta[p(y|\theta)]$, caso discreto.

2.1.2 Função de Verossimilhança

De acordo com Monfardini (2016), a função de verossimilhança pode ser interpretada como a função do vetor de parâmetros para um conjunto de dados fixo, em que mede o quanto os dados suportam uma hipótese sobre o parâmetro θ.

Definição 1. Seja $Y_1, Y_2, ..., Y_n$ uma amostra aleatória de uma família de distribuição $p(y|\theta)$, $\theta \in \Theta$. A distribuição de probabilidade conjunta é dada por:

$$L(y|\theta) = p(y_1, y_2, ..., y_n|\theta) \tag{2.5}$$

Em que $y = (y_1, y_2, ..., y_n)$ é o vetor de observações de $Y = Y_1, Y_2, ..., Y_N$. Quando os dados são independentes e identicamente distribuídos temos que:

$$L(y|\theta) = \prod_{i=1}^{n} p(y_i|\theta) \tag{2.6}$$

2.1.3 Distribuições à Priori

A distribuição a priori representa o conhecimento prévio, obtido antes do experimento, sobre o parâmetro θ desconhecido. Antes de observar os dados, é necessário que sejam feitos julgamentos sobre os parâmetros de interesse, essas informações acerca dos parâmetros podem ser obtidas por estudos anteriores, opiniões de especialistas, entre outras fontes. Em casos que não se tenha nenhuma informação dos parâmetros de interesse, pode-se optar por prosseguir as análises usando distribuição a priori não informativa.

2.1.3.1 Distribuições à Priori Conjugadas

Segundo Ehlers (2011) a distribuição à priori é representada por uma forma funcional, quando é possível definir um família paramétrica de densidades a partir do conhecimento acerca de θ, e os parâmetros são especificados seguindo o conhecimento. Nesse caso, os parâmetros indexadores da família de distribuição à priori são conhecidos e não são considerados variáveis aleatórias. Os parâmetros conhecidos são chamados de hiperparâmetros, para diferenciar dos parâmetros de interesse desconhecidos. O intuito é que a distribuição à priori pertença a mesma família de distribuições à posteriori, para que o conhecimento que se tem sobre o parâmetro θ envolva somente a modificação nos hiperparâmetros.

Definição 2. Se $F = \{p(x|\theta), \theta \in \Theta\}$ é uma classe de distribuições amostrais então uma classe de distribuições P é conjugada a F se $\forall p(x|\theta) \in F$ e $p(\theta) \in P \Rightarrow p(\theta|x) \in P$.

2.1.3.2 Distribuições à Priori não Informativas

As distribuições à priori não informativas são utilizadas quando não se tem informação disponível a respeito do vetor de parâmetros θ.

Definição 3. Seja θ uma parâmetro definido no intervalo (a,b) uma priori não informativa

uniforme seria:

$$P(\theta) = \begin{cases} \frac{1}{b-a}, & \text{se } a < \theta < b, a < b. \\ 0, & \text{c.c.} \end{cases} \tag{2.7}$$

No entanto, a distribuição *à priori* uniforme tem uma limitação. Seja $\Phi = \theta^2$, pela regra do Jacobiano mostra que a distribuição a *priori* uniforme deixa de ser não informativa quando se deseja estimar alguma função do parâmetro, de modo que é efetivamente não informativa apenas para o parâmetro de interesse (BOX e TIAO, 1992). Para contornar esse problema Jeffrey propôs, baseada na informação de Fisher, uma distribuição a priori imprópria.

Definição 4. Considere uma observação X com função de probabilidade $p(x|\theta)$. A informação de Fisher sobre o parâmetro $\theta|X$ é dada por:

$$IF(\theta) = E\left[\left(\frac{\partial \ln f(x|\theta)}{\partial \theta}\right)^2\right] \tag{2.8}$$

Em condições de regularidade, a informação de Fisher pode ser escrita por:

$$IF(\theta) = -E\left[\frac{\partial^2 \ln f(x|\theta)}{\partial \theta^2}\right] \tag{2.9}$$

Definição 5. Seja uma observação X com função de probabilidade $p(x|\theta)$. A priori não informativa de Jeffrey's tem função de densidade da forma:

$$P(\theta) \propto |IF(\theta)|^{\frac{1}{2}} \tag{2.10}$$

2.2 Método Monte Carlo via Cadeias de Markov

Para se obter a distribuição a posteriori é comum a utilização dos métodos de simulação. No entanto, nem sempre é simples obter uma densidade que seja uma aproximação adequada da distribuição a posteriori e ao mesmo tempo encontrar um algoritmo de rápida convergência.

Com os avanços computacionais foi possível a implementação dos Métodos de Monte Carlo via Cadeias de Markov (Monte Carlo Markov Chain - MCMC) que são amplamente utilizados em inferência Bayesiana quando a distribuição a posteriori não é facilmente obtida

de forma analítica, de forma que possibilita simular amostras grandes de uma determinada distribuição a posteriori e calcular suas estimativas amostrais.

A finalidade da utilização do método MCMC é aproximar a distribuição a posteriori e obter suas estimativas amostrais, tal procedimento se baseia em métodos de simulação iterativa através das cadeias de Markov.

Uma cadeia de Markov é um processo estocástico $X_0, X_1, ..., X_n$ em que a distribuição de X_t depende somente do estado anterior X_{t-1}, podendo ser assim expressada:

$$P(X_t \in A | X_0 = x_0, X_1 = x_1, ..., X_{t-1} = x_{t-1}) = P(X_t \in A | X_{t-1} = x_{t-1}).$$

O MCMC simula um passeio aleatório até alcançar uma distribuição estacionária, ou seja, quando não se tem mudança de probabilidade em relação ao tempo, assim, os métodos MCMC devem ter uma cadeia irredutível e aperiódica, de modo que cada estado pode ser alcançado a partir de qualquer outro estado.

A partir de um número de iterações t a cadeia estaciona, convergindo assim para a distribuição a posteriori. O método mais usual para geração de números aleatórios, baseados em cadeia de Markov é o Amostrador de Gibbs (EHLERS, 2011).

3 DADOS FALTANTES

Em um conjunto de dados, um registro é completo se todos os seus atributos (variáveis) estão apropriadamente preenchidos com seus dados. Um dado faltante indica que um atributo de um registro não está preenchido.

Antes do pesquisador proceder qualquer tipo de análise é fundamental conhecer se a presença de dados faltantes em uma variável está vinculada a algum processo identificável (LITTLE e RUBIN, 2002), isto é, saber o mecanismo que levou ao conjunto de dados a ter valores faltantes, uma vez que esse conhecimento servirá de auxílio para escolher a técnica apropriada para realizar a correta análise dos resultados.

O principal sistema de classificação foi criado por Donald Rubin em 1976. Os padrões de dados omissos são classificados como, MCAR, MAR e MNAR.

Com o intuito de representar matematicamente estes mecanismos de dados faltantes, parte-se do pressuposto que se tem uma matriz de dados coletada Z, com i linhas, que correspondem as observações, e j colunas que correspondem as variáveis. Pode-se dividir Z em dois conjuntos:

$$Z = \{Z_{obs}, Z_{mis}\}, \tag{3.1}$$

em que Z_{obs} é o conjunto com os indivíduos que contém todas as variáveis observadas e Z_{mis} é o conjunto com os indivíduos que contém variáveis não observadas. Desta forma tem-se que $z_i = (z_{i1}, z_{i2}, ..., z_{in})$ em que z_{ij} refere-se ao valor da observação i na variável j.

Correspondente a cada matriz de dados Z existe um identificador de dado faltante associado, denotado por R, o qual deve ter as mesmas dimensões de Z, onde $r_{ij} = 1$, se z_{ij} é observado, e $r_{ij} = 0$, caso contrário. Desta forma, o mecanismo de dados faltantes é caracterizado pela distribuição condicional de R dado Z, isto é $P(R|Z)$, a qual pode ser de três tipos (LITTLE e RUBIN, 2002).

a) Dado faltante completamente aletório - MCAR (*Missing Completely at Random*)
b) Dado faltante aleatório - MAR (Missing at Random)
c) Dado faltante não aleatório - NMAR (*Missing Not at Random*).

3.1 Dado Faltante Completamente Aletório (Missing Completely at Random)

Dados faltantes são completamente aleatórios quando a probabilidade de um registro que tem um valor em falta para um atributo não depende do valor observado dos dados tampouco

do valor faltante, ou seja, quando as razões para as perdas não são relacionadas a quaisquer respostas dos sujeitos, incluindo o valor faltante.

De modo que nenhuma das variáveis, dependente (Y) ou explicativa (X), tem valores faltando relacionados com os valores da própria variável (ALLISON, 2001). Em casos em que os dados faltantes não dependem dos valores de Z, faltantes ou observados, tem-se:

$$P(R|Z) = P(R). \tag{3.2}$$

Isso significa que a causa que levou aos dados faltantes é um evento aleatório, ou seja, o mecanismo é *Missing Completely at Random* (MCAR). Nesse caso, os valores faltantes para uma variável é uma simples amostra aleatória dos dados dessa variável, isto é, a distribuição dos valores faltantes é de mesma natureza da dos valores observados (ZHANG, 2003).

Na maioria das situações, o mecanismo é MCAR mesmo se os dados faltantes existam devido a algum evento que não é verdadeiramente aleatório, mas ocasionado por alguma variável. Geralmente acontece quando a causa é uma variável não correlacionada com a variável que possui valores faltantes (GRAHAM, 1995).

A grande vantagem de o mecanismo ser MCAR é que a causa que levou aos dados faltantes não precisa fazer parte da análise para controlar a influência destes nos resultados da pesquisa (GRAHAM, 1995).

Por exemplo, isso pode ocorrer se o processo para medir uma variável de um estudo é muito caro. Então, decide-se que ela será medida apenas para um subconjunto aleatório da amostra, o que implica nos dados serem MCAR para as demais variáveis dessa amostra (ALLISON, 2001).

3.2 Dado Faltante Aleatório (Missing at Random)

Dados faltantes são classificados como MAR quando o mecanismo de perda em uma variável é previsível a partir de outras variáveis no banco de dados e não é diretamente devido à variável específica na qual os dados são perdidos, de modo que os dados faltantes não dependem dos valores de Z_{mis} e apenas dos valores de Z_{obs}, tem-se:

$$P(R|Z) = P(R|Z_{obs}). \tag{3.3}$$

Neste caso, os dados faltantes são causados por alguma variável observada, disponível para análise no banco de dados e correlacionada com a variável que possui dados faltantes

(GRAHAM, 1995). Assim, os dados faltantes de uma variável são como uma amostra aleatória simples dos dados para essa variável dentro de subgrupos definidos por valores observados, e a distribuição dos valores faltantes é a mesma que a distribuição dos valores observados dentro de cada subgrupo (ZHANG, 2003).

Este é um mecanismo acessível, pois se a causa que levou aos dados faltantes pode ser medida e é incluída devidamente na análise, todas as influências causadas por eles podem ser consideradas (GRAHAM, 1995). Por exemplo, considere uma pesquisa na qual as mulheres são menos propensas a fornecer sua renda pessoal. Se conhecermos o sexo de todos os sujeitos e tivermos a renda para algumas mulheres, estimativas não viesadas da renda podem ser feitas. Isto porque a renda que se tem de algumas mulheres é uma amostra aleatória das rendas de todas as mulheres.

3.3 Dado Faltante não Aleatório (Missing Not at Random)

O mecanismo gerador de dados faltantes MNAR ocorre quando a probabilidade de um registo com um valor faltante em uma variável pode depender do valor da variável. Ou seja, quando a distribuição de R depende dos dados faltantes contidos na matriz Z (Z_{mis}), talvez pode depender dos dados observados (Z_{obs}), tem-se (ZHANG, 2003):

$$P(R|Z) \neq P(R|Z_{obs}). \tag{3.4}$$

Esta situação geralmente ocorre quando a causa dos dados faltantes em uma variável é seu próprio valor. Por exemplo, quando sujeitos com níveis de renda muito baixos ou muito altos têm probabilidade menor de responder sobre sua renda pessoal numa entrevista.

3.4 Métodos de Tratamento de Dados Faltantes

Nesta seção fazemos um resumo dos diferentes procedimentos para tratamento de dados faltantes. Dessa forma pretende-se dar ao leitor uma visão geral sobre à área.

3.4.1 Exclusão dos dados faltantes

Os métodos a seguir consistem em excluir valores faltantes do conjunto de dados. São procedimentos mais simples para tratar informações omissas, no entanto não são os mais

eficientes para alguns mecanismos de dados faltantes, principalmente quando não se dispõe de amostras grandes.

3.4.1.1 Caso Completo ou Listwise Deletion (LD)

O método *Listwise Deletion* ou Caso Completo, exclui todos os registros com algum atributo faltante, ou seja, somente são levados em consideração os casos completos (em que não possui registros em falta). Mesmo se houver apenas um atributo faltante, o objeto será descartado.

Em termos da matriz Z definida na seção 2.11, considera-se apenas Z_{obs}, isto é, o conjunto de indivíduos que contém todas as variáveis observadas em nível de registro.

Levando em consideração somente os casos completos, forma-se uma base de dados completa, na qual os procedimentos convencionais podem ser aplicados para análise (MCKNIGHT, 2007). Assim, o objetivo deste método não é propor uma metodologia de estimação para os valores faltantes, mas sim obter um banco de dados que possa ser analisado convencionalmente a fim de se obter estimativas dos parâmetros de interesse.

Este método tem a vantagem de tornar a análise simples, uma vez que as análises estatísticas convencionais podem ser aplicadas, sem nenhuma alteração e não exige métodos computacionais especiais (LITTLE e RUBIN, 2002)(ALLISON, 2001). No entanto, produz uma potencial perda de informação devido ao descarte dos casos incompletos, podendo causar perda de precisão e viés.

3.4.1.2 Casos Disponíveis ou Pairwise Deletion (PD)

O método de *Pairwise Deletion* ou casos disponíveis é semelhante ao LD, sendo que o PD não descarta os dados em nível de registro, mas sim em nível de atributo(s) de interesse. Ele inclui todos os objetos em que a variável de interesse está presente (LITTLE e RUBIN, 2002).

A principal desvantagem deste método é que a amostra-base é modificada de variável para variável, dependendo do padrão de dados faltantes, uma vez que é descartado os objetos em nível de variável e não em nível de observação.

De acordo com Allison (2001), se os dados são MCAR, PD possibilita estimativas não-viesadas dos parâmetros de interesse. Caso contrário, as estimativas podem ser viesadas. Little e Rubin (2002) destacam que, para o caso do padrão de dados omissos ser MCAR, amostras distintas são aceitáveis para estimativas de média e variância, mas não para estimativas

de covariância e correlação. Portanto, nenhum desses dois métodos geram resultados satisfatórios (LITTLE e RUBIN, 2002).

3.4.2 Imputação Única

Os métodos de imputação são utilizados para preencher os dados faltantes, podendo ser do tipo imputação simples e imputação múltipla. Métodos de imputação única são métodos utilizados para substituir dados faltantes em uma base de dados (ENGELS e DIEHR, 2003), para que depois seja possível analisar o banco de dados como se não tivesse havido observações faltantes. As estimativas dos parâmetros são obtidas por métodos convencionais já que o banco de dados já está completo, alguns métodos de imputação são apresentadas a seguir.

3.4.2.1 Imputação Dedutiva

Esse método depende de informações complementares existentes nos dados, de modo que o valor perdido em uma variável possa ser recuperado com base nas informações das variáveis informadas. Por exemplo, se a informação faltante for sexo e o indivíduo na mesma unidade amostral tiver respondido que fez exame de próstata fica evidente que a variável sexo a ser imputada é masculino.

3.4.2.2 Imputação pela Média Geral

Este método atribui o dado em falta pela média geral da mesma variável, de modo que não utiliza informações disponíveis de outras variáveis (DURRANT, 2005).

$$\tilde{x}_{ij} = \frac{\sum_{l=1}^{n_{obs}} x_{lj}}{n_{obs}} = \beta_j, i = 1, ..., n_{mis}, \tag{3.5}$$

em que n_{obs} é a quantidade de respondentes na variável X_j e n_{mis} é a quantidade de informações omissas.

Esse método geralmente é usado para variáveis contínuas e pode mudar a variabilidade dos dados reais sendo que todos os valores omissos serão substituídos pela média geral. Um exemplo seria a variável tempo de terapia de um indivíduo em avaliações de óbito em pacientes com câncer.

3.4.2.3 Imputação pela Média Dentro de Classes

Este método divide os dados em classes e imputa as informações omissas pela média das classes das suas unidades respondentes.

$$\tilde{x}_{ij/h} = \frac{\sum_{l=1}^{n_{obs/h}} x_{lj/h}}{n_{obs/h}} = \beta_{j/h}, i = 1, ..., n_{mis/h}, \tag{3.6}$$

em que h é a classe de imputação definida. Este método não pode ser utilizado em variáveis qualitativas e segundo Albieri (1989) exerce menos efeito sobre a distribuição da variável a imputar do que a imputação pela média geral.

3.4.2.4 Imputação pela Mediana

Este método usa a mediana, outra medida de tendência central, para preencher os dados omissos. Quando a distribuição dos dados desvia muito da distribuição normal padrão a mediana proporciona uma melhor síntese da distribuição, e, por conseguinte uma estimativa mais adequada para valores faltantes (MCKNIGHT, 2007).

3.4.2.5 Imputação Geral Aleatória

Este método define aleatoriamente para cada dado em falta uma informação disponível da mesma variável na base de dados. Por exemplo, supondo que um indivíduo não informou seu nível de escolaridade, assim é escolhido de modo aleatório dentro da variável escolaridade a informação de um outro indivíduo e feito o preenchimento do dado faltante.

3.4.2.6 Imputação Aleatória dentro de Classes

Este método também define aleatoriamente uma informação disponível na base de dados da mesma, mas dentro de classes semelhantes que são previamente definidas.

3.4.2.7 Imputação Hot-deck

Este método tem como base a especificação de um registro da matriz de dados em que os valores respondentes são similares em relação a uma determinada variável auxiliar X_j, $j = 1..., k$, de forma que são selecionados para a imputação. São os denominados doadores.

Em resumo, identifica o indivíduo com dado observado mais próximo com o indivíduo que possui informação omissa em relação às variáveis auxiliares e substitui-se o dado em

falta pelo valor do respondente pareado.

Dentre as principais vantagens do método de imputação por *hot-deck* é que a distribuição de valores imputados terá a mesma forma da distribuição dos dados observados (RUBIN, 1996). Além do mais, segundo Durrant (2005) outra vantagem é imputar valores que são observados na pesquisa e que é indicado para tratar não respostas em variáveis categorizadas.

A título de exemplo, em uma pesquisa sobre aprovação da atual administração do país onde o critério de análise é fechado pelo preenchimento de um instrumento validado, sendo consideradas cinco categorias de respostas, supondo que um item não preenchido será imputado por *hot-deck*. Primeiro é necessário criar uma matriz de padrões de respostas considerando como variáveis auxiliares, por exemplo, faixa etária e partido político. Identifica-se qual indivíduo respondente tem o mesmo padrão do não respondente em relação a faixa etária e partido político. Aquele de mesmo padrão é o doador e o dado em falta no item será preenchido com a resposta do doador. Quando tem-se mais de um indivíduo com o mesmo padrão do não respondente usa-se conhecimento do pesquisador para escolher o doador. Este método pode gerar estimativas viesadas se uma mesma unidade for usada como doadora com mais frequência que outras (DURRANT, 2005).

3.4.2.8 Imputação por Regressão (Média Preditiva)

Este método faz uso das informações disponíveis de variáveis auxiliares que entrará no modelo como covariáveis para ajustar a regressão de modo que o valor a ser imputado será a variável dependente, isto é, os valores imputados são preditos por meio de regressão simples ou múltipla, que pode ser utilizado uma ou mais variáveis auxiliares existentes para predizer os valores faltantes de outra variável correlacionada com as anteriores. Assim temos:

$$\tilde{x}_{ij} = \beta_0 + \sum_{l \neq j} \beta_1 x_{il}, \tag{3.7}$$

em que $i = 1, ..., n_{mis}$ e $j = 1, ..., k$.

Para o ajuste por regressão linear normal, os e_{ij} são considerados iguais a 0, ou seja, não são considerados efeitos aleatórios na estimação do valor imputado. Se a variável X_j a ser imputada for qualitativa utiliza-se regressão logística ou log-linear.

3.4.2.9 Imputação por Regressão Aleatória

Este método difere do anterior por considerar o erro aleatório da regressão na imputação. Dessa forma, tem-se:

$$\tilde{x}_{ij} = \beta_0 + \sum_{l \neq j} \beta_1 x_{il} + e_{ij}, \tag{3.8}$$

em que $i = 1, ..., n_{mis}$ e $j = 1, ..., k$.

O método de imputação por regressão aleatória impede de indivíduos que têm os mesmos valores nas mesmas covariáveis fiquem com o mesmo valor imputado (valor predito igual), uma vez que considerando o erro aleatório é adicionado ao valor predito um valor escolhido ao acaso de uma distribuição $N(0, \sigma^2)$, onde σ^2 é a variância residual da regressão.

3.4.2.10 Método Maximum Likelihood (ML)

Os métodos de Máxima Verossimilhança ou ML (*Maximum Likelihood*) em cálculos de estimativas dos parâmetros para dados faltantes são obtidos a partir dos dados observados, das relações existentes entre os registros observados e das restrições impostas pela suposição do modelo de distribuição (MCKNIGHT, 2007). Seu principal objetivo é estimar os parâmetros de interesse e não simplesmente atribuir valores aos dados faltantes diferente dos métodos de imputação.

De acordo com Allison (2001) o método ML tem como princípio básico escolher como estimativa dos parâmetros aqueles valores que maximizariam a probabilidade de obter o que, realmente, foi observado.

O procedimento ML consiste em considerar que os dados são gerados por um modelo descrito pela função de densidade $f(Y/\theta)$, em que Y são os dados e θ é um conjunto de parâmetros desconhecidos que rege a distribuição de Y, do qual sabe-se apenas pertencer a Ω_θ (LITTLE e RUBIN, 2002). Logo, dado o modelo considerado e uma vez calculado o vetor de parâmetros θ, $f(Y/\theta)$ pode ser usado para amostrar valores faltantes (LITTLE e RUBIN, 2002)(ALLISON, 2001).

A função de verossimilhança é dada como:

$$L(\theta|Y) = \prod f(Y_i|\theta) \tag{3.9}$$

Isto é, $L(\theta|Y)$ é uma função do vetor de parâmetros $\theta \in \Omega_\theta$ dado Y, proporcional à função de densidade (LITTLE e RUBIN, 2002). Sendo que em alguns casos é mais fácil trabalhar com a função $l(\theta|Y)$ (log-verossimilhança), que é o logaritmo natural (ln) da função de verossimilhança e que tem pontos de máximo nos mesmos pontos que a função original.

O método por máxima verossimilhança produz estimativas aproximadamente não viesadas para grandes amostras. E, quando se trata de amostragens repetidas, as estimativas têm aproximadamente uma distribuição normal, o que pode ser empregado para obter intervalos de confiança (MCKNIGHT, 2007) (ALLISON, 2001).

3.4.2.10.1 Algoritmo EM

O algoritmo EM (*Expectation maximization*) é bastante utilizado para obter estimativas ML em bases de dados incompletas (LITTLE e RUBIN, 2002). O intuito base é substituir uma difícil maximização da verossimilhança por uma sequência de maximizações mais simples, de modo que é projetado para encontrar estimadores de máxima verossimilhança (CASELLA e BERGER, 2010). Trata-se de um processo iterativo onde se repete dois passos, Estimação e Maximização, até que obtenha-se convergência.

Definição 6: Considere um conjunto de dados com informaçãoes observados e informações faltantes, com função de densidade dada por $p(\mathbf{y}^c|\theta)$.
De modo que, $l(\theta, \mathbf{y}^c)$ representam, respectivamente, a função log-verossimilhança dos dados completos e observados. O algoritmo sugere que em inicialmente encontremos o valor esperado do logaritmo da verossimilhança (passo E) e em seguida encontremos o seu máximo (passo M), isto é:
Passo E: Calcular $Q(\theta|\theta^{(k)}) = E(l^c(\theta, \mathbf{y}^c)|\mathbf{y}^c, \theta^{(k)})$;
Passo M: Encontrar $\theta^{(k+1)}$ que maximiza $Q(\theta|\theta^{(k)})$.
O processo é repetido até atingir convergência, podendo ser adotado um critério de parada, como por exemplo $||\theta^{(k+1)}\theta^{(k)}|| < \varepsilon$.

Em resumo, estimação: imputar valores para os dados faltantes usando como base os valores dos parâmetros (ALLISON, 2001) e maximização: estimar novos valores dos parâmetros (ALLISON, 2001). O método atinge a convergência quando a diferença entre os valores estimados dos parâmetros em duas iterações consecutivas é menor que o valor pré-estabelecido.

De acordo com Mcknight (2007) e Graham (1995) o procedimento EM tende a subestimar os erros padrões da amostra, que são críticos para os testes de hipóteses, podendo

gerar erros do Tipo I. Além disso, o algoritmo EM também não garante a convergência para o ótimo global quando a função de verossimilhança (ou log-verossimilhança) for multimodal, em que apresentam ótimos locais. De modo que é imprescindível uma boa escolha dos valores iniciais dos parâmetros para alcançar o ótimo global.

Uma das principais desvantagens do algoritmo EM é não permitir a obtenção de imputação para variáveis categóricas.

Os métodos citados acima nos permitem ter uma visão panorâmica da área de dados faltantes, porém neste estudo aplicaremos técnicas de algoritmo EM e Imputação Múltipla.

3.4.3 Imputação Múltipla

Nesta seção trataremos sobre uma das ferramentas utilizadas no nosso estudo. A imputação múltipla (IM) foi proposta por D.B. Rubin, na década de 70, na tentativa de resolver o problema de não-resposta em pesquisas. Esta técnica está sendo cada vez mais utilizadas devido aos avanços computacionais.

Essa técnica possibilita a inclusão da incerteza da imputação dos parâmetros estimados pontualmente na variância dos resultados estimados corrigindo o principal problema associado à imputação única (RUBIN, 1996). Assim, para cada dado faltante são imputados m valores, ao invés de um, formando assim m bases de dados. Ou seja, são obtidas m matriz de dados completos e em cada conjunto de dados usa-se procedimento de análise para dados completos. Posteriormente, tem-se a estimativa pontual do parâmetro que é encontrada por meio da média das múltiplas imputações e o erro padrão associado obtido através de sua variância.

O modelo utilizado para fazer as imputações será no melhor dos casos uma aproximação da realidade. Segundo Rubin (1996), para esse procedimento é necessário mais trabalho para produzir imputação múltipla em comparação à imputação simples, além de mais espaço para depositar um conjunto de dados múltiplo e mais trabalho empregado para análise do conjunto de dados múltiplo imputado do que um conjunto de dados simples imputado. Podendo também aparecer discrepância na variância quando é admitido pressupostos inadequados, como supor normalidade erroneamente, de modo que o modelo é inconsistente para imputar os dados.

No entanto, quando imputações são realizadas aleatoriamente com intuito de representar a verdadeira distribuição dos dados, a IM aumenta a eficiência da estimação. Outra vantagem da IM é que ao fazer m imputações sob um mesmo modelo para dados em falta, inferências válidas são obtidas combinando inferências de dados completos de forma simples. Além disso, é possível um estudo da sensibilidade das inferências para vários modelos de dados faltantes (RUBIN, 1996).

São necessários três passos para o método de Imputação Múltipla:

1) São gerados m conjunto de dados completados por meio de técnicas adequadas de imputação.
2) Utilizando procedimentos padrões são feitas m análises de dados completos.
3) Os resultados das m análises dos dados completos obtidas no passo dois são combinados para obter as inferências necessárias.

3.4.3.1 Combinação dos resultados

Após a imputação dos dados, são obtidas m estimativas para o parâmetro de interesse D_i, $i = 1, 2, ..., m$. Uma maneira de obter a estimativa global para um parâmetro de interesse D é através da média das estimativas produzidas para as m bases de dados (MCKNIGHT et al., 2007). Cada parâmetro estimado é chamado de $\hat{D}$ e a estimativa global é chamada de $\bar{D}$, dada por:

$$\bar{D} = \frac{1}{m}\sum_{i=1}^{m}\hat{D}_i. \tag{3.10}$$

Para calcular o erro padrão global, necessário para os testes de significância e para os intervalos de confiança (MCKNIGHT et al., 2007), calcula-se, inicialmente, a média dos erros padrão (que foram calculados através dos m conjuntos de dados completos) ao quadrado, chamado de *within-imputation variance*:

$$\bar{W} = \frac{1}{m}\sum_{i=1}^{m}\hat{W}_i, \tag{3.11}$$

onde W_i é o erro padrão ao quadrado calculado através do conjunto de dados completo i. Posteriomente, calcula-se a variância do parâmetro de interesse estimado, o que é chamado de *between-imputation variance*:

$$\bar{B} = \frac{1}{1-m}\sum_{i=1}^{m}(\hat{D}_i - \bar{D})^2. \tag{3.12}$$

E a variância combinada é obtida através da seguinte fórmula:

$$T = \bar{W} + \left(1 + \frac{1}{m}\right)B. \tag{3.13}$$

O erro padrão global pode finalmente ser calculado como $\sqrt{T}$. Este é necessário para se calcular intervalos de confiança e níveis de significância de D, os quais são construídos utilizando uma distribuição de referência t com:

$$df = (m-1)(1+r^{-1})^2, \tag{3.14}$$

graus de liberdade, onde r representa o aumento relativo na variância devido aos dados omissos e é dado por:

$$r = \frac{(1+m^{-1})B}{W}. \tag{3.15}$$

Sendo que quanto maior o valor de *r*, menor a estabilidade nos parâmetros estimados, refletindo em maior incerteza estatística (MCKNIGHT et al., 2007).

Assim, de forma padrão, um intervalo com $100(1-\alpha)\%$ de confiança de D é:

$$IC_{\hat{D}t_{df}} = \hat{D} \pm (\alpha/2)\sqrt{T}, \tag{3.16}$$

onde $t_{df}(\alpha/2)$ é um quantil da distribuição t de Student com df graus de liberdade. De acordo com Schafer e Graham (2002), quando df é grande, a distribuição de t é normal, a variância total é bem estimada e pouco se ganha em aumentar o valor de m (SCHAFER GRAHAM, 2002).

A combinação dos resultados imputados será feita através da função PROCMI ANALYSE da ferramenta SAS que será utilizada como auxílio para a análise de dados.

3.4.3.2 Informação faltante

Veroneze (2011) afirma que para ter uma percepção do impacto dos dados faltantes nas estimativas dos parâmetros e nas conclusões das estatísticas geradas, é interessante calcular um valor entre 0 e 1, o qual é chamado por taxa de informação faltante (λ).

$$\lambda = \left(r + \frac{2}{gl+3}\right)\frac{1}{r+1}, \tag{3.17}$$

em que r é o aumento relativo na variância em consequência dos dados faltantes definido na seção anterior, e gl são os graus de liberdade. O r representa a estabilidade dos parâmetros, então quanto maior o seu valor menor é a certeza que se tem sobre os resultados.

3.4.3.3 Eficiência Relativa

Não é necessário para se ter estimativas eficientes grandes quantidades ($m \to \infty$) de imputações, pois é um processo demorado e exige custos elevados e muitos recursos computacionais. Schafer (1997) afirma que a quantidade de imputações necessárias para que uma estimativa de conjunto de dados tenha eficiência é dado por:

$$RE = \sqrt{1 + \frac{\lambda}{m}}, \tag{3.18}$$

sendo λ a taxa de informação faltante e m a quantidade de conjuntos de dados completados.

Definir a quantidade de conjunto de dados imputados é a parte mais importante da IM, uma vez que as técnicas de imputação aplicadas têm o papel de preservar a relação entre as observações faltantes e observadas.

3.4.3.4 Método da Regressão Linear Bayesiana (BLR – Bayesian Linear Regresssion)

Regressão linear é bastante utilizada para predizer Y_i de um conjunto de covariáveis X_i . Então tem-se

$$Y_i \sim N(X_i\beta, \sigma^2). \tag{3.19}$$

Assim, a especificação para $f(Y_i/X_i, \theta)$, $\theta = (\beta, \log\sigma)$, β um vetor de q componentes, onde q é o número de preditores, e σ um escalar. Admitindo-se uma distribuição a priori não-informativa para θ, $P(\theta) \propto 1$, para evitar grandes complexidades assume-se $n_1 > q$, onde n_1 é o número de respondentes. Assim de acordo com Rubin (1987) a distribuição *a posteriori* de θ envolve apenas os Y_i observados. De modo que,

$$\widehat{\sigma}_1^2 = \frac{\sum_{obs}(Y_i - X_i\widehat{\beta}_1)^2}{n_1 - q}, \tag{3.20}$$

sendo:

$$\widehat{\beta}_1 = V\left[\sum_{obs} X_i^T Y_i\right], \tag{3.21}$$

e

$$V = \left[\sum_{obs} X_i^T X_i\right]. \tag{3.22}$$

Logo a distribuição à *posteriori* de θ descrita em termos de distribuições padrão pode-se estimar os parâmetros a serem usados na imputação.

Por fim, a tarefa de imputação para esse modelo pode ser descrita pelos três passos a seguir:

1. Simular uma variável aleatória σ_*^2 qui-quadrado, $X^2_{n_1-q}$, g e seja:

$$\sigma_*^2 = \frac{\widehat{\sigma}_1^2(n_1-q)}{g}. \tag{3.23}$$

2. Simular q variáveis independentes $N(0,1)$ para criar um vetor Z de q componentes e seja:

$$\beta^* = \widehat{\beta}_1 + \sigma_*[V]^{1/2}Z, \tag{3.24}$$

em que $[V]^{1/2}$ é a raiz quadrada de V tal como a raiz quadrada triangular obtida pela fatoração de Cholesky.

3. Simular os n_0 valores dos Y_{mis} como:

$$Y_i^* = X_i\beta^* + z_i\sigma^*, \tag{3.25}$$

onde os n_0 desvios normais z_i são simulados independentemente. Para um novo valor a ser imputado para Y_{mis} simula-se um novo valor para o parâmetro σ_*^2. Assim, se m imputações são desejadas, esses três passos são repetidos m vezes independentemente.

3.4.3.5 Método da Média Preditiva (PMM – Predictive Mean Matching)

O método PMM é parecido ao método BLR, porém no terceiro passo é alterado da forma:

a) Gera-se os n_0 valores preditos dos Y_{mis} como $Y_i^* = X_i\beta^*$, $i \in$ dados em falta.
b) Para cada Y_i^*, $i \in$ faltantes, encontra-se o respondente cujo Y_i ($i \in$ observados) de modo que esteja o mais próximo de Y_i^*. O valor do Y_i será usado para próxima imputação.

Esse método calcula a variabilidade entre imputações desde que os passos 1 e 2 para simular β^* do método BLR e um modelo linear para nortear a escolha dos valores a serem imputados sejam utilizados.

3.4.3.6 MCMC (Markov Chain Monte Carlo)

O método de Monte Carlo é baseado em Cadeias de Markov (MCMC) tendo como objetivo simular distribuições multivariadas as quais o limite é uma cadeia de Markov estacionária

que possui a distribuição que se interessa encontrar (GILKS, 1996), conforme já explicado na seção 2.2.

Quando a função de verossimilhança conjunta dos dados observados não pode ser fatorada em funções de verossimilhança independentes não é possível aplicar este método.

3.4.3.7 Métodos FCS para Conjuntos de Dados com Padrões Arbitrários Ausentes

Quando se tem um conjunto de dados com um padrão de dados arbitrário ausente, pode-se usar métodos FCS para imputar valores faltantes para todas as variáveis, assumindo a existência de uma distribuição conjunta para essas variáveis (BRAND, 1999; VAN BUUREN, 2007).

3.4.3.7.1 Algoritmo MICE

Multivariate Imputation by Chained Equation (MICE) é uma técnica de imputação múltipla (RAGHUNATHAN ET AL., 2001 ; VAN BUUREN, 2007) e funciona sob a suposição de que, dadas as variáveis usadas no procedimento de imputação, os dados faltantes são *Missing At Random* (MAR), caso contrário pode resultar em estimativas tendenciosas.

Muitos dos procedimentos de imputação múltipla inicialmente desenvolvidos assumiram um grande modelo conjunto para todas as variáveis, como uma distribuição normal conjunta. A imputação multivariada por equações encadeadas (MICE) é uma abordagem que permite maior flexibilidade e é uma alternativa a esses modelos conjuntos.

No procedimento MICE uma série de modelos de regressão é ajustada, sendo cada variável com dados omissos modelada condicionalmente às outras variáveis da base de dados. De modo que cada variável é modelada de acordo com sua distribuição, por exemplo, variáveis binárias modeladas usando regressão logística e variáveis contínuas modeladas usando regressão linear. Os passos do algoritmo MICE para a imputação são:

$\theta_1^{(0)} \sim P(\theta_1|Y_{1(obs)})$

$Y_{1(*)}^{(0)} \sim P(Y_1|\theta 1^{(0)})$.

$Y_1^{(0)} = (Y_{1(obs)}, Y_{1(*)}^{(0)})$.

.

.

.

$$\theta_p^{(0)} \sim P(\theta_p|Y_1,...,Y_{p-1}^{(0)},Y_{p(obs)})$$

$$Y_{p(*)}^{(0)} \sim P(Y_p|\theta_p^{(0)}).$$

$$Y_p^{(0)} = (Y_{p(obs)},Y_{p(*)}^{(0)}).$$

Em que $Y_{p(obs)}$ é o conjunto de valores observados Y_p, $Y_{p(*)}^{(0)}$ é o conjunto de Y_p valores preenchidos, $Y_p^{(0)}$ é o conjunto de ambos, e $\theta_p^{(0)}$ é o conjunto de parâmetros simulados para a distribuição condicional de Y_p variáveis dadas $Y_1,Y_2,...,Y_{p-1}$.

A fase de imputação substitui esses valores preenchidos por valores $Y_{p(*)}^{(0)}$ imputados para cada variável sequencialmente em cada iteração. Isto é, com p variáveis $Y_1,Y_2,...,Y_p$ (nessa ordem), os valores ausentes são imputados com a sequência na iteração $t+1$ da forma,

$$\theta_1^{(0)t+1} \sim P(\theta_1|Y_{1(obs)},Y_2^{(t)},...,Y_p^{(t)})$$

$$Y_{1(*)}^{(t+1)} \sim P(Y_1|\theta 1^{(t+1)}).$$

$$Y_1^{(t+1)} = (Y_{1(obs)},Y_{1(*)}^{(t+1)}).$$

.

.

.

$$\theta_p^{(t+1)} \sim P(\theta_p|Y_1,...,Y_{p-1}^{(t+1)},Y_{p(obs)})$$

$$Y_{p(*)}^{(t+1)} \sim P(Y_p|\theta_p^{(t+1)}).$$

$$Y_p^{(t+1)} = (Y_{p(obs)},Y_{p(*)}^{(t+1)}).$$

Em que $Y_{p(obs)}$ é o conjunto de Y_p valores observados, $Y_{p(*)}^{t+1}$ é o conjunto de valores imputados na iteração $(t+1)$, $Y_{p(*)}^{t}$ é o conjunto de Y_p valores preenchidos $(t=0)$ ou o conjunto de valores imputados na iteração $t(t>0)$, $Y_{p(*)}^{t+1}$ é o conjunto dos valores Y_p observados e imputados na iteração $t+1$, e $\theta_p^{(t+1)}$ é o conjunto de parâmetros simulados para a distribuição condicional de dados Y_p co-variáveis construídas a partir $Y_1,...,Y_{p-1},Y_{p+1},...,Y_p$.

Em cada iteração, um modelo especificado é ajustado para cada variável com informações faltantes usando observações observadas para essa variável, que podem incluir observações com valores imputados para outras variáveis. Esse modelo resultante é usado para imputar valores faltantes variável imputada.

O método FCS requer menos iterações do que o método MCMC, geralmente usa-se de cinco ou dez iterações para produzir resultados satisfatórios (VAN BUUREN e OUDSHOORN, 1999)(BRAND, 1999).

4 METODOLOGIA

As informações utilizadas na pesquisa foram cedidos por Andrade (2017), pesquisador da área de neurologia, foi realizado um tratamento prévio da base de dados para uma pré seleção de variáveis relevantes para explicar o desfecho do modelo, ou seja, para o desenvolvimento da transformação hemorrárica, sequela do Acidente Vascular Cerebral. A seção 5 trará mais informações sobre a aplicação realizada no estudo.

Nas subseções a seguir descreveremos brevemente as técnicas mais comuns de análise. Abordaremos técnicas para ajuste de modelos preditivos assim como também métodos de avaliação e seleção de modelos.

4.1 Regressão Linear

Entre as técnicas estatísticas utilizadas para análise de dados, os modelos de regressão são muito úteis quando tem-se o interesse de expressar, por meio de uma equação, a relação entre uma variável de interesse e um conjunto de variáveis preditoras. As aplicações da análise de regressão podem ocorrer em quase todas as áreas de atuação.

Definição :7 Seja X uma variável aleatória contínua com média α em que $-\infty < x < \infty$, e $\sigma > 0$. Pode-se dizer que X possui uma distribuição normal, assim $X \sim N(\alpha, \sigma^2)$ com função densidade de probabilidade dada por:

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{x-\alpha}{\sigma}\right)^2}. \tag{4.1}$$

Considere o modelo de regressão linear simples é dado por:

$$Y = E(Y/X = x) + \varepsilon = \alpha + \beta x + \varepsilon. \tag{4.2}$$

Em que Y é a variável resposta ou dependente (aleatória), X é a variável explicativa ou independente medida sem erro (não aleatória), α é o coeficiente de regressão, que representa o intercepto (parâmetro desconhecido do modelo à estimar), β é o coeficiente de regressão, que representa a inclinação da reta (parâmetro desconhecido do modelo à estimar) e ε é o erro aleatório ou estocástico, onde se procuram incluir todas as influências no comportamento da variável Y que não podem ser explicadas linearmente pelo comportamento da variável X.

Considerando que a média e a variância da variável ε são, $E(\varepsilon) = 0$ e $Var(\varepsilon) = \sigma^2$, respectivamente. Temos que a média e variância da variável resposta, dado um valor fixo da

variável preditora, são, respectivamente,

$$E(Y/X) = \alpha + \beta x, \tag{4.3}$$

e

$$Var(Y/X) = \sigma^2. \tag{4.4}$$

Portanto, a curva de regressão apresenta o valor esperado da variável resposta Y para um dado valor da variável preditora X . Logo, o intercepto, α, é o valor médio da variável Y quando a variável preditora X é igual a zero. Isto é, quando a variável preditora não tem peso algum na média da variável resposta.

Por outro lado, o coeficiente de inclinação, β, pode ser interpretado como a mudança na média da variável Y para a variação de uma unidade na variável x. Este modelo implica que existe uma distribuição para Y dado qualquer valor de x. A variância σ^2 é uma medida que informa a quantidade de informação não explicada pelo modelo. Valores pequenos de σ^2 retornam observações próximas da reta de regressão.

Em situações em que existe mais de uma variável explicativa ou preditora tem-se o Modelo de Regressão Linear Múltipla.

Considere $(y_1, y_2, ..., y_n)$ uma amostra aleatória selecionada da população Y, e $x_i = (x_{i1}, x_{i2}, ..., x_{ik})$, $i = 1, 2, ..., n$ seus respectivos vetores de variáveis preditivas. O modelo de regressão linear múltipla com k variáveis preditoras é então dado por:

$$Y = E(Y/X = x_i) + \varepsilon = \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + ... +_k x_{ik} + \varepsilon_i. \tag{4.5}$$

Em que Y é a variável resposta ou dependente (aleatória), X é o vetor de variável explicativas ou independentes medida sem erro (não aleatória), α é o coeficiente de regressão, que representa o intercepto (parâmetro desconhecido do modelo à estimar), $\beta = (\beta_1, \beta_2, ..., \beta_k)$ é o vetor de coeficientes de regressão, que representa a inclinação da reta (parâmetro desconhecido do modelo à estimar) e ε é o vetor de erros aleatórios.

4.2 Regressão Logística

Em situações em que a variável resposta possui apenas duas categorias, ou seja, natureza dicotômica é necessário técnicas de análise de modelos de regressão binária.

Na prática, ocorrências envolvendo estas variáveis binárias são bastante comuns. No contexto médico, o resultado poderia ser presença ou ausência de uma determinada doença,

câncer de pulmão, por exemplo. Assim como, em um cenário financeiro, por exemplo, as previsões podem ser feitas para o resultado dicotômico do sucesso ou insucesso, em uma operação de crédito. O modelo de regressão logística, é um dos modelos amplamente aplicados para a análise de proporções observadas e taxas.

O modelo de regressão logístico é um dos mais utilizado para a análise deste tipo de conjunto de dados. Dentre as razões, está a facilidade na interpretação de seus parâmetros por meio de razão de chances (odds ratio). Paula (2010) afirma que a regressão logística tem se constituído em um dos métodos predominantes de modelagem estatística para dados com respostas dicotômicas.

Isto pode ocorrer até mesmo quando a resposta de interesse não é a princípio binária. Neste caso, a variável resposta é dicotomizada de tal forma que esta possa ser modelada através da regressão logística.

Suponha que a variável resposta assuma apenas dois valores, que representa a presença ou ausência de uma característica de interesse. Estes valores são, por exemplo, denotados por 0 na ausência e por 1 na presença, da característica de interesse. Este tipo de resposta (binária) ocorre em muitos contextos. Por exemplo, a variável resposta poderia ser o resultado de uma transação de crédito em que o cliente de uma instituição financeira poderia ter se tornado inadimplente (variável resposta $y = 1$) ou ter quitado sua dívida (variável resposta $y = 0$). Esta inadimplência poderia ter sido causada por algumas variáveis, denominadas variáveis explicativas (ou preditoras), por exemplo, renda e estado civil.

Considerando que a probabilidade de um cliente Y se tornar inadimplente é p,independente dos outros individuos, temos que Y segue uma distribuição Bernoulli com esperança e a variância dadas, respectivamente, por

$$E(Y) = P[Y = 1] = p. \tag{4.6}$$

e

$$Var(Y) = p(1-p). \tag{4.7}$$

Ou seja, a esperança da variável resposta é igual à probabilidade do cliente estar inadimplente, a qual está contida no intervalo $(0,1)$. Portanto, considerando que x é uma variável explicativa, podemos definir o modelo de regressão logístico linear simples por

$$\ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x, \tag{4.8}$$

em que $Y \sim Bernoulli(p)$, com $p = P(Y = 1|x)$ a probabilidade de sucesso, β_0 e β_1 os coeficientes da regressão e log é o logaritmo natural (base , frequentemente denotado por ln). De modo similar, este modelo pode ser descrito por meio da chance (*odds*) da probabilidade de sucesso.

$$\frac{p}{1-p} = \exp(\beta_0 + \beta_1 x). \tag{4.9}$$

Portanto, a probabilidade de sucesso é dada por:

$$P[Y = 1|x] = \frac{\exp(\beta_0 + \beta_1 x)}{1 + \exp(\beta_0 + \beta_1 x)}. \tag{4.10}$$

Isto é, a probabilidade de sucesso p pode ser representada como a função de distribuição acumulada de uma distribuição logística padrão avaliada no ponto $\eta = \beta_0 + \beta_1 x$.

Quando tem-se mais de uma variável explicativa ou preditora tem-se o Modelo de Regressão Logística Múltipla.

Considere $y_1, y_2, ..., y_n$ uma amostra aleatória selecionada da população de interesse Y e $x_i = (x_{i1}, x_{i2}, ..., x_{ik})$, $i = 1, ..., n$ seus respectivos vetores de variáveis preditivas. O modelo de regressão logística múltiplo com k variáveis preditoras pode ser representado por

$$P(y_i = 1|x) = \frac{\exp(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_k x_{ik})}{1 + \exp(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_k x_{ik})}, \tag{4.11}$$

em que $\beta = (\beta_0, \beta_1, ..., \beta_k)$ é o vetor de parâmetros do modelo, os coeficientes da regressão. O modelo de regressão logística também pode ser escrito através da seguinte transformação,

$$\log\left[\frac{p_i}{1-p_i}\right] = \eta_i, \tag{4.12}$$

em que $p_i = P(y_i = 1|x)$ é a probabilidade de sucesso do i-ésimo indivíduo e o preditor linear do modelo é $\eta_i = \beta_0 + \beta_1 x_{i1} + \beta_1 x_{i2} + ... + \beta_2 x_{ik}$. Esta transformação é chamada de transformação *logit* da probabilidade de sucesso, e a razão uma transformação chamada odds. Esta transformação é bastante utilizada em estudos epidemiológicos, financeiros e áreas afins.

4.3 Interpretação dos Parâmetros de Modelos Logísticos

Na regressão logística há uma medida de chance relativa chamado *odds*, em que encontra a chance de um evento ocorrer em relação a chance do mesmo evento não ocorrer. Considere A um evento de interesse para encontrar o *odds* fixa-se uma característica como

referência, casela de referência, diz-se que a chance de ocorrência de uma característica da variável A em relação a outra característica da mesma variável é dada por:

$$Odds(A) = \frac{p(A)}{1-p(A)}. \tag{4.13}$$

$$Odds(A) \begin{cases} = 1, & \text{não há relação entre as características de interesse;} \\ > 1, & \text{a probabilidade de ocorrência da característica é maior do que a referência;} \\ < 1, & \text{a probabilidade de ocorrência da característica é menor do que a referência} \end{cases}$$

Para a interpretação dos parâmetros da regressão logística é utilizado o *odds ratio* (OR) para estimação, uma vez que representa a razão de chance de um evento acontecer em relação a ocorrência de outro evento da mesma categoria de interesse. O *odds ratio* é um número não negativo, geralmente é tomado $OR = 1$ como base para comparação e é definido como a razão entre as probabilidades para $x = 1$ e as probabilidades para $x = 0$, e é dada pela equação:

$$OR = \frac{\pi(1)/[1-\pi(1)]}{\pi(0)/[1-\pi(0)]} \tag{4.14}$$

Substituindo as expressões para o modelo de regressão logística obtemos os resultados da Tabela 1.

Tabela 1 – Expressões para o modelo logístico.

	Variável Independente (X)	
Variável Resposta (Y)	x=1	x=0
y=1	$\pi(1) = \frac{\exp^{\beta_0+\beta_1}}{1+\exp^{\beta_0+\beta_1}}$	$\pi(0) = \frac{\exp^{\beta_0}}{1+\exp^{\beta_0}}$
y=0	$1-\pi(1) = \frac{1}{1+\exp^{\beta_0+\beta_1}}$	$1-\pi(0) = \frac{1}{1+\exp^{\beta_0}}$
Total	1	1

Então, tem-se:

$$OR = \frac{\exp^{\beta_0+\beta_1}}{\exp^{\beta_0}} = \exp^{(\beta_0+\beta_1)-\beta_0} = \exp^{\beta_1} \tag{4.15}$$

Portanto, para a regressão logística com uma variável independente dicotômica codificada 1 e 0, a relação entre a razão de chances e o coeficiente de regressão é dada por:

$$OR = \exp^{\beta_1}. \tag{4.16}$$

4.4 Premissas do Modelo de Regressão Logística

Com intuito de obter um bom ajuste do modelo de regressão logística, algumas premissas devem ser satisfeitas antes de serem realizadas inferências. O afastamento destas premissas pode invalidar os resultados obtidos no ajuste do modelo. Dentre as principais premissas estão:

1. Independência entre as observações da variável resposta;
2. A variável resposta tem distribuição Bernoulli $Y \sim bernoulli(p_i)$;
3. A relação entre o logito da probabilidade de sucesso e as variáveis preditoras é linear. Isto é,

$$\log\left[\frac{p_i}{1-p_i}\right] = \beta_0 + \beta_1 x_{i1} + \beta_1 x_{i2} + ... + \beta_2 x_{ik}, \tag{4.17}$$

Uma alternativa para o modelo de regressão logística é o modelo de regressão probito que considera a relação entre a probabilidade de sucesso e as variáveis explicativas $p_i = N(\eta_i)$, onde $N(t) = \int_{-\infty}^{t} \frac{1}{\sqrt{2\pi}} \exp -\frac{1}{2}u^2 dt$ é a função de distribuição acumulada da distribuição normal padrão avaliada no ponto t.

No modelo de regressão logística não são feitas suposições sobre as distribuições das variáveis explicativas, mas elas não devem ser altamente correlacionadas.

Outra alternativa é a discretização (categorização) de variáveis para ajudar a criar uma relação mais simples entre as variáveis resposta e explicativas. De modo que as dependências não lineares possam ser modeladas como lineares.

4.5 Independência entre as Observações da Variável Resposta

Diferente do que ocorre em modelos de regressão linear, em que a independência entre as variáveis respostas podem ser verificadas analisando a distribuição dos resíduos, os resíduos obtidos a partir do modelo logístico não fornecem a mesma interpretação devido à natureza dicotômica da variável resposta.

Paula (2010) sugere o uso da função desvio (*Deviance*) na verificação da independência entre as variáveis respostas, afirmando que quando o número de graus de liberdades do modelo ajustado é menor que o desvio, pode haver indícios de superdispersão (*overdispersion*) no modelo. A superdispersão ocorre quando a variância da variável resposta é superior à variação nominal, por exemplo, no modelo de regressão logística a variância nominal é $p(1-p)$. Esta superdispersão pode ocorrer quando existe dependência entre as variáveis resposta. Uma

alternativa neste caso é o uso de modelos de quase-verossimilhança (WEDDERBURN, 1974). Estes modelos de quase-verossimilhança são amplamente utilizados no ajuste de conjunto de dados com superdispersão (PAULA, 2010).

4.6 Processo de Seleção das Variáveis do Modelo

Existem vários procedimentos e critérios para a seleção de um subconjunto de variáveis explicativas para serem incorporadas aos modelos de regressão. Os procedimentos apresentados neste sessão serão AIC, BIC e Stepwise.

4.6.1 Critério de informação de Akaike (AIC)

O critério de informação de Akaike (1974) e é definido como:

$$AIC = -2logL(\hat{\theta}) + 2(p) \tag{4.18}$$

em que $L(\hat{\theta})$ é a função de máxima verossimilhança e p o número de variáveis explicativas. Admite a existência de um modelo que descreve os dados que é desconhecido, e tenta escolher dentre um grupo de modelos avaliados, o que minimiza a divergência de Kullback-Leibler (K-L).

A distância de Kullback Leibler é uma medida da distância entre duas distribuições de probabilidade. A distância de Kullback Leibler é dada por:

$$D(w||q) = \sum w_i log(w_i/q_i) = \sum w_i log(1/q_i) w_i log(1/w_i) \tag{4.19}$$

é uma medida da ineficiência dada por assumir que a distribuição de probabilidades q sendo que a verdadeira distribuição é w. Nessa equação, p_i e q_i indicam as probabilidades do evento i de uma variável aleatória discreta nas distribuições de probabilidade p e q.

Esta divergência está relacionada à informação perdida por se usar um modelo aproximado e não a verdadeira distribuição de dados.

O AIC também pode ser definido como:

$$AIC = nln\left(\frac{SQRes}{n}\right) + 2p. \tag{4.20}$$

Em que $SQRes$ é a soma de quadrados dos resíduos. O modelo com menor valor de AIC é considerado o modelo de melhor ajuste.

4.6.2 Critério de Informação Bbayesiano - BIC (BIC)

O Critério de Informação Bayesiano(BIC), proposto por Schwarz (1978) é definido como a estatística que maximiza a probabilidade de se identificar o verdadeiro modelo dentre os avaliados. O valor do critério BIC para um determinado modelo é dado por:

$$BIC = -2logf(x_n|\theta) + plog(n), \tag{4.21}$$

em que $f(x_n|\theta)$ é o modelo escolhido, p é o número de parâmetros e n tamanho da amostra.

O modelo com menor valor de BIC é considerado o modelo de melhor ajuste.

O BIC também pode ser escrito da forma:

$$BIC = nln\left(\frac{SQRes}{n}\right) + 2(p+2)q - 2q^2, \tag{4.22}$$

Em que é dado por :

$$q = \frac{n\hat{\sigma^2}}{SQRes} + 2(p+2)q - 2q^2, \tag{4.23}$$

4.6.3 Método Stepwise

O método stepwise é usado na construção de modelos para identificar um subconjunto de preditores. Trata-se de um processo que adiciona a variável mais significativa ou remove a variável menos significativa durante cada etapa. Após cada etapa de adição de uma variável, pode-se descartar uma variável já selecionada. Considerando a estatística:

$$[SQReg(comp) - SQReg(red)]/\sigma^2 \tag{4.24}$$

Os passos do método são:

Passo 1: Ajustar o modelo reduzido de m variáveis e obter o $SQReg(reg)$;

Passo 2: Para cada variável não pertencente ao modelo do passo 1, considerar o modelo completo com a adição desta variável extra e calcular o $SQReg(comp)$ e para obter o valor da estatística(4.24);

Passo 3: Achar o máximo dos valores de (4.23) obtidos no passo 2, denotado por F_{max};

Passo 4: Seja F_{in} o quantil especificado da distribuição F com 1 e $(n-m-2)$ graus de liberdade:

- Se $F_{max} > F_{in}$, passar ao passo 5, com modelo completo composto por $(m+1)$ variáveis – as m variáveis do modelo do passo 1 e a variável cuja estatística (4.23) é igual a F_{max};

- Se $F_{max} < Fin$, passar ao passo 5, com modelo completo igual ao modelo do passo 1 ou encerrar o processo se no passo 8 da etapa anterior, nenhuma variável tiver sido eliminada;

Passo 5: Ajustar o modelo completo de k variáveis – sendo $k = m$ ou $k = (m+1)$ e obter o $SQReg(comp)$;

Passo 6: Para cada uma das k variáveis do modelo completo do passo 5, considerar o modelo reduzido, retirando esta variável e calcular o SQReg(red) para obter o valor da estatística (4.23);

Passo 7: Achar o mínimo dos k valores de (4.23) obtidos no passo 6, denotado por F_{min};

Passo 8: Seja F_{out} o quantil especificado da distribuição F com 1 e $(n-k-1)$ graus de liberdade:

- Se $F_{min} > F_{out}$ não eliminar nenhuma variável e voltar ao passo 1, iniciando nova etapa com modelo reduzido com k variáveis ou encerrar o processo se no passo 4 nenhuma variável tiver sido anexada;
- Se $F_{min} < F_{out}$ eliminar a variável cuja estatística (4.23) é igual a F_{min} e voltar ao passo 1 iniciando nova etapa com modelo reduzido com $(k-1)$ variáveis.

O procedimento do *stepwise* chega ao fim quando nenhuma variável é incluída ou descartada.

4.7 Avaliação do Modelo

Nesta sessão serão abordadas algumas ferramentas estatísticas de avaliação do poder preditivo de modelos ajustados.

4.7.1 Curva ROC

Outra forma de avaliar o modelo é através da curva Características de Operação do Receptor (*Receiver Operating Characteristic* – ROC) ou Diagrama de Lorentz (AGRESTI, 1990). A Curva ROC surgiu no campo das comunicações como uma forma de demonstrar as relações entre sinal e ruído e é uma ferramenta poderosa para medir e especificar problemas no desempenho do diagnóstico uma vez que possibilita estudar a variação da sensibilidade e especificidade para diferentes valores de corte.

Após ajustar o modelo e se atribuir um score para cada indivíduo da amostra, define-se o ponto de corte PC, tal que o i- ésimo indivíduo da amostra será classificado como não teve transformação hemorrágica se score $i >$ PC, e teve transformação hemorrágica caso contrário. Depois, constrói-se a chamada matriz de confusão, representada na Tabela 2, a qual servirá como

base para as demais medidas a serem apresentadas a seguir.

Tabela 2 – Matriz de Contingência

	Classificação Real		
Valor Previsto	Não teve hemorragia	Teve hemorragia	Total
Não teve hemorragia	VNTH	FNTH	TNTH
Teve hemorragia	FTH	VTH	TTH
Total	RNTH	RTH	N

Fonte: Autoria Própria

Os valores a serem representados na tabela são:

VNTH: número de pacientes que não teve transformação hemorrágica classificados como não teve transformação hemorrágica;

FNTH: número de pacientes teve transformação hemorrágica classificados como não teve transformação hemorrágica;

FTH: número de pacientes que não teve transformação hemorrágica classificados como teve transformação hemorrágica;

VTH: número de pacientes que teve transformação hemorrágica classificados como teve transformação hemorrágica;

RNTH: número de pacientes não teve transformação hemorrágica;

RTH: número de pacientes teve transformação hemorrágica;

TNTH: número de pacientess classificados como não teve transformação hemorrágica;

TTH: número de pacientes classificados como teve transformação hemorrágica;

N: número total de pacientes.

Com base nessas definições, é possível definir novas medidas de desempenho, conforme visto abaixo.

Especificidade: proporção de pacientes que não teve transformação hemorrágica classificados corretamente.

$$E = \frac{VNTH}{RNTH}. \tag{4.25}$$

Sensibilidade: proporção de pacientes que teve transformação hemorrágica classificados corretamente.

$$S = \frac{VTH}{RTH}. \tag{4.26}$$

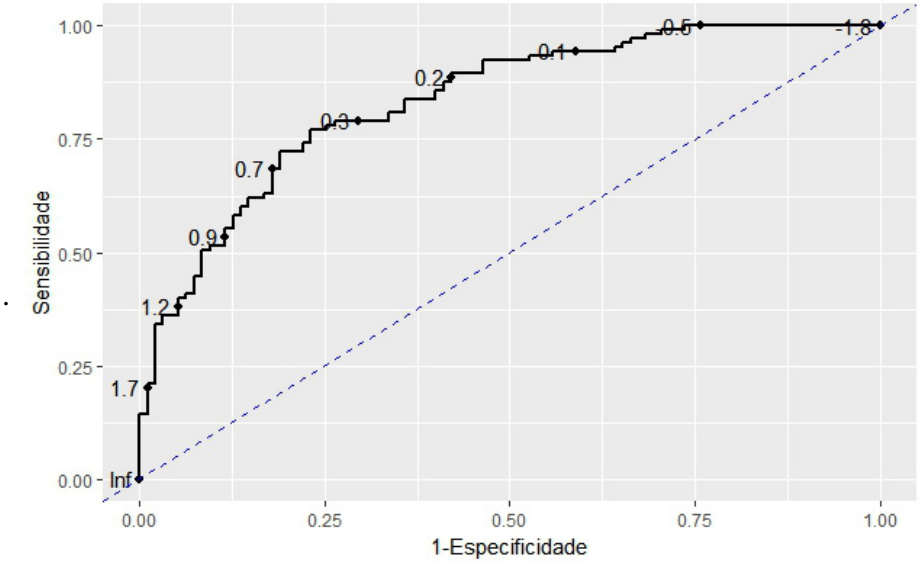
Acurácia: proporção total de acertos.

$$A = \frac{(VTH + VNTH)}{N}. \tag{4.27}$$

De modo que especificidade e sensibilidade são representadas em um gráfico, que apresenta os valores da sensibilidade no eixo das ordenadas e o complemento da especificidade (1 – especificidade) no eixo das abscissas. A linha diagonal $(x = y)$ indica uma classificação aleatória, isto é, um modelo com um poder preditivo nulo.

Na Figura 1 observa-se a curva de ROC.

Figura 1 – Curva ROC



Fonte: Autoria Própria

Quanto maior a distância entre a curva ROC da linha diagonal, melhor será o modelo.

4.7.2 DFBETA

DFBETA é uma medida que mensura a influência que a observação i tem sobre o coeficiente de X_j. Esta é definida da seguinte forma:

$$DFB_{j(i)} = \frac{\hat{\beta}_j - \hat{\beta}_{j(i)}}{\sqrt{QME_i c_{jj}}}, \qquad j = 0, 1, ..., p. \tag{4.28}$$

Em que c_{jj} é o j-ésimo elemento da diagonal de $(X'X)^{-1}$. Quando tem-se valor alto para a medida DFBETA é indícios que a observação i influência na estimativa do coeficiente angular da variável explicativa X_j.

São apontadas como observações influentes aquelas que,

1 $|DFBETA| > 1$, para amostras pequenas.

2 $|DFBETA| > 2/\sqrt{n}$, para amostras grandes.

4.8 Análise Discriminante

Discriminação e classificação são técnicas multivariadas relacionadas com separação de objetos (ou observações) em conjuntos distintos e alocação de novos objetos para grupos previamente definidos.

A análise discriminante é um procedimento separativo e frequentemente empregado como base para investigar as diferenças observadas quando as relações causais não são explícitas.

Os procedimentos de classificação são menos exploratórios uma vez que eles levam a regras bem definidas, que podem ser usadas para atribuir novos objetos.

A terminologia discriminante foi introduzida por R.A. Fisher no primeiro tratamento moderno de problemas separativos. Um termo mais empregado para esse objetivo, no entanto, é a separação.

4.8.1 Separação e Classificação para Duas Populações

Suponha que há interesse em (1) separar duas classes de objetos ou (2) atribuir um novo objeto a uma das duas classes. Isto é conveniente para rotular as classes π_1 e π_2. Os objetos são normalmente separados ou classificados com base em medições, por exemplo, de p variáveis aleatórias associadas a $X' = [X_1, X_2, ..., X_p]$. Os valores observados de X diferem em certa medida uma classe para a outra. Podemos pensar na totalidade dos valores da primeira classe como

sendo a população de valores x para π_1 e aqueles da segunda classe como a população de x valores para π_2. Estas duas populações podem então ser descritas por probabilidade das funções de densidade $f_1(x)$ e $f_2(x)$ e, consequentemente, podemos falar em atribuir observações para populações ou objetos para classes intercambiáveis. Tais objetos devem ser separados em duas classes rotuladas com base nos valores de variáveis que os caracteriza.

As regras de alocação ou classificação geralmente são desenvolvidas a partir de amostras de *learning*. Características medidas de objetos selecionados aleatoriamente conhecidos das duas populações tem suas diferenças examinadas. Essencialmente, o conjunto de todos os possíveis resultados amostrais são divididos em duas regiões, R_1 e R_2, de modo que, se a observação cai no R_1 é alocado para a população π_1, e se cai em R_2, alocamos para a população π_2.

Outro aspecto da classificação é o custo. Suponha que classificando um objeto π_1 como pertencer a π_2 representa um erro mais grave do que classificar um objeto π_2 como pertencente para π_1. Então, deve-se ter cuidado ao fazer a designação anterior. Um exemplo seria, deixar de diagnosticar uma doença potencialmente fatal é substancialmente mais "caro"do que concluir que a doença está presente quando, na verdade, não está. Um ótimo procedimento de classificação deve, sempre que possível, contabilizar os custos associados com erro de classificação.

Seja $f_1(x)$ e $f_2(x)$ as funções de densidade de probabilidade associadas ao vetor $px1$ da variável aleatória X para as populações π_1 e π_2, respectivamente. Um objeto com medidas associadas x deve ser atribuído a π_1 ou π_2. Seja Ω o espaço amostral, isto é, todas as observações possíveis x. Seja o R_1 o conjunto de valores x para os quais classificamos objetos como π_1 e $R_2 = \Omega - R_1$ os demais valores de x para os quais classificamos objetos como π_2. Como todo objeto deve ser atribuído a uma e apenas uma das duas populações, os conjuntos R_1 e R_2 são mutuamente exclusivos.

Para $p = 2$, a probabilidade condicional, $P(2|1)$, de classificar um objeto como π_2 quando, de fato, é de π_1 é

$$\mathbb{P}(2|1) = \mathbb{P}(X \in R_2|\pi_1) = \int_{R_2=\Omega-R_1} f_1(x)dx \tag{4.29}$$

e de modo similar, a probabilidade condicional, $p(1|2)$, de classificar um objeto como π_1 quando é realmente de π_2 é

$$\mathbb{P}(1|2) = \mathbb{P}(X \in R_1|\pi_2) = \int_{R_1=\Omega-R_2} f_2(x)dx \tag{4.30}$$

Seja p_1 a probabilidade priori de π_1 e p_2 a probabilidade priori de π_2, onde $p_1 + p_2 = 1$. Então as probabilidades totais de classificação correta e incorreta são dadas por:

- P(a observação é corretamente classificada como π_1 = P(a observação vem do π_1 e é classificado corretamente como π_1)= $P(X \in R_1|\pi_1)P(\pi_1) = p(1|1)p_1$.
- P(a observação é classificada erroneamente como π_1) = P(a observação vem de π_2 e é erroneamente classificado como π_1)= $P(X \in R_1|\pi_2)P(\pi_2) = P(1|2)p_2$.
- P(a observação é corretamente classificada como π_2) = P(a observação vem de π_2 e está corretamente classificado como π_2) = $P(X \in R_2)P(\pi_2) = P(2|2)p_2$.
- P(a observação é classificada erroneamente como π_2) = P(a observação vem de π_1 e é erroneamente classificado como π_2)= $P(X \in R_2|\pi_1)P(\pi_1) = P(2|1)p_1$.

Os custos de erros de classificação podem ser definidos por uma matriz de custos:

	π_1	π_2
π_1	0	$c(2\|1)$
π_2	$c(1\|2)$	0

Os custos são (1) zero para a classificação correta, (2) $c(1|2)$ quando uma observação de π_2 é classificado incorretamente como π_1 e (3) $c(2|1)$ quando uma observação π_1 é incorretamente classificado como π_2. Para qualquer regra, o custo médio ou esperado da classificação de classe (ECM) é fornecido multiplicando as entradas fora da diagonal por suas probabilidades de ocorrência. Assim temos,

$$ECM = c(2|1)P(2|1)p_1 + c(1|2)P(1|2)p_2 \tag{4.31}$$

Logo tem-se, as regiões R_1 e R_2 que minimizam o ECM são definidas pelo valores x para os quais as seguintes desigualdades são válidas:

$$R_1 : \left(\frac{f_1(x)}{f_2(x)}\right) \geq \left(\frac{c(1|2)}{c(2|1)}\right)\left(\frac{p_2}{p_1}\right) \tag{4.32}$$

e

$$R_2 : \left(\frac{f_1(x)}{f_2(x)}\right) < \left(\frac{c(1|2)}{c(2|1)}\right)\left(\frac{p_2}{p_1}\right) \tag{4.33}$$

Uma regra de classificação razoável deve ter um ECM tão pequeno quanto possível.

4.8.2 Classificação com Duas Populações Multivariadas Normais

Geralmente, procedimentos de classificação baseados em populações normais predominam por causa de sua eficiência razoavelmente alta em uma ampla variedade de modelos populacionais. Assumindo que $f_1(x)$ e $f_2(x)$ são densidades normais multivariadas, a primeira com vetor médio μ_1 e matriz de covariância Σ_1 e a segundo com vetor médio μ_2 e matriz de covariância Σ_2. O caso especial de matrizes de covariâncias iguais leva a uma linear linear estatística de classificação.

4.8.2.1 Classificação de Populações Normais com $\Sigma_1 = \Sigma_2 = \Sigma$

Suponha que as densidades conjuntas de $X' = [X_1, X_2, ..., X_p]$ para as populações π_1 e π_2 são dadas por:

$$f(x) = \frac{1}{(2\pi)^{\frac{p}{2}}|\Sigma|^{\frac{1}{2}}} \exp\left[-\frac{1}{2}(x-\mu_i)'\Sigma^{-1}(x-\mu_i)\right] \tag{4.34}$$

Para $i = 1, 2$. Suponha também que os parâmetros populacionais μ_1, μ_2 e Σ sejam conhecidos. Então depois cancelando os termos $(2\pi)^{\frac{p}{2}}|\Sigma|^{\frac{1}{2}}$ as regiões mínimas da ECM são:

$$R_1 : \exp\left[-\frac{1}{2}(x-\mu_i)'\Sigma^{-1}(x-\mu_i) + \frac{1}{2}(x-\mu_i)'\Sigma^{-1}(x-\mu_i)\right] \geq \left(\frac{c(1|2)}{c(2|1)}\right)\left(\frac{p_2}{p_1}\right) \tag{4.35}$$

e

$$R_2 : \exp\left[-\frac{1}{2}(x-\mu_i)'\Sigma^{-1}(x-\mu_i)+\frac{1}{2}(x-\mu_i)'\Sigma^{-1}(x-\mu_i)\right] < \left(\frac{c(1|2)}{c(2|1)}\right)\left(\frac{p_2}{p_1}\right) \tag{4.36}$$

De modo que a regra de alocação que minimiza o ECM que aloca x_0 para π_1 é dada por:

$$(\mu_1-\mu_2)'\Sigma^{-1}x_0-\frac{1}{2}(\mu_1-\mu_2)'\Sigma^{-1}(\mu_1+\mu_2) \geq ln\left[\left(\frac{c(1|2)}{c(2|1)}\right)\left(\frac{p_2}{p_1}\right)\right] \tag{4.37}$$

e x_0 aloca π_2 caso contrário. Quando tem-se medidas desconhecidas trabalhamos com suas respectivas estimativas.

4.8.2.2 Classificação de Populações Normais com $\Sigma_1 \neq \Sigma_2$

Considere as densidades normais multivariadas em (6) com $\Sigma_i, i = 1,2$, substituindo Σ. Assim, as matrizes de covariância, como os vetores médios, são diferentes entre si para as duas populações. Como vimos, as regiões de ECM mínima e probabilidade total mínima de erro de classificação (TPM) dependem da razão das densidades,$f_1(x)/f_2(x)$, ou, equivalentemente, do logaritmo natural da razão de densidade, $\ln[f_1(x)/f_2(x)] = \ln[f_1(x)] - \ln[f_2(x)]$. Quando as densidades normais multivariadas têm diferentes estruturas de covariância, os termos na razão de densidade envolvendo $|\Sigma_i|^{1/2}$ não se cancelam como quando $\Sigma 1 = \Sigma 2$. De modo que as regiões de classificação são dadas por:

$$R_1 : -\frac{1}{2}x'(\Sigma_1^{-1}-\Sigma_2^{-1})x+(\mu_1'\Sigma_1^{-1}-\mu_2'\Sigma_2^{-1})x-k \geq \left[\left(\frac{c(1|2)}{c(2|1)}\right)\left(\frac{p_2}{p_1}\right)\right] \tag{4.38}$$

e

$$R_2 : -\frac{1}{2}x'(\Sigma_1^{-1}-\Sigma_2^{-1})x+(\mu_1'\Sigma_1^{-1}-\mu_2'\Sigma_2^{-1})x-k < \left[\left(\frac{c(1|2)}{c(2|1)}\right)\left(\frac{p_2}{p_1}\right)\right] \tag{4.39}$$

Em que,

$$k = \frac{1}{2}\ln\left(\frac{|\Sigma_1|}{|\Sigma_2|}\right)+\frac{1}{2}(\mu_1'\Sigma_1^{-1}-\mu_2') \tag{4.40}$$

As regiões de classificação são definidas por funções quadráticas de x. Quando $\Sigma_1 \neq \Sigma_2$, o termo quadrático, $\frac{1}{2}x'(\Sigma_1-\Sigma_2)x$ desaparece, e as regiões definidas por(4.37 - 4.38) reduzir para aqueles definidos por (4.31 - 4.32).

4.8.3 Separação e Classificação para Várias Populações

Essa abordagem não leva a conclusões gerais uma vez que as propriedades dependem onde as populações estão localizadas.

Considere $f_i(x)$ a densidade associada à população $\pi_i, i = 1,2,...,g$. Seja π_i = a probabilidade à priori da população $\pi_i, i = 1,2,...,g$. $c(k|i) =$ o custo de alocar um item para π_k quando, de fato, ele pertence π_i, em que $k,i = 1,2,...,g$. Para $k = i$, $c(i|i) = O$. Finalmente, seja R_k o conjunto de $x's$ classificado como π_k e

$$P(k|i) = P(\pi_k|\pi_i) = \int_{R_k} f_1(x)\partial x \tag{4.41}$$

para $k,i = 1,2,...,g$ com $P(i|i) = 1 - \sum_{k=1,k\neq i}^{g} P(k|i)$.

O custo esperado condicional de classificar de forma errada um x de π_1 para π_2 ou π_g é dado por:

$$ECM(l) = P(2|1)c(2|1) + P(3|1)c(3|1) + ... + P(g|l)c(g|1) = \sum_{k=2}^{g} P(k|l)c(k|1) \tag{4.42}$$

De uma forma similar, podemos obter os custos condicionais esperados de $ECM(2),...,ECM(g)$. Multiplicando cada ECM condicional pela sua probabilidade de ocorrência, assim tem-se:

$$ECM = p_1ECM(1) + p_2ECM(2) + ... + p_gECM(g) = \sum_{i=1}^{g} p_i \left(\sum_{k=1,k\neq i}^{g} P(k|i)c(k|i) \right) \tag{4.43}$$

O procedimento de classificação ideal é aquele em que há o mínimo possível de separação das regiões de classificação $R_1, R_2, ..., R_g$.

4.8.3.1 Classificação com Populações Normais

Um caso especial ocorre quando tem-se:

$$f(x) = \frac{1}{(2\pi)^{\frac{p}{2}} |\Sigma_i|^{\frac{1}{2}}} \exp\left[-\frac{1}{2}(x-\mu_i)'{\Sigma_i}^{-1}(x-\mu_i) \right], \tag{4.44}$$

com $i = 1,2,...,g$. são densidades normais multivariadas com vetores médias μ_i e matrizes de covariância σ_i. E se, $c(i|i) = 0$, $c(k|i) = 1$, $k \neq i$. Alocar x para π_k se

$$lnp_kf_k(x) = lnp_k - \left(\frac{p}{2}\right) ln(2\pi) - \frac{1}{2}ln|\Sigma_k| - \frac{1}{2}(x-\mu_k)'{\Sigma_i}^{-1}(x-\mu_k) = max_i lnp_if_i(x), \tag{4.45}$$

A constante $(p/2)$ e $ln(2\pi)$ pode ser suprimida uma vez que é a mesma para todas as populações. Portanto, definimos o escore de discriminação quadrática para a i-população como sendo

$$d_i^Q(x) = -\frac{1}{2}ln|\Sigma_i| - \frac{1}{2}(x-\mu_k)'{\Sigma_i}^{-1}(x-\mu_k) + lnp_i. \tag{4.46}$$

Em que $i = 1,2,...,g$. O escore quadrático d_i^Q é composto de contribuições da variância generalizada $|\Sigma_i|$, da probabilidade prévia π_i e do quadrado da distância de x à média populacional π_i.

4.8.3.2 Regra de Probabilidade Mínima Total de Erros de Classificação (TPM) para Populações Normais - Diferentes Σ_i

Aloca-se x para π_k se o escore quadrático $\hat{d}_i^Q(x) = max(\hat{d}_i^1(x), \hat{d}_i^2(x), ..., \hat{d}_i^g(x))$ onde $\hat{d}_i^Q(x)$ é dado por:

$$\hat{d}_i^Q(x) = -\frac{1}{2} ln|S_i| - \frac{1}{2}(x-\mu_k)' \S_i^{-1}(x-\mu_k) + lnp_i, \tag{4.47}$$

com $i = 1, 2, ..., g$. Uma simplificação é possível se as matrizes de covariância populacional, $|\Sigma_i|$ são iguais. Quando $\Sigma_i = \Sigma$, para $i = 1, 2, ..., g$, o escore discriminante em (4.45) se torna,

$$d_i^Q(x) = -\frac{1}{2} ln|\Sigma| - \frac{1}{2} x'\Sigma^{-1}x + \mu_i'\Sigma^{-1}x - \frac{1}{2}\mu_i'\Sigma^{-1}\mu_i + lnp_i. \tag{4.48}$$

Simplificando as constantes, os termos restantes consistem em uma constante $c_i = lnp_i \frac{1}{2}\mu_i'\Sigma^{-1}\mu_i$ e uma combinação linear dos componentes de x. Assim temos que o escore discriminante linear é dado por:

$$d_i(x) = \mu_i'\Sigma^{-1}x - \frac{1}{2}\mu_i'\Sigma^{-1}\mu_i + lnp_i, \tag{4.49}$$

para $i = 1, 2, ..., g$. A estimativa $\hat{d}_i(x)$ do escore discriminante linear $d_i(x)$ é baseado na estimativa conjunta de Σ dado por:

$$S_c = \frac{1}{n_1 + n_2 + ... + n_g - g}((n_1 - 1)S_2 + (n_2 - 1)S_2 + ... + (n_g - 1)S_g), \tag{4.50}$$

Assim tem-se:

$$\hat{d}_i(x) = x_i'S_c^{-1}x - \frac{1}{2}x_i'S_c^{-1}\bar{x}_i + lnp_i, \tag{4.51}$$

para $i = 1, 2, ..., g$.

4.8.3.3 Regra de TPM Mínima Estimada para Populações Normais de Igualdade de Covariância

Alocar x para π se o escore discriminante linear $\hat{d}_i^Q(x) = max(\hat{d}_i^1(x), \hat{d}_i^2(x), ..., \hat{d}_i^g(x))$ em que $\hat{d}_i(x)$ foi expressa em (4.50). Um classificador equivalente para o caso de covariância igual pode ser obtido de (4.45) sem utilizar o termo constante, $\frac{1}{2} ln|\Sigma|\cdot$ O resultado, com estimativas de amostra inseridas para quantidades populacionais desconhecidas, também pode ser interpretado em termos das distâncias quadradas de modo que tem-se:

$$D_i^2(x) = (x - \bar{x}_i)'S_c^{-1}(x - \hat{x}_i). \tag{4.52}$$

Em que x é o vetor de médias da amostra $\hat{x_i}$. A regra de alocação é então atribuir x à população π_i quando $-\frac{1}{2}D_i^2(x)+lnp_i$ é o maior. De modo que essa regra é equivalentemente a expressão (4.46) que atribui x à população "mais próxima". Se as probabilidades anteriores são desconhecidas, o procedimento usual é definir $p_1 = p_2 = p_g = 1/g$ e a observação é então atribuída à população mais próxima.

4.9 Ferramentas Utilizadas

O *software* utilizado para as análises foi o *Statistical Analysis System Studio* (SAS Studio) e nesta seção será explicado as funções usadas na aplicação.

1. PROC MI

   O procedimento do MI realiza várias imputações de informações faltantes. Ele cria conjuntos de dados com imputação múltipla para dados multivariados incompletos, usando métodos que incorporam variabilidade apropriada entre as *m* imputações. O método de escolha depende dos padrões de ausência das informações. Tais métodos seram definidos a seguir:

   - BY: em que especifica grupos nos quais análises de imputação múltiplas são executadas.
   - CLASS: lista as variáveis de classificação na instrução VAR, podendo ser string ou numérico.
   - EM: usa o algoritmo EM para calcular a estimativa de probabilidade máxima dos dados com valores faltantes, assumindo uma distribuição normal multivariada para os dados.
   - FCS: usa uma imputação multivariada pelo método de equações encadeadas para imputar valores para um conjunto de dados com um padrão arbitrário ausente, supondo que exista uma distribuição conjunta para os dados.
   - FREQ: especifica a frequência de ocorrência de outros valores na observação.
   - MCMC: usa um método de Monte Carlo da cadeia de Markov para imputar valores para um conjunto de dados com um padrão arbitrário omisso, assumindo uma distribuição normal multivariada para os dados.
   - MONOTONE: especifica métodos monótonos para imputar variáveis contínuas e de classificação para um conjunto de dados com um padrão omisso monótono.
   - TRANSFORM: especifica as variáveis a serem transformadas antes do processo de

imputação.

- VAR: lista as variáveis numéricas a serem analisadas. Caso não seja utilizada a declaração VAR, todas as variáveis numéricas não listadas em outras declarações serão usadas.

Na Tabela 3 tem-se o resumo dos argumentos da função PROC MI.

Tabela 3 – Resumo dos argumentos da função PROC MI

Função	Descrição
DATA	especifica o conjunto de dados de entrada
OUT	especifica o conjunto de dados de saída com valores imputados.
NIMPUTE	especifica o número de imputações.
SEED	especifica semente para iniciar o gerador de números aleatórios.
ROUND	especifica unidades para arredondar valores de variáveis imputadas.
MAXIMUM	especifica valores máximos para valores de variáveis imputados.
MINIMUM	especifica valores mínimos para valores variáveis atribuídos.
MINMAXITER	especifica o número máximo de iterações para imputar no intervalo especificado.
SINGULAR	especifica tolerância de singularidade.
ALPHA	especifica o nível para o intervalo de confiança $(1-\alpha)$.
MU_0	especifica significa sob a hipótese nula.
NOPRINT	suprime toda a saída exibida.
SIMPLE	exibe estatísticas e correlações univariadas.

2. PROC MIANALYSE

   O procedimento MIANALYZE combina os resultados das análises de imputações e gera inferências estatísticas válidas. O procedimento MIANALYZE lê estimativas de parâmetros e erros padrão associados ou matriz de covariância que são calculados pelo procedimento estatístico padrão para cada conjunto de dados imputado, posteriormente deriva uma inferência univariada válida para esses parâmetros. Com uma suposição adicional sobre a população entre e dentro de matrizes de covariância de imputação.

   Na Tabela 4 observa-se o resumo dos principais argumentos da função PROC MIANALYZE.

Tabela 4 – Resumo dos argumentos da função PROC MIANALYZE

Função	Descrição
BY	especifica o conjunto de dados de entrada.
CLASS	lista as variáveis de classificação na instrução MODELEFFECTS.
MODELEFFECTS	lista os efeitos a serem analisados.
STDERR	lista os erros padrão associados aos efeitos na instrução MODELEFFECTS.
TEST	testa hipóteses lineares sobre os parâmetros.

3. PROC LOGISTIC

   O procedimento LOGISTIC ajusta modelos de regressão logística linear para dados de resposta binárias pelo método de máxima verossimilhança. As estimativas da razão de chances são apresentadas juntamente com as estimativas dos parâmetros. Intervalos de confiança para os parâmetros de regressão e *odds ratios* podem ser calculados com base na função de quasi-verossimilhança ou na normalidade assintótica dos estimadores de parâmetros.

   Na Tabela 5 tem o resumo dos principais argumentos da função PROC LOGISTIC.

Tabela 5 – Resumo dos argumentos da função PROC LOGISTIC

Função	Descrição
BY	especifica o conjunto de dados de entrada.
CLASS	lista as variáveis de classificação.
CONTRAST	valores de efeito.
EFFECT	tipo de efeito.
EFFECTPLOT	tipo de plot.
ESTIMATE	estimativas dos parâmetros.
EXACT	intercepto.
ODDSRATIO	odds radio das estimativas dos parâmetros.
OUTPUT	conjunto de dados.
ROC	curva ROC.

4. PROC GENMOD

   O procedimento GENMOD ajusta um modelo linear generalizado dos dados por estimativa de máxima verossimilhança do vetor de parâmetros. Quando não existe solução de forma fechada para as estimativas de máxima verossimilhança dos parâmetros, o procedimento GENMOD estima os parâmetros do modelo numericamente por meio de um processo de ajuste iterativo.

Tabela 6 – Resumo dos argumentos da função PROC GENMOD

Função	Descrição
BY	especifica o conjunto de dados de entrada.
CLASS	lista as variáveis de classificação.
CONTRAST	valores de efeito.
DEVIANCE	variação do modelo.
EFFECTPLOT	tipo de plot.
ESTIMATE	estimativas dos parâmetros.
EXACT	intercepto.
OUTPUT	conjunto de dados.
BAYES	uma das funções disponíveis.

5. PROC GLMSELECT

   O procedimento GLMSELECT executa a seleção de variáveis dos modelos, este procedimento oferece recursos abrangentes para a seleção com uma ampla variedade de critérios de seleção e interrupções.

   Na Tabela 7 tem-se o resumo dos argumentos da função PROC GLMSELECT.

Tabela 7 – Resumo dos argumentos da função PROC GLMSELECT

Função	Descrição
DATA	nomeia um conjunto de dados a ser usado para a regressão.
MAXMACRO	define o número máximo de variáveis produzidas.
TESTDATA	nomeia um conjunto de dados contendo dados de teste.
VALDATA	nomeia um conjunto de dados contendo dados de validação.
PLOTS	visões e gráficos.
OUTDESIGN	solicita um conjunto de dados contendo a matriz.
NAMELEN	define o comprimento de efeitos em tabelas e conjuntos de dados de saída.
NOPRINT	suprime a saída exibida incluindo gráficos.
SEED	define a semente usada para geração de números pseudo-aleatórios.

5 APLICAÇÃO

O Acidente Vascular Cerebral (AVC) é umas das principais causas de morte e incapacidade no mundo. Os riscos para o desenvolvimento de AVC aumentam a medida que as pessoas ficam mais velhas, principalmente naquelas com idade superior aos 55 anos, porém em pessoas mais jovens o aparecimento dessa doença geralmente está associado à condições genéticas. Existe dois tipos de AVC: o isquêmico, quando ocorre a obstrução dos vasos sanguíneos acarretando em uma parada do sangue que chega ao cérebro, e o hemorrágico, que ocasiona hemorragia no cérebro.

A principal sequela cerebral do AVC isquêmico é a transformação hemorrágica (TH). Encontrar soluções para reduzir o risco de TH é uma das preocupações no âmbito médico, com o intuito de aumentar a confiabilidade do tratamento no desenvolvimento de estudos é interessante detectar na seleção de pacientes aqueles com maiores riscos de desenvolverem TH.

Neste trabalho apresenta-se uma aplicação do problema. Para tal desenvolvimento será utilizada a técnica de regressão logística para elaboração de ferramenta preditiva do risco de transformação hemorrágica em pacientes com Acidente Vascular Cerebral isquêmico em uma unidade hospitalar pública de referência em Fortaleza, Ceará, na qual dentre suas covariáveis, algumas delas possuem uma quantidade representativa de dados omisso. Dessa forma, que o objetivo principal do estudo é aplicar técnicas diferentes de tratamentos de dados faltantes para cada variável de acordo com sua natureza e ajustar um modelo preditivo e posteriormente comparar com uma base de dados mais completa obtida em outr momento da pesquisa.

A base de dados do estudo foi retirada de uma instituição hospitalar, disponibilizada por um neurologista e obtidos por meio dos registros hospitalares presentes no prontuário do paciente em estudo. Foram considerados aptos a participar da pesquisa todos os pacientes com diagnóstico inicial de AVC isquêmico admitidos pelo serviço de Emergências Médicas do Hospital Geral de Fortaleza, unidade pública de saúde referência em atendimento neurológico para o estado, no contexto da Secretaria de Saúde do Estado do Ceará (ANDRADE, 2017).

A ferramenta *SAS Studio* versão estudante, que é gratuito, foi utilizada como auxílio computacional das análises realizadas.

5.1 Análise do Banco de Dados Incompleto

Nesta seção serão apresentados resultados referentes a análise da base de dados incompleta, isto é, com informações omissas. O tratamento aplicado foi Imputação Múltipla uma vez que temos o interesse de avaliar por meio de comparações com a base de dados completa a precisão dos resultados dos modelos.

5.1.1 Análise Descritiva

Com intuito de evitar informações redundante no modelo e redução do custo computacional foram selecionadas pelo pesquisador as variáveis consideradas importantes para o desenvolvimento de TH no paciente.

Das 43 variáveis selecionadas, somente 19 variáveis não possuíram nenhuma informação faltantes, ou seja, todas as observações do indivíduo foram preenchidas.

Tabela 8 – Variáveis do conjunto de dados sem informações faltantes.

Idade	Sexo	Leucócitos	Tabagismo	Extabagista
Estilismo atual	Insuciência cardíaca	Coronaripatia	AAS na internação	Dias início clopidogel
Prolaxia	Dias início prolaxia	Tipo prolaxia	Dose profilaxia	AAS e clopidogel
Dias início aas e clopidogel	Toast	Bamford	NIHSS	

Na Tabela 9 tem-se as variáveis que possuem observações faltantes e suas quantidades de informações observadas.

Foram retiradas as variáveis dependentes de outras variáveis preditoras restando 33 variáveis para serem analisadas. Por exemplo, pacientes que usaram AAS na internação (resposta é sim ou não), e na outra questão é perguntado os dias de início do uso de AAS, questão que será respondida somente por pacientes que tiveram uso de AAS. Após a pré seleção de variáveis as demais foram testadas no modelo e o critério utilizado para a escolha das variáveis para o modelo final foi o método de *stepwise*. Assim, as variáveis selecionadas para o modelo final estão descritas na Tabela 10.

Tabela 9 – Quantidade de informações faltantes e observadas segundo as variáveis do conjunto de dados.

Variáveis	Observadas	Faltantes
Plaquetas	379	1
Hipertensão arterial sistêmica	379	1
AAS	379	1
DRC	379	1
Clopidogel na internação	379	1
WAKEUP	378	2
DM	378	2
Estatina	378	2
Etilismo prévio	378	2
Creatinina	372	8
Hipodensidade	372	8
Glicemia	370	10
Dose AAS na internação	350	30
RANKIN	347	33
Dias início AAS	344	36
TTPA	336	44
ASPECTS	331	49
TAP	323	57
Tipo estatina na internação	284	96
Dose estatina	284	96
Dias início estatina	276	104
Território da artéria	274	106
ASPECTS TC	169	211
Dose AAS prévio	104	276

Fonte: Feita no SAS Studio

De modo que temos a variável glicemia (*mg/dL*) como sendo a única do tipo contínua a entrar no modelo. Já as variáveis sexo, hipertensão arterial sistêmica, diabetes, etilismo prévio, clopidogel na internação e insuciência cardíaca são classicadas como qualitativas nominais, em que a primeira possui característica como feminino e masculino e as demais variáveis possuem resposta característica sim ou não. Por fim, a variável Bamford também é tida como qualitativa nominal, podendo ter como resposta: POCS, TAC, PACS e LACS, e as variáveis Alberta Stroke Program Early CT Score e NIHSS são classicadas como variáveis qualitativas ordinais, em que a variável Alberta Stroke Program Early CT Score é uma classicação segundo o tamanho do AVC, essa escala varia de 1 a 10.

Em virtude da quantidade de pacientes com o tamanho do AVC entre 1 a 4 ser muito reduzida então formou-se uma nova característica juntando os pacientes que tiveram o tamanho do AVC entre essa classicação, de modo que foram comparadas com as demais classicações da variável em questão.

A variável Bamford é mensurada de acordo com a classicação clínica e anatômica das áreas cerebrais infartadas. Podendo ser classificadas como Infarto de circulação posterior, Infarto de circulação anterior total, Infarto de circulação anterior parcial, Infarto lacunar. Os

Tabela 10 – Descrição das variáveis utilizadas no modelo de regressão logística.

Variável	Categoria	Descrição
Glicemia		Glicemia (mg/dL) na admissão
Sexo	Masculino	Sexo masculino
	Feminino	Sexo feminino
Hipertensão arterial sistêmica	Sim	Com hipertensão arterial sistêmica
	Não	Sem hipertensão arterial sistêmica
Diabetes	Sim	Com diabetes
	Não	Sem diabetes
Estilismo prévio	Sim	Qualquer ingestão de álcool
	Não	Parou com qualquer ingestão de álcool nos últimos 30 dias
Clopidogel na internação	Sim	Uso de clopidogel na internação
	Não	Não uso de colpidogel na internação
Insuficiência cardíaca	Sim	Com insuficiência cardíaca
	Não	Sem insuficiência cardíaca
Bamford	POCS	Infarto de circulação posterior
	TACS	Infarto de circulação anterior total
	PACS	Infarto de circulação anterior parcial
	LACS	Infarto lacunar
NIHSS	>20	Faixa maior que 20
	0 - 8	Faixa entre 0 e 8
	9 - 14	Faixa entre 9 e 14
	15 - 20	Faixa entre 15 e 20
Alberta Stroke Program Early CT Score	1	Maiores tamanhos de AVC
	5	5º maior tamanho de AVC
	6	6º maior tamanho de AVC
	7	7º maior tamanho de AVC
	8	8º maior tamanho de AVC
	9	9º maior tamanho de AVC
	10	Menores tamanhos de AVC

pacientes classicados com Infarto de circulação anterior total são os casos mais graves.

A variável NIHSS é uma escala que classifica as faixas de gravidade mensuradas na admissão do paciente, essa escala é mensurada de acordo com a variação: 0 à 8, 9 à 14, 15 à 20 e maiores que 20, em que quanto maior a faixa maior a gravidade do paciente ao ser admitido no hospital.

As variáveis sexo, insuciência cardíaca, Bamford e NIHSS não apresentaram dados faltantes. No conjunto de variáveis selecionadas, a que apresentou maior quantidade de informações faltantes foi a Alberta Stroke Program Early CT Score totalizando 49 observações.

Na Tabela 11 estão dispostas a quantidade de informações faltantes em cada variável selecionada para o modelo.

Tabela 11 – Quantidade de informações faltantes segundo as variáveis selecionadas para o modelo final.

Variáveis	Quantidade	
Glicemia	10	2,63%
Sexo	0	0,00%
Hipertensão	1	0,26%
Diabetes	2	0,52%
Etilismo prévio	2	0,52%
Clopidogel na internação	1	0,26%
Insuficiência cardíaca	0	0,00%
Bamford	0	0,00%
NIHSS	0	0,00%
Alberta Stroke Program Early CT Score	49	12,89%

Fonte: Feita no SAS Studio

Dos 380 pacientes observados, 78 desenvolveram TH, os outros 302 pacientes que tiveram AVC não desenvolveram TH.

Os indivíduos do sexo masculino, proporcionalmente, apresentaram uma maior quantidade no desenvolvimento de TH do que indivíduos do sexo feminino conforme pode-se notar na tabela 12. Aqueles que não tem hipertensão arterial sistêmica apresentaram uma proporção maior de desenvolver TH do que pacientes que tem hipertensão arterial sistêmica. A proporção em relação a desenvolver TH é maior para pacientes que tem diabetes do que indivíduos que não tem diabetes.

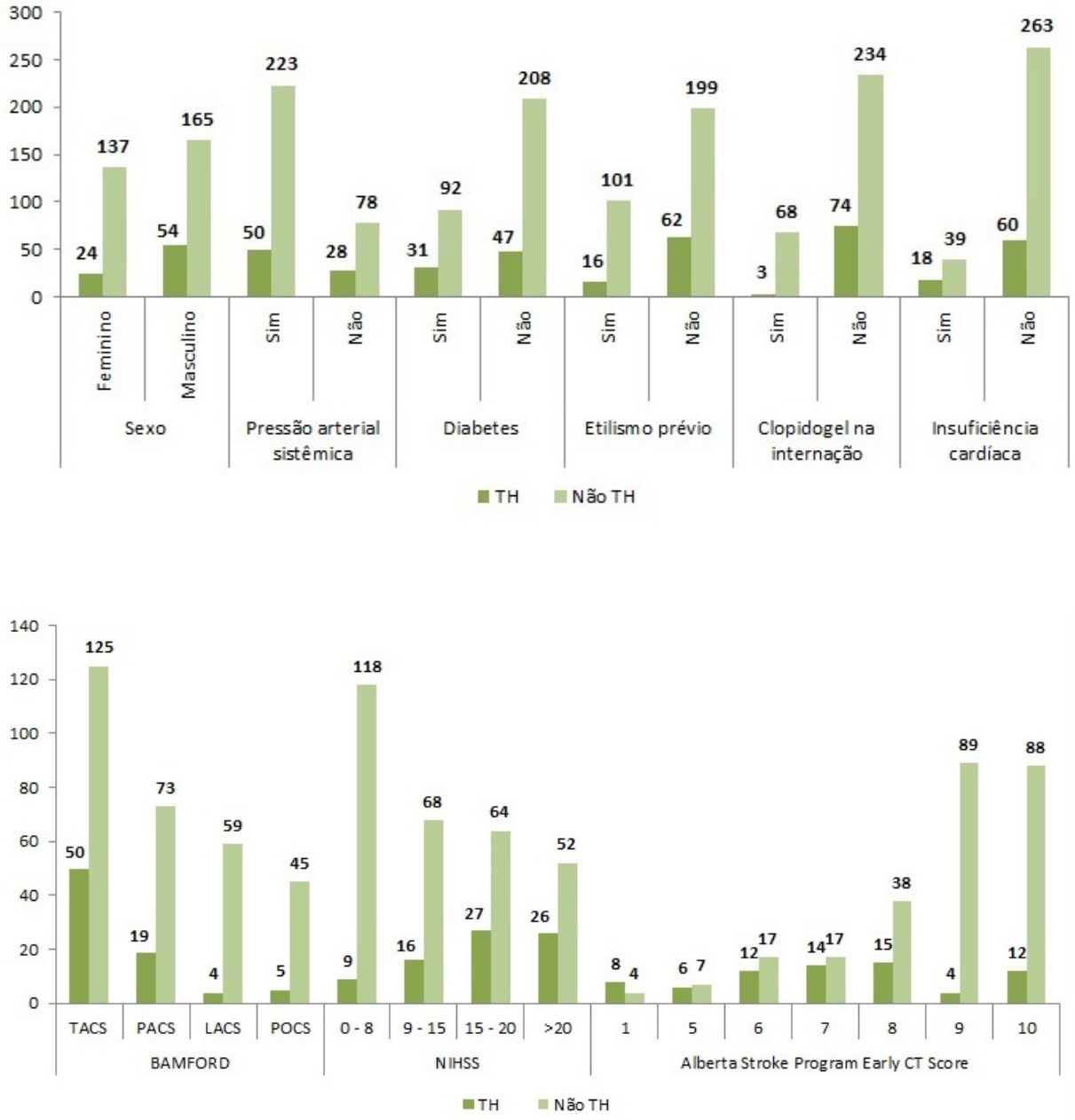
Na análise descritiva é possível visualizar uma prévia acerca do comportamento dos pacientes em relação ao desenvolvimento de transformação hemorrágica.

Nota-se pela tabela 12 que indivíduos do sexo masculino, sem hipertensão arterial sistêmica, com diabetes, insuciência cardíaca, não apresentaram etilismo prévio, não tomaram como medicação clopidogel na internação, apresentando valores acima de 20 na escala NIHSS, classicados entre os 4 maiores tamanhos de AVC e com Infarto de circulação anterior total aparentemente possuem maiores chances de desenvolverem transformação hemorrágica do que pacientes com pelo menos uma característica diferente.

Tabela 12 – Quantidade de indivíduos e percentual do total de cada variável que desenvolveram transformação hemorrágica segundo as variáveis do modelo final anterior a imputação.

Variáveis	Categoria	Transformação hemorrágica	
		Sim	Não
Sexo	Feminino	24 (6,32%)	137 (36,05%)
	Masculino	54 (14,21%)	165 (43,42 %)
Pressão Arterial Sistêmica	Sim	50 (13,19%)	223 (58,84%)
	Não	28 (7,39%)	78 (20,58%)
Diabetes	Sim	31 (8,21%)	92 (24,34%)
	Não	47 (12,43%)	208 (55,03%)
Etilismo prévio	Sim	16 (4,23%)	101 (26,72%)
	Não	62 (16,40%)	199 (52,65%
Clopidogel na internação	Sim	3 (0,79%)	68 (17,94%)
	Não	74 (19,53%)	234 (61,74%)
Insficiência cardíaca	Sim	18 (4,74%)	39 (10,26%)
	Não	60 (15,79%)	263 (69,21%)
BAMFORD	TACS	50 (13,16%)	125 (32,89%)
	PACS	19 (5,00%)	73 (19,21%)
	LACS	4 (1,05%)	59 (15,53%)
	POCS	5 (1,32%)	45 (11,84%)
NIHSS	0 - 8	9 (2,37%)	118 (31,05%)
	9 - 15	16 (2,37%)	68 (17,89%)
	15 - 20	27 (7,11%)	64 (16,84%)
	>20	26 (6,84%)	52 (13,68%)
Alberta Stroke Program Early CT Score	1	8 (2,42%)	4 (1,21%)
	5	6 (1,81%)	7 (2,11%)
	6	12 (3,63%)	17 (5,14%)
	7	14 (4,23%)	17 (5,14%)
	8	15 (4,53%)	38 (11,48%)
	9	4 (1,21%)	89 (26,89%)
	10	12 (3,64%)	88 (26,59%)

Figura 2 – Quantidade de indivíduos que desenvolveram transformação hemorrágica ou não segundo as variáveis do modelo final anterior a imputação.



Embora a média de glicemia em pacientes com e sem TH serem valores próximos, é notável que pacientes com TH apresentaram uma maior dispersão em relação aos pacientes que não tiveram TH. Na Tabela 13 tem-se medidas de resumo e dispersão da variável glicemia.

Tabela 13 – Medidas de resumo e dispersão referente aos valores de glicemia (mg/dL) segundo o desenvolvimento de transformação hemorrágica.

		Mín.	1 Q.	Med.	Média	3 Q.	Máx.	NA	Desvio Padrão
Tranformação hemorrágica	Sim	90,00	110,20	138,50	169,50	211,00	434,00	0	81,90
	Não	65,00	109,30	124,00	144,40	161,50	568,00	10	64,80

Os valores maiores para a variável glicemia na admissão são de pacientes que desenvolveram TH assim como também seu desvio padrão.

5.1.2 Análise Inferencial

O método da imputação múltipla foi aplicado para o tratamento de dados faltantes devido sua eficiência em relação aos demais. Foram utilizadas diferentes técnicas de imputação para cada procedimento uma vez que conjunto de dados possui nas variáveis explicativas valores contínuos e categóricos.

O processo de imputação múltipla foi realizado com o auxílio da ferramenta *SAS Studio*. Com a função PROC MI, com declaração FCS em que especifica uma imputação multivariada por métodos de especificação condicionais. Esse método realiza a imputação das observações faltantes mediante análise multivariada baseando-se em todas as informações observadas das variáveis especificadas pelo comando VAR. De acordo com o tipo de variável foi realizada um método de imputação, para as variáveis qualitativas nominais imputação por método de análise discriminante fazendo uso do comando DISCRIM, para as variáveis qualitativas ordinais imputação por regressão logística através do comando LOGISTIC, e para as variáveis contínuas imputação por métodos de regressão utilizando o comando REG. Foram realizadas cinco imputações especificadas no comando NIMPUTE.

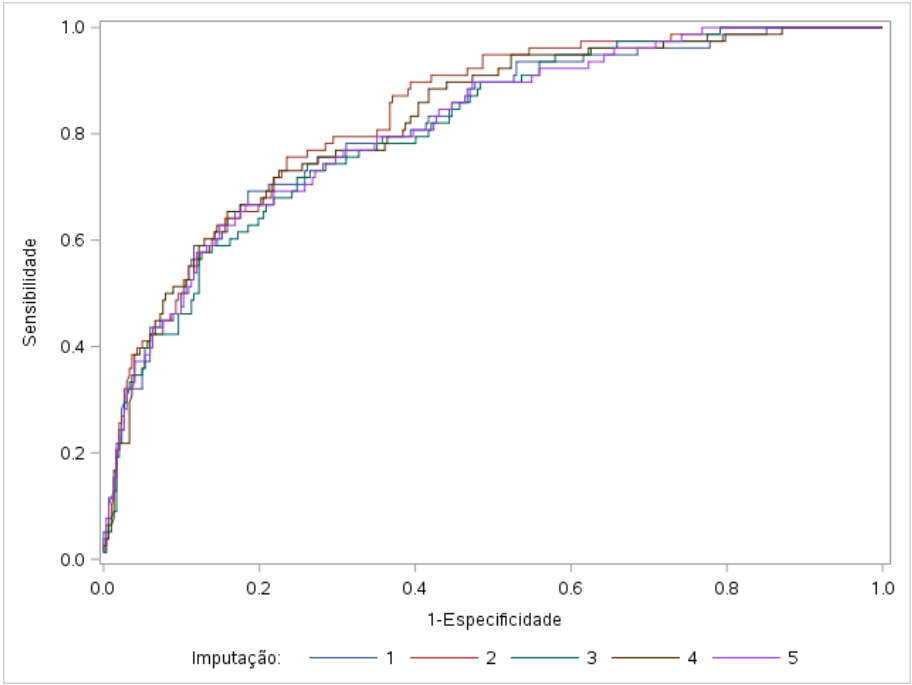
Combinação dos Resultados Posterior a obtenção dos 5 bancos de dados completos, foi realizado a seleção de variáveis com intuito de selecionar as variáveis mais importantes. Para isso foi utilizada a função PROC GLMSELECT, aplicando o comando BY para que esse procedimento fosse realizado para cada conjunto de dados gerados pela imputação, a técnica utilizada para seleção de variáveis foi a *stepwise* com critério de decisão BIC.

Analisando o modelo com as variáveis selecionadas, foi concluído que as variáveis que construíram o modelo final foram as descritas anteriormente na Tabela 10. Logo depois da seleção das variáveis, foi realizada a estruturação do modelo bayesiano, o qual foi construído

utilizando a função PROC GENMOD. Uma vez que a distribuição dos dados é desconhecida optou-se por fazer uso da distribuição a priori não informativa de Jeffrey. O modelo logístico foi o mais adequado devido a variável resposta ser categórica e a distribuição especificada para o modelo foi a Binomial com função de ligação *logit*. Foram geradas amostras através do MCMC e o critério de parada foi a distribuição atingir a estacionaridade.

Com intuito de validar o modelo final, os valores preditos dos modelos obtidos no PROC GENMOD foram calculados. Foi construído para cada conjunto de dados uma curva Roc, avaliando assim a acurácia do teste.

Figura 3 – Curva ROC para os modelos gerados para as 5 imputações.



Fonte: Feita no SAS

Tabela 14 – Área sob a curva para as 5 imputações.

Imputação	Área sobre a curva
1	0,81
2	0,83
3	0,80
4	0,82
4	0,81

Fonte: Feita no SAS Studio

De acordo com a Figura 3, a curva ROC das diferentes imputações apresentaram comportamento semelhante, a área sob a curva dos modelos indicaram valores satisfatórios, isto é, acima de 0,80, indicando que o modelo proposto apresentou um bom desempenho.

Nota-se que ponto de corte ideal é 0,2107. Utilizando este ponto de corte a sensibilidade é de 75,64% e 1-especificidade de 24,17%, ou seja, aproximadamente 75,64% de todas as amostras de pacientes com transformação hemorrágica seriam corretamente identificadas como tal, e 24,17% de todas as amostras de pacientes sem a transformação hemorrágica poderiam ser incorretamente identificadas como tendo desenvolvido TH.

Tabela 15 – Estimativas e seus respectivos erros padrões.

Variável	Categoria	Estimativas	Erro padrão
Intercepto		0,2406	0,2150
Glicemia		1,0055	0,0025
Sexo	Feminino		
	Masculino	3,4253	1,2050
Hipertensão arterial sistêmica	Não		
	Sim	0,4301	0,1571
Diabetes	Não		
	Sim	1,5098	0,5848
Etilismo prévio	Não		
	Sim	0,4563	0,4563
Clopidog na internação	Não		
	Sim	0,4563	0,1763
Insuficiência cardíaca	Não		
	Sim	2,0288	0,8743
BAMFORD	POCS		
	TACS	0,5371	0,3552
	PACS	0,9051	0,5921
	LACS	0,4983	0,4055
NIHSS	> 20		
	0 - 8	0,3593	0,2269
	9 - 14	0,4643	0,2299
	15 - 20	0,9569	0,3919
Alberta Stroke ProgramEarly CT Score	10		
	1	9,3874	6,9147
	5	4,6453	3,4733
	6	2,9980	1,6850
	7	3,1117	1,7228
	8	1,8261	0,9429
	9	0,4280	0,2390

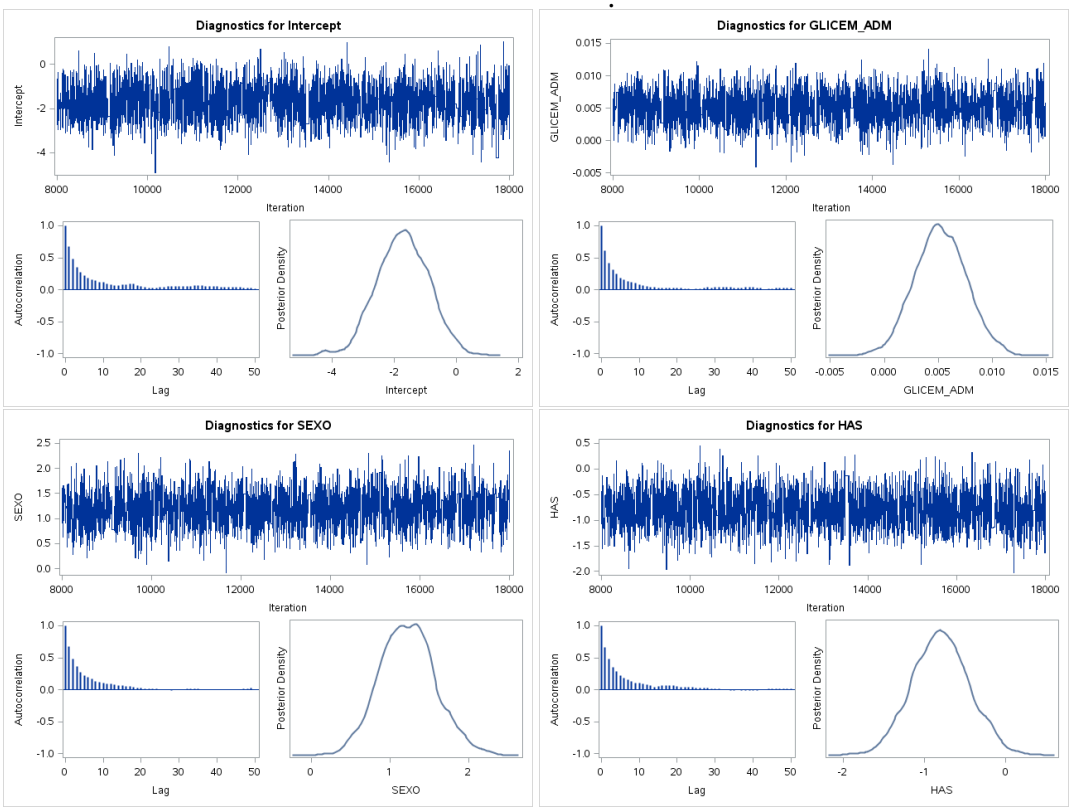
5.1.3 Análise de Diagnóstico

Do ponto de vista bayesino o diagnóstico do modelo ajustado é fundamentalmente avaliado na prática, de modo que esperasse retornar um alto poder preditivo. No entanto, uma das ferramentas utilizadas para a avaliação é a convergência da cadeia utilizada na simulação assim como a autocorrelação e distribuição *à posteriori* das estimativas do modelo.

Os resultados foram obtidos através da função GENMOD e em função dos resultados e gráficos obtidos pelo processo de imputação múltipla serem semelhantes em relação a configuração dos 5 conjuntos de dados, as análises de convergência a seram apresentadas nesta sessão são referentes ao conjunto de dados da imputação 1, os demais referentes as outras imputações podem ser vistas no apêndice A.

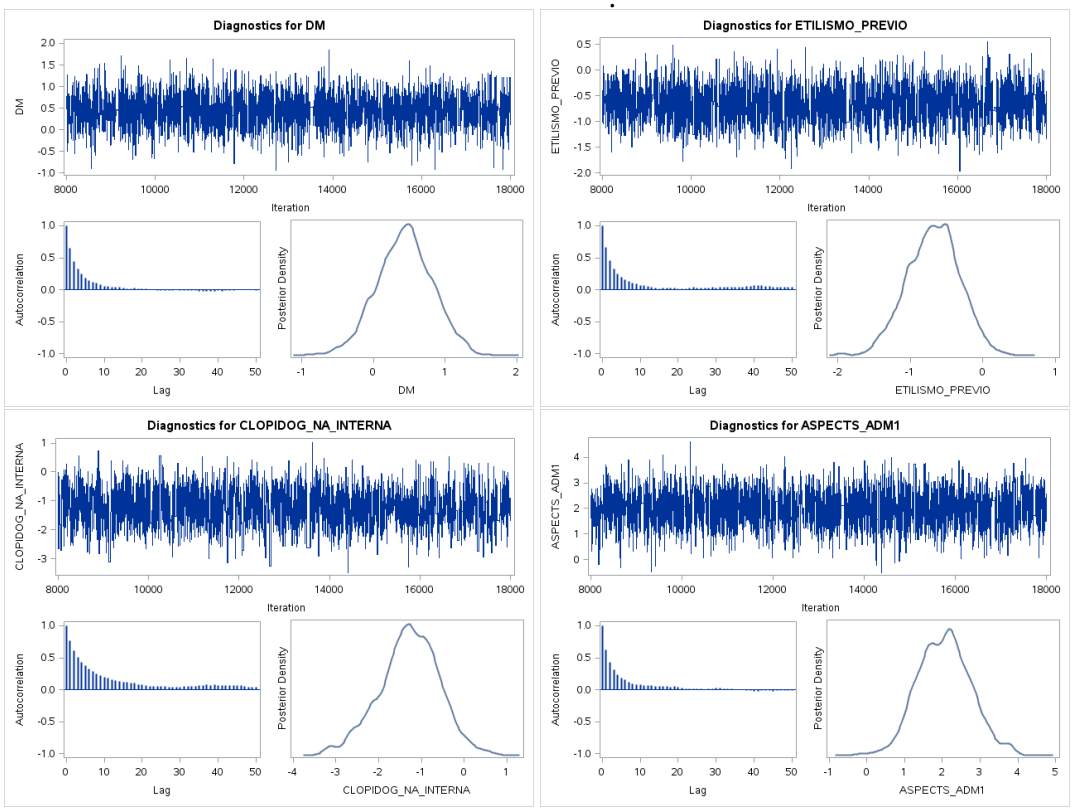
Nas Figuras 4 a 8 nota-se a cadeia de convergência para cada parâmetro estimado pelo modelo.

Figura 4 – Convergência dos parâmetros estimados pelo Modelo.



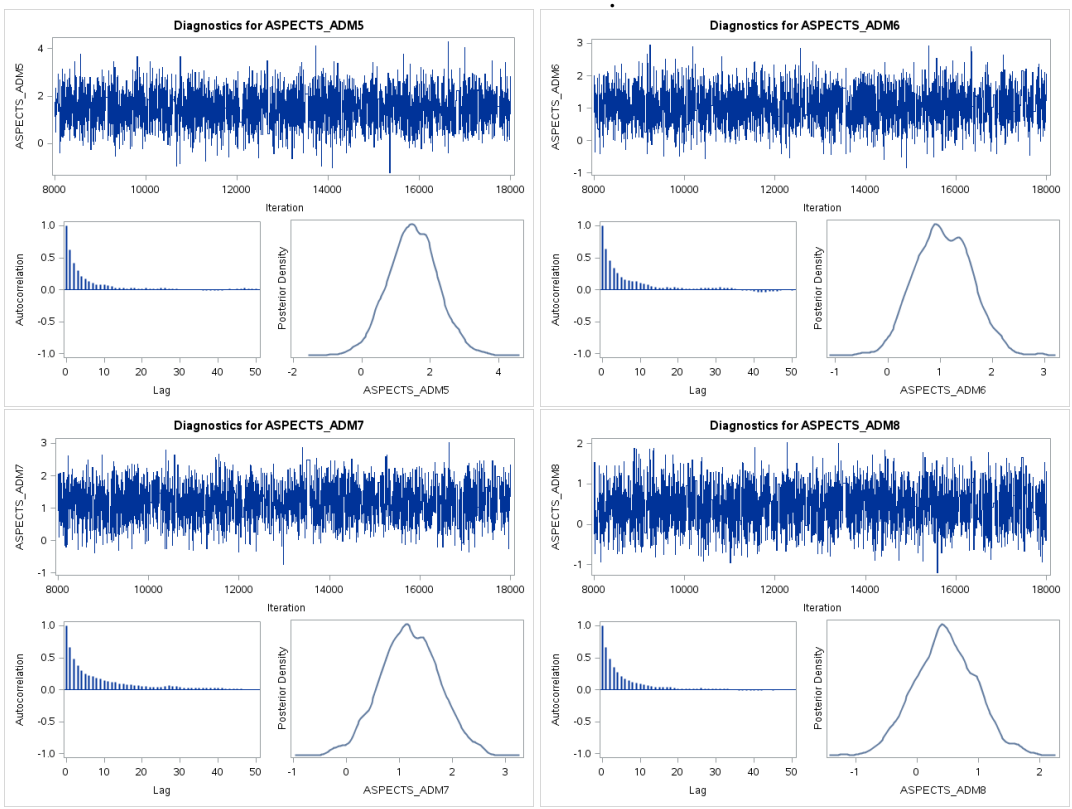
Fonte: Feita no SAS.

Figura 5 – Convergência dos parâmetros estimados pelo Modelo.



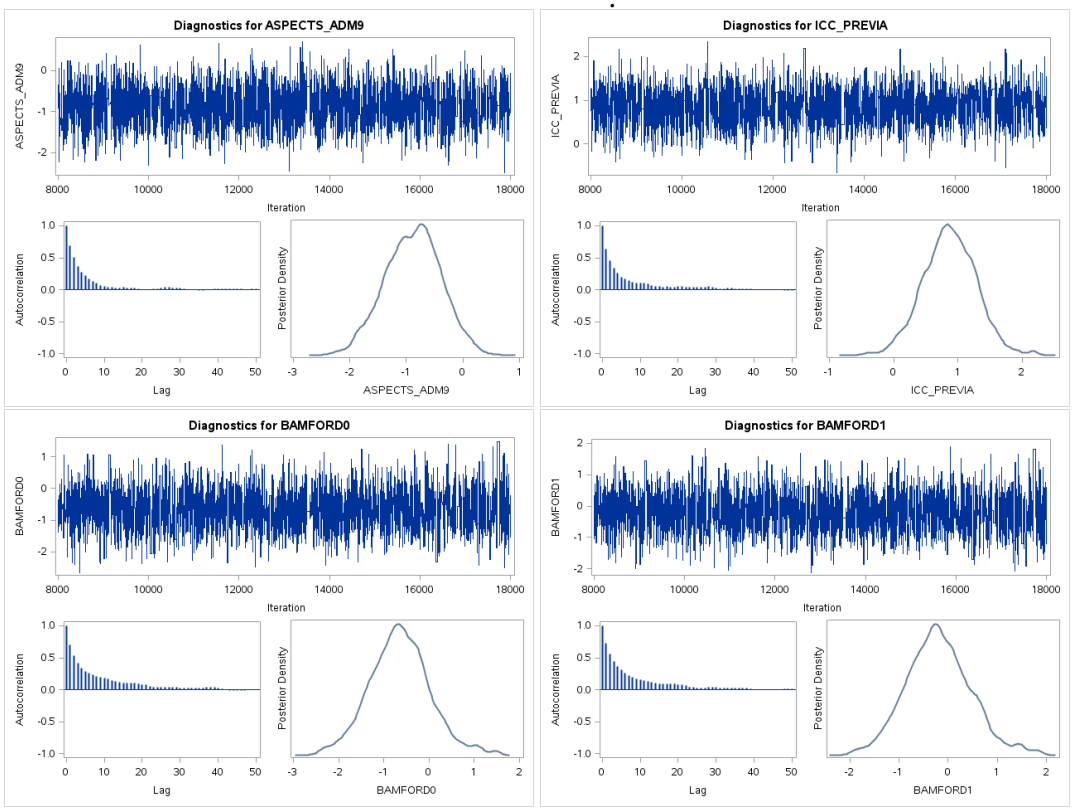
Fonte: Feita no SAS.

Figura 6 – Convergência dos parâmetros estimados pelo Modelo.



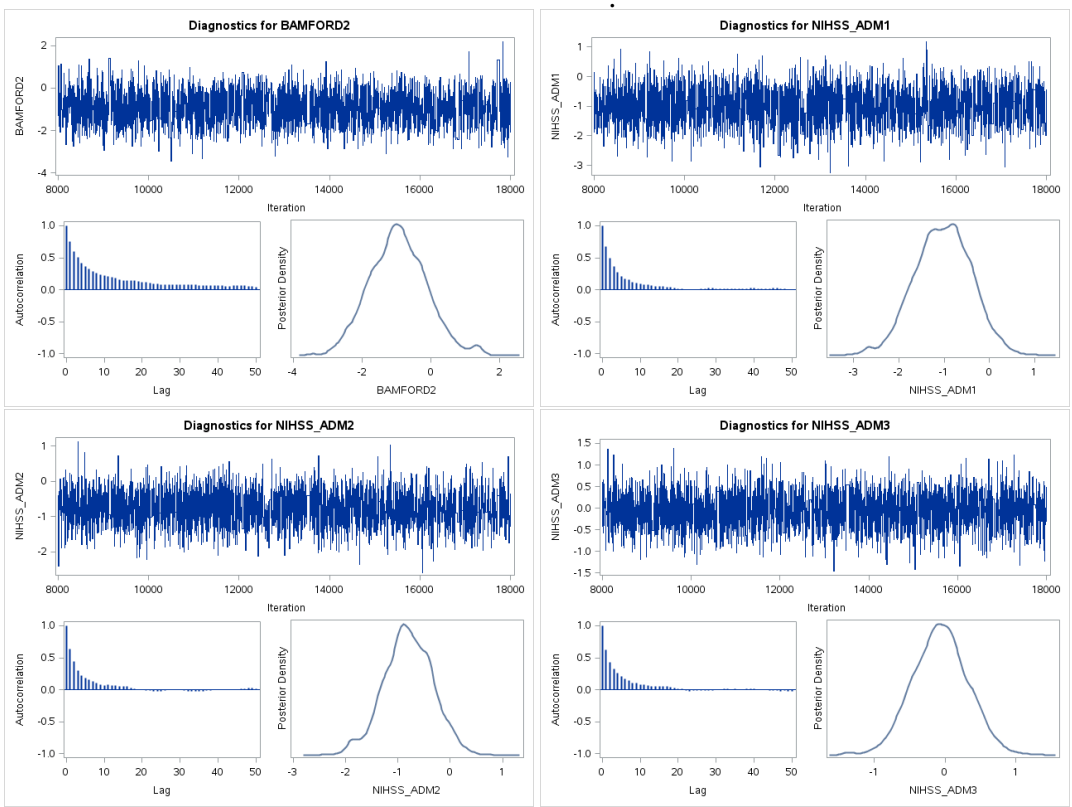
Fonte: Feita no SAS.

Figura 7 – Convergência dos parâmetros estimados pelo Modelo.



Fonte: Feita no SAS.

Figura 8 – Convergência dos parâmetros estimados pelo Modelo.



Fonte: Feita no SAS.

5.1.4 Análise de Sensibilidade

Algumas medidas de análises de diagnóstico usualmente aplicadas em estatística frequentista foram calculadas para verificar o comportamento dos dados e analisar as observações influentes para o modelo proposto.

Devido os resultados e gráficos obtidos pelo processo de imputação múltipla serem muito semelhantes em relação a configuração dos 5 conjuntos de dados, as análises de sensibilidade mostradas nesta sessão são referentes ao conjunto de dados da imputação 1, os demais referentes as outras imputações podem ser vistas no apêndice A.

Através da função PROC LOGISTIC com o preditor linear da função do PROC GENMOD foram obtidos os resultados a seguir. Para a avaliação das medidas de diagnóstico foram utilizados métodos gráficos, os valores foram obtidos pela função INFLUENCE e os gráficos referente a essas informações pela função IPLOT.

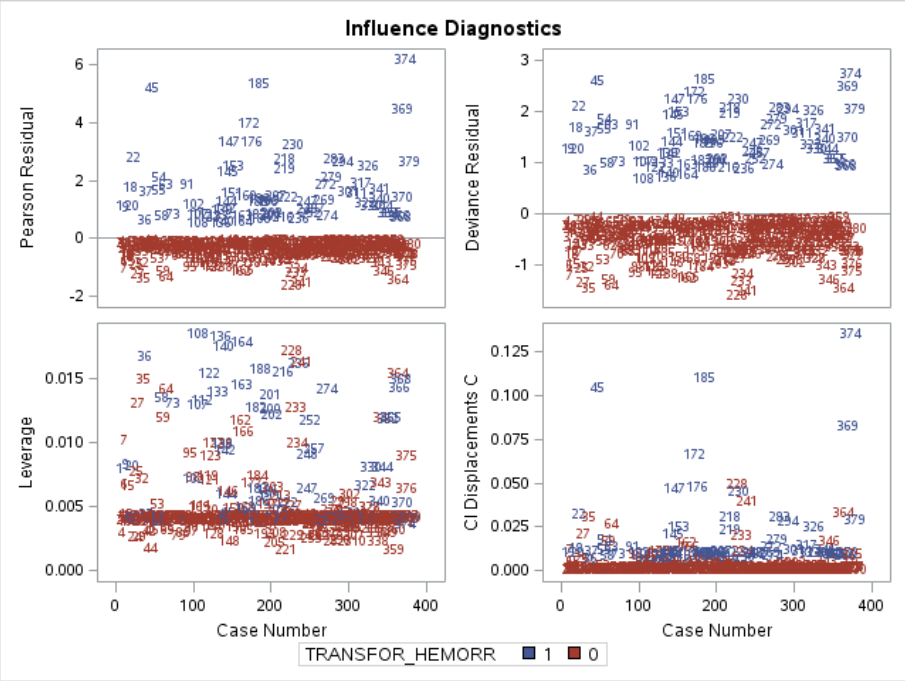
A Figura 9 aponta os resíduos de Pearson, função desvio, medidas de alavanca e a medida de deslocamento dos intervalos de confiança.

De acordo com Agranonik (2005) os desvios podem ser definidos como a distância entre os valores estimados e os valores observados, e são empregados para identificar observações que não estão sendo corretamente explicadas pelo modelo, podendo ser identificadas como pontos influentes. Os resíduos deviance avaliam se o modelo ajustado está adequado. Os pontos de alavanca vericam a distância estão os indivíduos em relação as demais obervações, de modo que apresentam características diferente das demais em relação as variáveis explicativas.

Agranonik (2005) defende a análise desses pontos devido haver a possibilidade desses pontos de alavanca implicarem em mudanças significativas nos valores estimados dos parâmetros.

De acordo com a Figura 9 há indícios de que as observações #45, #185 e #374 são possíveis pontos influentes, isto é, essas observações estão distante das demais e aparentemente não estão bem ajustadas pelo modelo proposto, tendo grande potencial de causar alterações nas estimativas dos parâmetros e suas respectivas interpretações caso estas observações sejam retiradas do conjunto de dados.

Figura 9 – Diagnóstico de observações influentes.

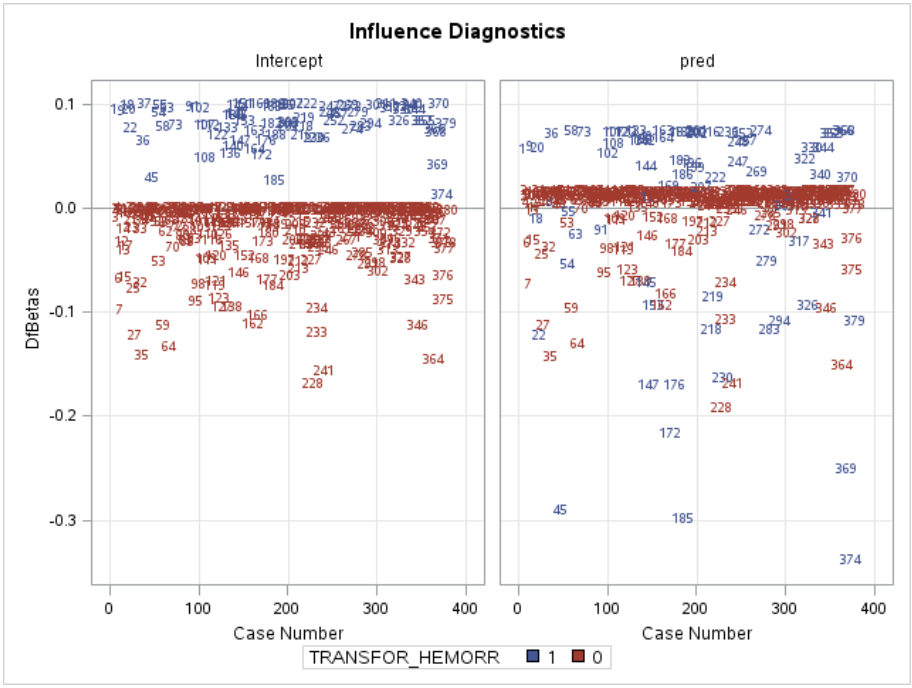


Fonte: Feita no SAS.

Segundo Barbiere (2012) as medidas DFBETAS permitem a realização da análise de diagnóstico para cada observação, medindo para cada coeficiente de regressão relacionado ao preditor o quanto alterado permanece quando esse indivíduo é excluído.

Pela Figura 10 nota-se as observações #45, #185 e #374 são medidas de influência de acordo com os DFBETAS, de modo que podem ser influentes em relação ao preditor linear, ou seja, não estão bem ajustadas, discordando entre os valores observados e preditos pelo modelo.

Figura 10 – Diagnóstico de observações influentes.



Fonte: Feita no SAS.

Através da análise das observações influentes é verificado que os valores das variáveis explicativas para esses indivíduos indicariam uma baixa probabilidade de desenvolver transformação hemorrágica, porém, esses indivíduos desenvolveram TH.

Tabela 16 – Observações influentes para o modelo segundo as covariáveis e predição.

Variável	#45	#185	#374
Sexo	Masculino	Masculino	Masculino
Hipertensão arterial sistêmica	Sim	Sim	Sim
Diabetes	Não	Não	Não
Etilismo Prévio	Não	Não	Não
Clopidogel na internação	Não	Não	Sim
Insuficiência cardíaca	Não	Não	Sim
NIHSS	0-8	>20	0-8
Alberta Stroke Program Early CT Score	9	9	9
Bamford	LACS	LACS	LACS
Transformação Hemorrágica	Sim	Sim	Sim
Probabilidade	0,028	0,026	0,018

Com intuito de verificar a influência dessas observações nas estimativas do modelo, foi retirada cada observação individualmente e depois as três em conjunto. Na Tabela 16 é apresentado as estimativas (*odds ratio*), o erro padrão assim como a taxa de variação das estimativas de cada parâmetro do modelo bayesiano logístico ao retirar tais observações. Caso todas as observações ocasionassem o mesmo impacto no modelo, então a variação para cada

estimativa (*odds ratio*) esperada seria de $\frac{1}{380} * 100 = 0,26\%$.

Retirando as observações individualmente e em conjunto, houve uma variação das estimativas (*odds ratio*) dos parâmetros em relação ao modelo completo, principalmente se retiradas as três observações em conjunto.

Analisando cada uma separadamente é observado que em todos os casos a variável contínua glicemia mostrou uma alteração bem pequena, menor do que o impacto esperado com a retirada de alguma observação.

Uma vez que houve alterações nas estimativas *odds ratio* pode-se dizer que há indícios de que as observações #45, #185 e #374 são influentes para o modelo, analisando as estimativas sem a transformação para o *odds ratio* é notado que apesar das estimativas mudarem houveram poucos casos em que o sinal das estimativas alteraram, apenas alguns casos na variável Bamford de classificação PACS e LACS (quando retirada as três observações conjuntamente) em que o sinal da estimativa foi modificado, e no caso das estimativas *odds ratio* essas observações alteraram na interpretação dos parâmetros em que anteriormente eram menos prováveis de ocorrer em relação a casela de referência, e posterior a retirada essa classificação passou a ser mais provável em comparação a referência.

Tabela 17 – Estimativas(*oddsratio*) dos parâmetros, erro padrão e variação(%) das respectivas estimativas para o modelo completo e sem as observações influentes.

Variável	Categoria	Completo	Sem #45	Sem #185	Sem #347	Sem #45, #185 e #347
Intercepto		0,24±0,22	0,14±0,12(-43,83)	0,12±0,11(-49,72)	0,21±0,19(-12,94)	0,12±0,25(-51,10)
Glicemia		1,01±0,002	1,01±0,002(-0,001)	1,006±0,002(0,004)	1,01±0,003(0,01)	1,01±0,01(0,03)
Sexo	Masculino	3,43±1,21	3,04±1,06(-11,15)	3,08±1,06(-10,24)	3,21±1,133(-6,19)	2,86±3,00(-16,57)
Hipertensão Arterial	Sim	0,43±0,16	0,46±0,16(5,87)	0,47±0,17(9,20)	0,42±0,16(-1,75)	0,38±-0,37(-11,92)
Diabetes	Sim	1,51±0,59	1,59±0,62(5,65)	1,572±0,60(4,10)	1,59±0,62(5,25)	1,702±0,91(12,75)
Estilismo Prévio	Sim	0,46±0,17	0,53±0,20(14,98)	0,51±0,19(12,11)	0,48±0,18(5,70)	0,53± -0,33(15,71)
Clopidogel na internação	Sim	0,25±0,18	0,32±0,23(28,37)	0,296±0,21(18,14)	0,18±0,15(-28,38)	0,21± -0,33(-15,45)
Alberta Stroke ProgramEarly CT Score	1	9,39±6,92	8,80±6,62(-6,31)	8,88±6,65(-5,41)	8,69±6,43(-7,45)	7,95±16,49(-15,29)
	5	4,65±3,47	4,39±3,25(-5,58)	4,63±3,42(-0,33)	4,36±3,25(-6,15)	3,80±5,07(-18,26)
	6	2,10±1,69	2,84±1,57(-5,27)	2,90±1,60(-3,42)	2,86±1,61(-4,55)	2,57±2,44(-14,06)
	7	3,11±1,72	3,06±1,67(-1,83)	3,14±1,71(0,81)	2,85±1,58(-8,30)	2,61±2,51(-16,08)
	8	1,83±0,94	1,63±0,82(-10,60)	1,68±0,85(-8,22)	1,69±0,88(-7,42)	1,52±0,64(-16,70)
	9	0,43±0,24	0,35±0,20(-18,86)	0,39±0,21(-9,90)	0,40±0,24(-7,20)	0,28± -0,36(-35,58)
Insuficiência Cardíaca	Não	2,03±0,87	2,33±0,98(14,63)	2,26±0,95(11,16)	1,83±0,80(-9,71)	1,88±1,19(-7,22)
Bamford	TACS	0,54±0,36	0,70±0,49(27,00)	0,72±0,50(33,08)	0,66±0,47(23,51)	1,39±0,46(159,33)
	PACS	0,91±0,59	1,28±0,90(41,19)	1,18±0,80(30,292)	1,14±0,80(26,38)	2,46±2,21(171,74)
	LACS	0,50±0,41	0,68±0,59(36,63)	0,56±0,47(12,86)	0,69±0,59(37,41)	1,48±0,58(197,36)
NIHSS	0-8	0,36±0,23	0,25±0,165(-29,74)	0,35±0,23(-2,83)	0,32±0,21(-9,96)	0,30± -0,36(-17,04)
	9-14	0,46±0,23	0,433±0,22(-6,74)	0,48±0,24(3,77)	0,465±0,23(0,20)	0,50± -0,35(8,56)
	15-20	0,96±0,39	0,874±0,352(-8,65)	0,93±0,38(-2,726)	0,95± 0,39(-0,57)	0,98± -0,02(1,96)

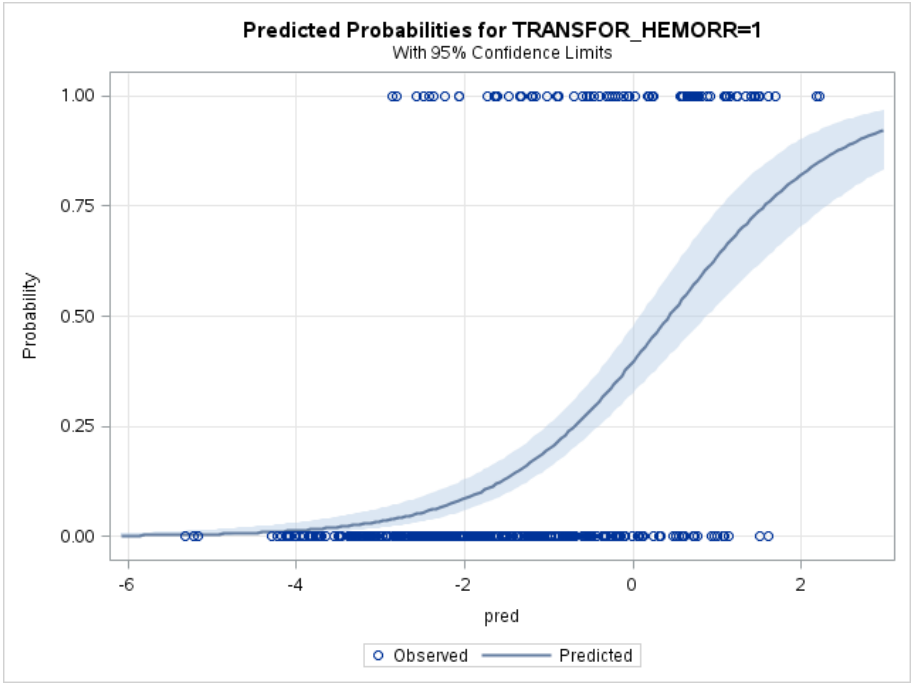
Através da Tabela 17 nota-se que retirando individualmente cada observação o impacto na avaliação do modelo não é grande, no entanto é perceptível um aumento leve do desempenho do modelo. Por outro lado, caso se retire essas observações influentes de forma conjunta é vista uma redução no desempenho do modelo, assim como mostra o valor da área sob a curva.

Tabela 18 – Área sob a curva para as 5 imputações do modelo completo e retirando as observações influentes.

Imputação	Completo	Sem #45	Sem #185	Sem #347	Sem #45, #185 e #347
1	0,8140	0,8210	0,8250	0,8210	0,7870
2	0,8350	0,8350	0,8310	0,8340	0,8110
3	0,8070	0,8170	0,8190	0,8140	0,7830
4	0,8220	0,8280	0,8260	0,8270	0,7920
5	0,8120	0,8190	0,8210	0,8160	0,7780

Pela Figura 11 nota-se a relação entre os valores preditos segundo a probabilidade de ocorrência do paciente desenvolver transformação hemorrágica em relação ao não desenvolvimento. Uma vez que os valores preditos aumentam maior é a probabilidade do paciente desenvolver TH.

Figura 11 – Valores observados e ajustados de desenvolvimento de transformação hemorrágica.



Fonte: Feita no SAS.

5.2 Análise da Base de Dados Completo

Nesta seção serão apresentados os resultados da análise feita na base de dados com as informações mais completas, ou seja, com menos *missings* obtidas em um segundo momento da pesquisa, em tal base foi ajustada um modelo de regressão logística escolhido utilizando o método *stepwise* para a seleção das variáveis.

5.2.1 Análise Descritiva

A Tabela 19 mostra as variáveis da base mais completa e a quantidade de informações observadas e faltantes.

Tabela 19 – Quantidade de informações faltantes e observadas segundo as variáveis do conjunto de dados.

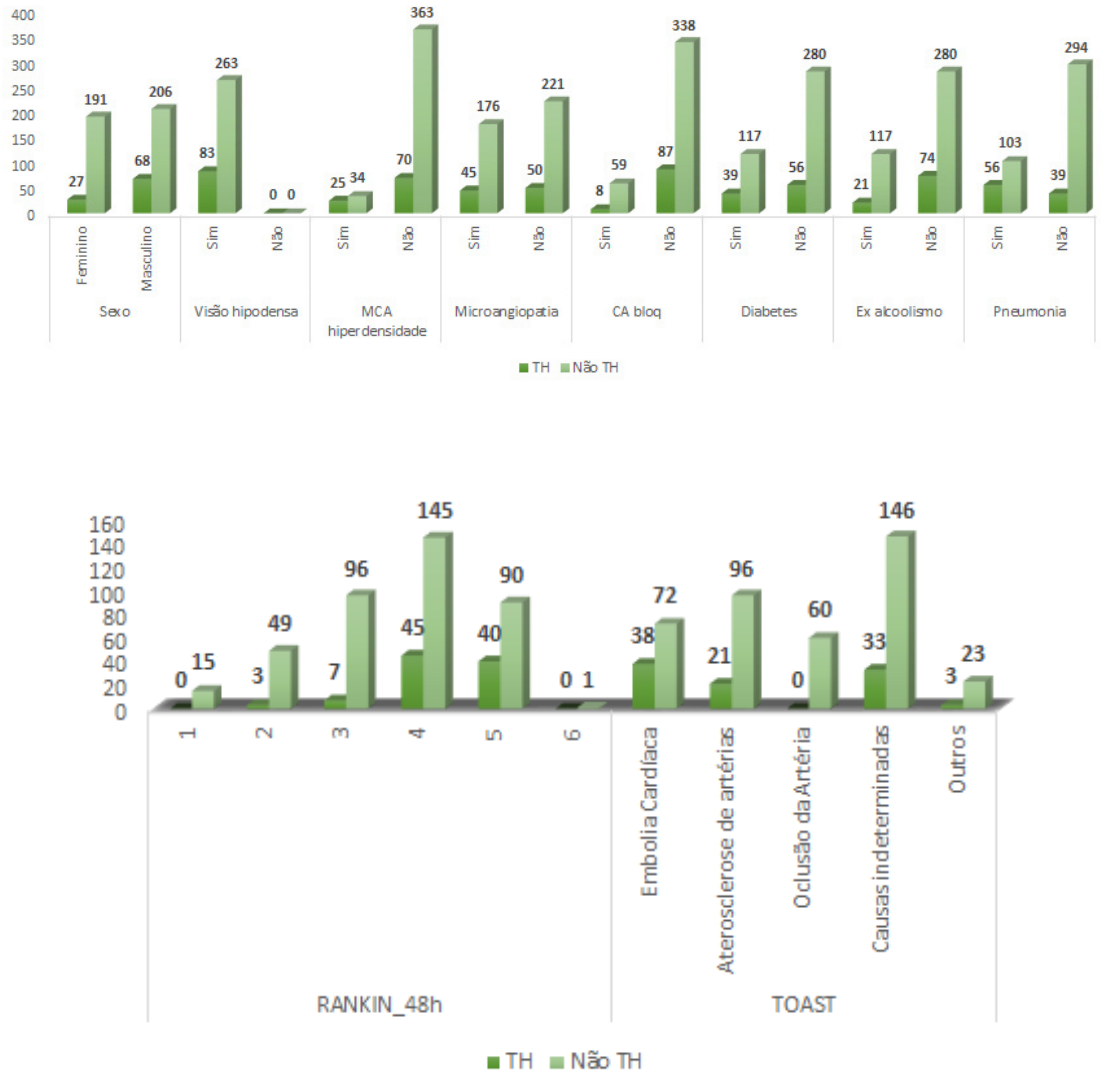
Variáveis	Observadas	Faltantes
Tempo_início_adm	491	1
Síncope_adm	491	1
Apreensão_adm	491	1
RANKIN_48h	491	1
TAP_adm	491	1
WBC_adm	491	1
Plaquetas_adm	491	1
ASPECTS_adm	491	1
CA_bloq_past	491	1
Alcoolatra	491	1
Doença cardíaca	491	1
TIA_past	491	1
Colesterol_total	491	1
HDL_colesterol	491	1
Pneumonia	491	1
Apreensão	491	1
ASA_hospital	491	1
Estatina_hospital	491	1
VTE_profilaxia	491	1
VTE_time_to_use	491	1
TOAST	491	1
BAMFORD	491	1
Idade	490	2
Dor_cabeça_adm	490	2
Glicemia_adm	490	2
Delírio	490	2
Microangiopatia	439	3
Ieca_bra_past	439	3
Fribilição_atrial	439	3
Outras_complicações	439	3
LDL_colesterol	438	4
Sexo	437	5
FC_adm	437	5
Insuficiencia_cardíaca	437	5
Triglicerídeos	437	5
ASPECTS_final	437	5
PAD_adm	436	6
Dor_cabeça	436	6
Diuretico_passado	435	7

Na Tabela 20 tem-se os pacientes que desenvolveram transformação hemorrágica e suas frequências segundo as variáveis do modelo final.

Tabela 20 – Quantidade de indivíduos e percentual do total de cada variável que desenvolveram transformação hemorrágica segundo as variáveis do modelo final da base mais completa.

Variáveis	Categoria	Transformação Hemorrágica	
		Sim	Não
Sexo	Feminino	27 (5,48%)	191(38,82%)
	Masculino	68 (13,82%)	206 (41,86%)
Visão hipodensa	Sim	83 (23,98%)	263 (76,01%)
	Não	0 (0,00%)	0 (0,00%)
MCA hiperdensidade	Sim	25 (5,08%)	34 (6,91%)
	Não	70 (14,22%)	363 (73,78%)
Microangiopatia	Sim	45(9,14%)	176 (35,77%)
	Não	50 (10,16%)	221 (44,91%)
CA bloq	Sim	8 (1,62%)	59 (11,99%)
	Não	87 (17,68%)	338 (68,69%)
Diabetes	Sim	39 (7,92%)	117 (23,78%)
	Não	56 (11,38%)	280 (56,91%)
Etilismo prévio	Sim	21 (4,26%)	117 (23,78%)
	Não	74 (15,04%)	280 (56,91%)
Pneumonia	Sim	56 (11,38%)	103 (20,93%)
	Não	39 (7,92%)	294 (59,75%)
TOAST	Embolia Cardíaca	38 (7,77%)	72 (14,63%)
	Aterosclerose de grandes artérias	21 (4,26%)	96 (19,51%)
	Oclusão da Pequena Artéria	0 (0,00%)	60 (12,19%)
	Causas indeterminadas	33 (3,70%)	146 (29,67%)
	Outros	3 (0,06%)	23 (4,67%)
RANKIN_48h	1	0 (0,00%)	15 (3,04%)
	2	3 (0,60%)	49 (9,95%)
	3	7 (1,42%)	96 (19,51%)
	4	45 (9,14%)	145 (29,47%)
	5	40 (8,13%)	90 (18,29%)
	6	0 (0,00%)	1 (0,20%)

Figura 12 – Quantidade de indivíduos que desenvolveram transformação hemorrágica ou não segundo as variáveis do modelo final da base mais completa.



Observa-se que para as variáveis que foram significativas para os dois modelos, isto é, na base com imputações e na base mais completa, as variáveis possuem valores similares na análise descritiva.

As informações referente a variável contínua glicemia e a variável numérica idade estão dispostas na Tabela 21.

Na próxima seção serão realizadas análises inferenciais através do modelo logístico bayesiano.

Tabela 21 – Medidas de resumo e dispersão referente aos valores de glicemia (mg/dL) e da idade segundo o desenvolvimento de transformação hemorrágica.

	TH	Mín.	1 Q.	Med.	Média	3 Q.	Máx.	Desvio Padrão
Glicemia	Sim	65	106	125,5	149,8592	169,75	568	69,33967
	Não	65	106	125,5	149,751	168,75	568	69,33546
Idade	Sim	15	56	67	64,93075	76	99	14,92605
	Não	15	56	67	64,94501	76	99	14,92276

5.2.2 Análise Inferencial

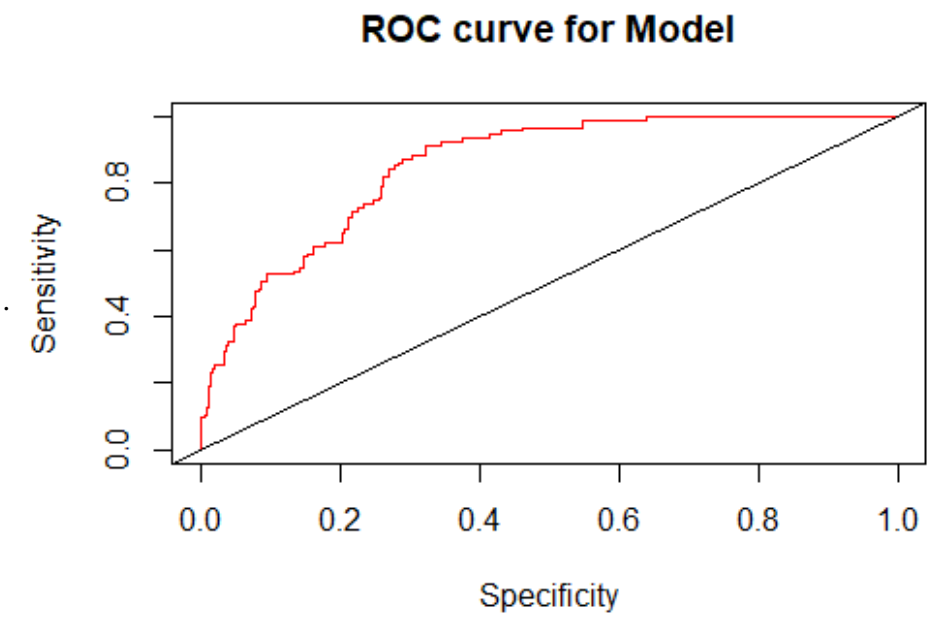
Nesta seção observa-se os resultados da regressão logística aplicada na base de dados mais completa, em que não aplicou-se imputações, foi ajustado uma regressão logística e utilizado o método *stepwise* para a seleção das variáveis do modelo mais apropriado. A Tabela 22 apresenta as estimativas dos parâmetros ajustados.

Tabela 22 – Estimativas dos parâmetros, erros padrões, estatística Wald, *odds ratio*, e intervalo de credibilidade das respectivas estimativas para o modelo final da base de dados completo.

Variáveis	Estimativas(β)	Erro padrão	Wald	Prob	exp(β)	IC (2,5% ; 97,5%) exp(β)	
Intercepto	-4,686294	1,14132	-4,106	4,03E-05	0,009221	0,000985	0,08635
Idade	-0,029834	0,010642	-2,804	0,00505	0,970607	0,950573	0,991063
Sexo	1,564473	0,318301	4,915	8,87E-07	4,780155	2,561558	8,920313
RANKIN_48h	0,625065	0,178727	3,497	0,00047	1,868367	1,316223	2,652131
Glicemia	0,00569	0,002227	2,555	0,01061	1,005706	1,001326	1,010104
Hipodensidade adm	1,094584	0,378035	2,895	0,00379	2,987939	1,424255	6,268385
Hiperdensidade adm	0,986734	0,365094	2,703	0,00688	2,682459	1,311488	5,486581
Microangiopatia	0,838395	0,296237	2,83	0,00465	2,312652	1,294054	4,133025
Ca_bloq_past	-1,217969	0,485396	-2,509	0,0121	0,29583	0,114254	0,765972
Diabetes	0,729937	0,357744	2,04	0,04131	2,07495	1,029189	4,18331
Etilismo prévio	-0,764678	0,339197	-2,254	0,02417	0,465484	0,239431	0,90496
Pneumonia	0,936834	0,301611	3,106	0,0019	2,551889	1,412961	4,608858
TOAST	-0,352567	0,108016	-3,264	0,0011	0,702881	0,568772	0,868612

Com intuito de validar o modelo final, os valores preditos dos modelo obtido, foi construído uma curva ROC, avaliando assim a acurácia. A Figura 13 mostra a curva ROC do modelo final do banco de dados mais completo.

Figura 13 – Curva ROC para o modelo final.

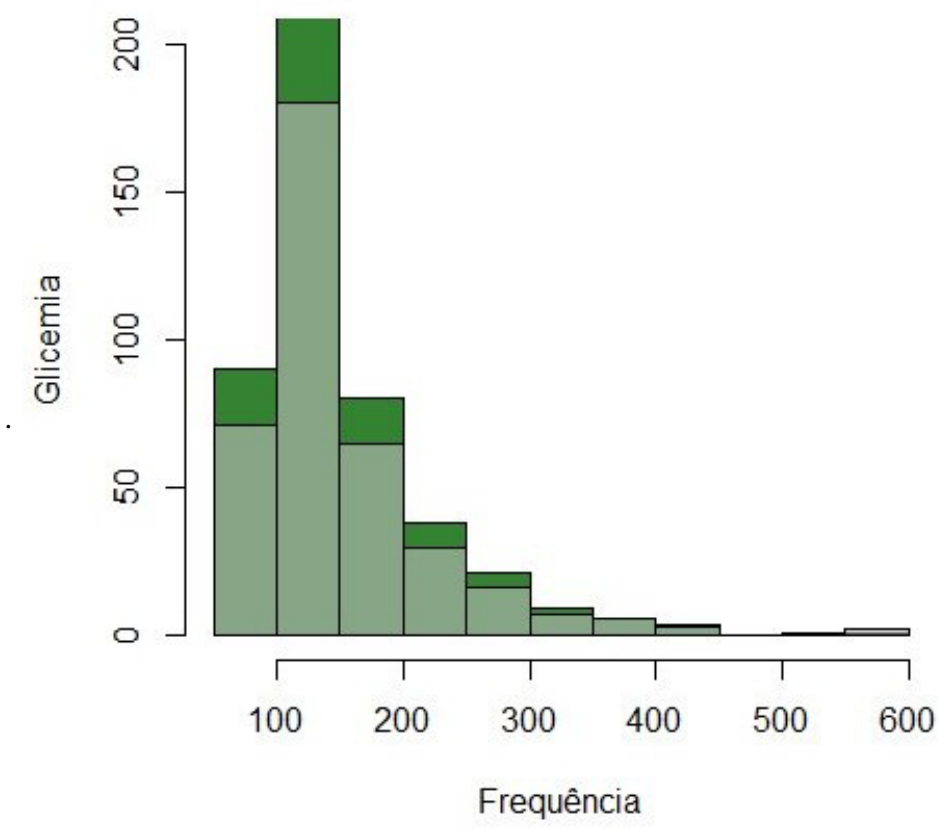


A área sob a curva do modelo é de 0,82 de modo que aparenta ser um valor satisfatório, uma vez que está acima de 0,80, indicando que o modelo proposto apresentou um bom desempenho.

5.3 Comparação de Resultados

Analisando os resultados, nota-se que as variáveis Glicemia, Sexo e Estilismo prévio foram as variáveis significativas nos dois modelos finais, sendo eles ajustados com a base de dados com dados faltantes e o outro com base de dados mais completa. Na Figura 14 tem-se os histogramas da variável glicemia sobrepostos em que o verde são os valores da base mais completa e o cinza os valores da base de dados imputada.

Figura 14 – Histograma da variável Glicemia.



Nota-se que a maior frequência está concentrada entre 100 e 150 e que os valores se comportame de maneira semelhante no conjunto de dados coletados nos dois momentos distintos, primeiramente com muitos dados faltantes e posteriormente com uma base mais completa.

Na Tabela 23 tem-se a frequência e a proporção dos valores observados na base de dados mais completa comparado aos valores dos dados imputados para as variáveis categoricas sexo e estilismo prévio, ambas significativas nos dois modelos finais.

Tabela 23 – Frequência e proporção dos valores observados na base de dados mais completa e valores dos dados imputados da variável Diabetes.

Diabetes - Base Mais Completa					
	Não	Sim	Total	fr(nd)	fr(d)
Quantidade	336	156	492	0,682926829	0,317073171
Diabetes - Base Imputada					
	Não	Sim	Total	fr(nd)	fr(d)
Quantidade	257	123	380	0,676315789	0,323684211

Observando a Tabela 23 notamos que a proporção dos pacientes que não tinham diabetes representa 32,3% e que não tinham diabetes aproximadamente 67,7% do total para os valores imputados enquanto que na base de dados mais completa é 31,7% para o pacientes com diabetes e 68,3% para pacientes diabético, isto é, valores bem próximos nas duas bases.

Tabela 24 – Frequência e proporção dos valores observados na base de dados mais completa e valores dos dados imputados da variável Estilismo Prévio.

Etilismo prévio - Base Mais Completa					
	Não etilismo prévio	Etilismo prévio	Total	fr(ne)	fr(e)
Quantidade	354	138	492	0,719512	0,280488
Etilismo prévio - Base Imputada					
	Não etilismo prévio	Etilismo prévio	Total	fr(ne)	fr(e)
Quantidade	263	117	380	0,692105	0,307895

Através da Tabela 24 observa-se que a proporção dos pacientes com estilismo prévio representa 30,7% e com estilismo prévio 69,3 % do total para os valores imputados enquanto que na base de dados mais completa é 28% para pacientes com estilismo prévio e aproximadamente 72% para estilismo prévio, navamente tem-se valores bem próximos nas duas bases.

Outra maneira de comparar o desempenho do tratamento de imputação múltipla aplicada é por meio do percentual de acerto, ou seja, a percentagem de acerto do valor imputado em relação aos valores da base mais completa.

Na Tabela 25 observamos o percentual de acerto da imputação para as variáveis selecionadas para o modelo final.

Tabela 25 – Percentual de acerto dos valores imputados em relação aos valores da base mais completa para as variáveis do modelo final.

Variável	Quantidade imputada	Percentual de acerto
Glicemia	10	40%
Hipertensão	1	100%
Diabetes	2	50%
Etilismo prévio	2	100%
Clopidogrel	1	100%
Hipodensidade	8	62%
TOAST	4	50%

Nota-se pela Tabela 25 que a taxa de acerto para os valores imputados foram altas, um vez que para todas as variáveis o percentual de acerto foi maior ou igual a 50% do total para a variáveis categóricas e 40% para a variável continua Glicemia.

6 CONSIDERAÇÕES FINAIS

Este trabalho teve como finalidade descrever procedimentos para análise de banco de dados com informações faltantes com uma aplicação com dados reais comparando as estimativas dos parâmetros com o modelo ajustado com a base de dados completa.

O método de imputação utilizado foi imputação múltipla, a qual possibilitou a estruturação das análises sem ser necessário a exclusão das observações faltantes, de modo que possibilitou a construção de um modelo preditivo e realização de inferências satisfatórias com todos os indivíduos do conjunto de dados. Inicialmente tentou-se a comparação da técnica de imputação por algoritmo EM mas há a limitação de não ser possível uma vez que apenas é possível a aplicação quando não se tem variáveis categóricas.

Com a utilização de metodologias bayesiana foi construído um modelo preditivo que indica características que descreva a probabilidade do paciente que teve AVC isquêmico desenvolver transformação hemorrágica. Para validação do modelo utilizou-se a curva Roc que indicou um bom desempenho do modelo.

A aplicação foi realizada com base no trabalho de Andrade (2017), em que resultaram em algumas conclusões semelhantes. Os homens tiveram maiores chances de desenvolver TH do que mulheres, assim como pessoas classificadas com NIHSS acima de 20 tiveram chances maiores do que pessoas com essa classificação abaixo de 20. Nota-se também que a média das imputações para a variável glicemia, significativa nos dois modelos finais, foi próxima a médias dos valores observados na base mais completa para os mesmos individuos, o que pode ser um forte indicativo de que o método de imputação utilizado trás bons resultados, isto é, próximos da realidade, outra medida também avaliada foi a taxa percentual de acerto das variáveis imputadas em relação a base mais completa em que podemos observar que os resultados foram bem próximos dos valores observados na base mais completa indicando um bom desempenho do método de tratamento de dados faltantes aplicado.

Como nesta abordagem tratamos apenas de dados omissos quando estas ocorrem nas variáveis explicativas então utilizou-se somente o desfecho transformação hemorrágica, no qual todas as informações estavam preenchidas, o que possibilita pesquisas posteriores com o mesmo conjunto de dados com foco quando os desfechos, assim como as variáveis explicativas, possuem informações faltantes.

REFERÊNCIAS

AGRANONIK, M. **Técnicas de diagnóstico aplicadas ao modelo de regressão logística.** 2005**,** 66 f. Monografia (Bacharel em Estatística) - Universidade Federal do Rio Grande do Sul, Porto Alegre, 2005.

AGRESTI, A.; KATERI, M. **Categorical data analysis**. [*S. l.*]: Springer, 2011.

ALBIERI, S. **A ausência de resposta em pesquisas: uma aplicação de métodos de imputação.** 1989. 138 p. Tese (Doutorado em Estatística) — Instituto de Matemática Pura e Aplicada, Rio de Janeiro, 1989.

ALLISON, P. D. **Missing data**. [*S. l.*]: Sage publications, 2001.

ANDRADE, J. B. C. **Transformação hemorrágica espontânea: há como prever?.** 2017**,** 106 f. TCC (Curso de Medicina) - Universidade Estadual do Ceará, Fortaleza, 2017.

ASSIS, D. F. **Modelo bayesiano aplicado ao tratamento de dados faltantes.** 2017**,** 78 f. TCC (Curso de Estatística) - Centro de Ciências, Universidade Federal do Ceará, Fortaleza, 2017.

AZUR, M. J.; STUART, E. A.; FRANGAKIS, C.; LEAF, P. J. Multiple imputation by chained equations: what is it and how does it work? **International journal of methods in psychiatric research**, Wiley Online Library, n. 1, p. 40–49, 2011.

BARBIERI, N. B. **Estimação robusta para o modelo de regressão logística** 2012**,** 59 f. TCC (Curso de Estatística) - Universidade Federal do Rio Grande do Sul, Porto Alegre, 2012.

BOX, G. E.; TIAO, G. C. **Bayesian inference in statistical analysis**. [*S. l.*]: John Wiley & Sons, 2011.

BUUREN, S. V.; BOSHUIZEN, H. C.; KNOOK, D. L. Multiple imputation of missing blood pressure covariates in survival analysis. **Statistics in medicine**, Wiley Online Library, n. 6, p. 681–694, 1999.

BUUREN, S. V.; OUDSHOORN, K. **Flexible multivariate imputation by MICE**. [*S. l.*]: Leiden: TNO, 1999.

CASELLA, G.; BERGER, R. L. **Statistical inference**. [*S. l.*]: Duxbury Pacific Grove, CA, 2002.

CHEN, F. Missing no more: Using the mcmc procedure to model missing data. **Proceedings of the SAS Global Forum 2013 Conference. Cary, NC: SAS Institute**, Citeseer, n. 1, p. 1–23, 2013.

COLANTONIO, A.; PIETRO, R. D.; OCELLO, A.; VERDE, N. V. Abba: Adaptive bicluster-based approach to impute missing values in binary matrices. **Proceedings of the 2010 ACM Symposium on Applied Computing**, ACM, n. 1, p. 1026–1033, 2010.

DURRANT, G. B. *et al.* Imputation methods for handling item-nonresponse in the social sciences: a methodological review. **ESRC National Centre for Research Methods and Southampton Statistical Sciences Research Institute**, NCRM Methods Review Papers NCRM/002, p. 1–36, 2005.

EHLERS, R. S. Inferência bayesiana. **Notas de Aula - Departamento de Matemática Aplicada e Estatıstica, ICMC-USP**, p. 64, 2011.

ENGELS, J. M.; DIEHR, P. Imputation of missing longitudinal data: a comparison of methods. **Journal of clinical epidemiology**, Elsevier, n. 10, p. 968–976, 2003.

FARHANGFAR, A.; KURGAN, L. A.; PEDRYCZ, W. A novel framework for imputation of missing values in databases. **IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans**, IEEE, n. 5, p. 692–709, 2007.

GILKS, W. R.; RICHARDSON, S.; SPIEGELHALTER, D. **Markov chain Monte Carlo in practice**. [*S. l.*]: Chapman and Hall/CRC, 1995.

GRAHAM, J. W. *et al.* Analysis with missing data in prevention research. **The science of prevention: Methodological advances from alcohol and substance abuse research**, American Psychological Association, p. 325–366, 1997.

INSTITUTE, S. **SAS/STAT user's guide: version 6**. [*S. l.*]: Sas Inst, 1990.

LITTLE, R. J.; RUBIN, D. B. **Statistical analysis with missing data**. [*S. l.*]: John Wiley & Sons, 2019.

MCKNIGHT, P. E.; MCKNIGHT, K. M.; SIDANI, S.; FIGUEREDO, A. J. **Missing data: A gentle introduction**. [*S. l.*]: Guilford Press, 2007.

MENDONÇA, T. S. **Modelos de regressão logística clássica, bayesiana e redes neurais para credit ccoring.** 2008**,** 177 f. Dissertação (Mestrado em Estatística) — Universidade Federal de São Carlos, São Carlos, 2008.

MENDONÇA, T. S. **Estratégias para tratamento de variáveis com dados faltantes durante o desenvolvimento de modelos preditivos.** 2012, 74 f. Dissertação (Mestrado em Estatística) — Universidade de São Paulo, São Paulo, 2012.

MONFARDINI, F. **Modelos lineares generalizados bayesianos para dados longitudinais.** 2016**,** 79 f. Dissertação (Mestrado Matemática Aplicada e Computacional) — Universidade Estadual Paulista, Presidente Prudente, 2016.

NUNES, L. N. **Métodos de imputação de dados aplicados na área da saúde.** 2007**,** 120 f. Tese (Doutorado em Epidemiologia) — Universidade de Federal do Rio Grande do Sul, 2007. Disponível em: <http://hdl.handle.net/10183/11422>. Acesso em.: 11 de outubro de 2017.

PARK, T.; CASELLA, G. The bayesian lasso. **Journal of the American Statistical Association**, Taylor & Francis, n. 482, p. 681–686, 2008.

PAULA, G. Modelos de regressão com apoio computacional. **São Paulo-SP: Instituto de Matemática e Estatística (IME), Universidad de São Paulo**, 2010.

R Core Team. **R: A Language and Environment for Statistical Computing**. Vienna, Austria, 2019. Disponível em: <https://www.R-project.org/>.

RUBIN, D. B. Multiple imputation after 18+ years. **Journal of the American statistical Association**, Taylor & Francis Group, n. 434, p. 473–489, 1996.

SCHAFER, J. L. Imputation of missing covariates under a multivariate linear mixed model. **Tech**, National Institutes of Health, n. 1, p. 1–26, 1997.

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

SILVA, E. D. Análise de custos através de uma linguagem paramétrica usando o modelo"backward elimination". **Contabilidade Vista & Revista**, Platform workflow by OJS/PKP, n. 2, p. 26–27, 1992.

TIBSHIRANI, R. Regression shrinkage and selection via the lasso. **Journal of the Royal Statistical Society: Series B (Methodological)**, Wiley Online Library, n. 1, p. 267–288, 1996.

VERONEZE, R. **Tratamento de dados faltantes empregando biclusterização com imputação múltipla.** 2011**,** 203 f. Dissertação (Mestrado em Engenharia Elétrica) — Universidade Estadual de Campinas, Presidente Prudente, 2011.

WEDDERBURN, R. W. Quasi-likelihood functions, generalized linear models, and the gauss—newton method. **Biometrika**, Oxford University Press, n. 3, p. 439–447, 1974.

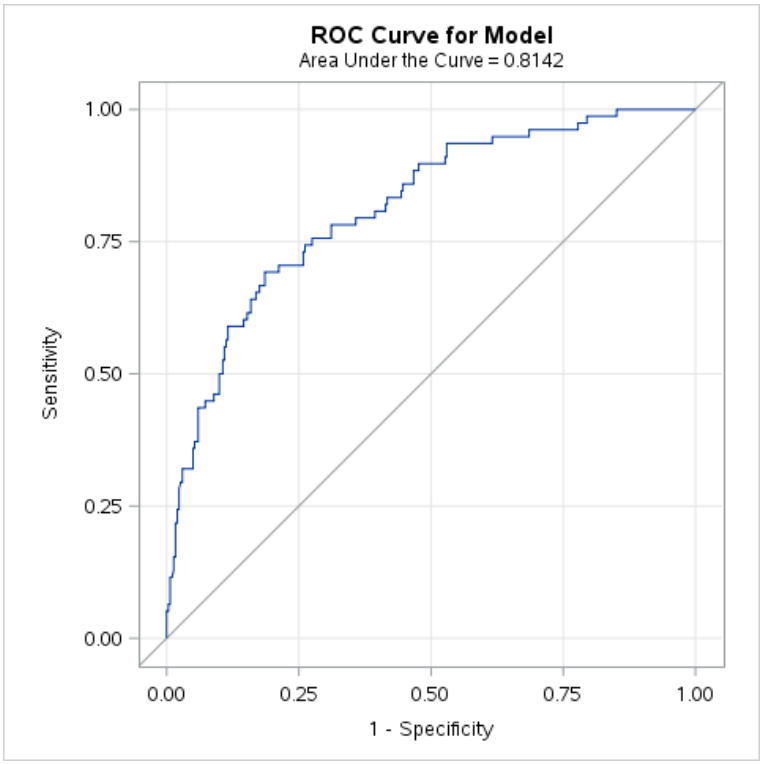
ZHANG, P. Multiple imputation: theory and method. **International Statistical Review**, Wiley Online Library, n. 3, p. 581–592, 2003.

APÊNDICE A – RESULTADOS DAS IMPUTAÇÕES.

No Apêndice A estão dispostos os resultados das 5 imputações realizadas, incluindo indicadores de ajustes, estimativas dos parâmetros dos modelos ajustados e análise de diagnóstico.

Modelo Logístico Ajustado - Imputação 1

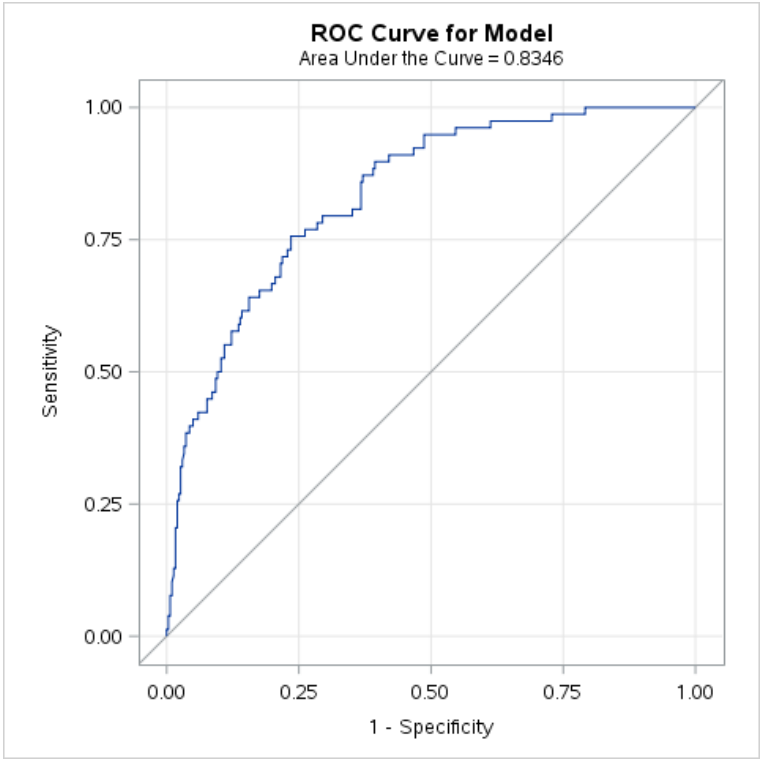
Figura 15 – Curva ROC para o modelo gerado na primeira imputação.



Fonte: Feita no SAS.

Modelo Logístico Ajustado - Imputação 2

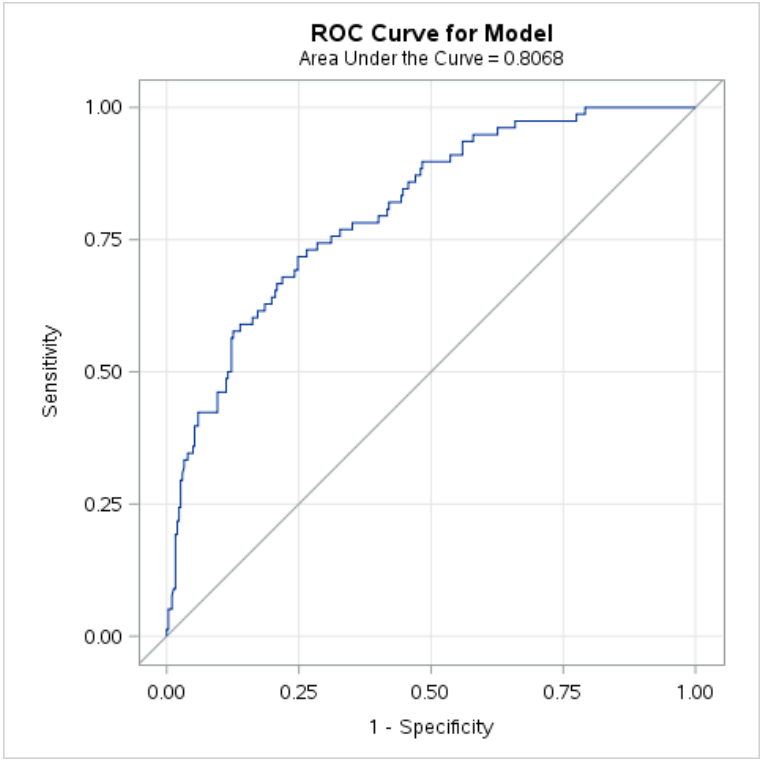
Figura 16 – Curva ROC para o modelo gerado na segunda imputação.



Fonte: Feita no SAS.

Modelo Logístico Ajustado - Imputação 3

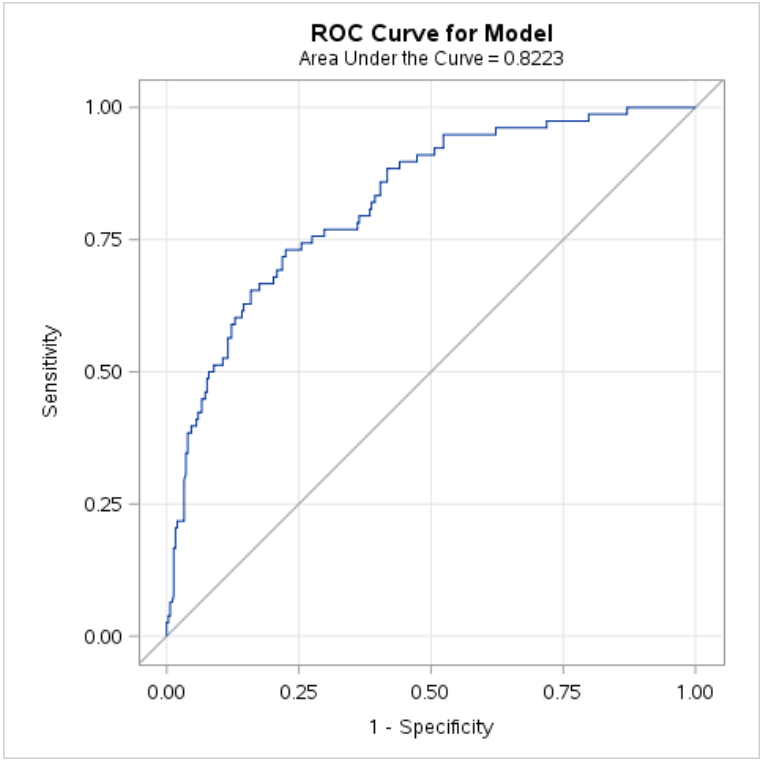
Figura 17 – Curva ROC para o modelo gerado na terceira imputação.



Fonte: Feita no SAS.

Modelo Logístico Ajustado - Imputação 4

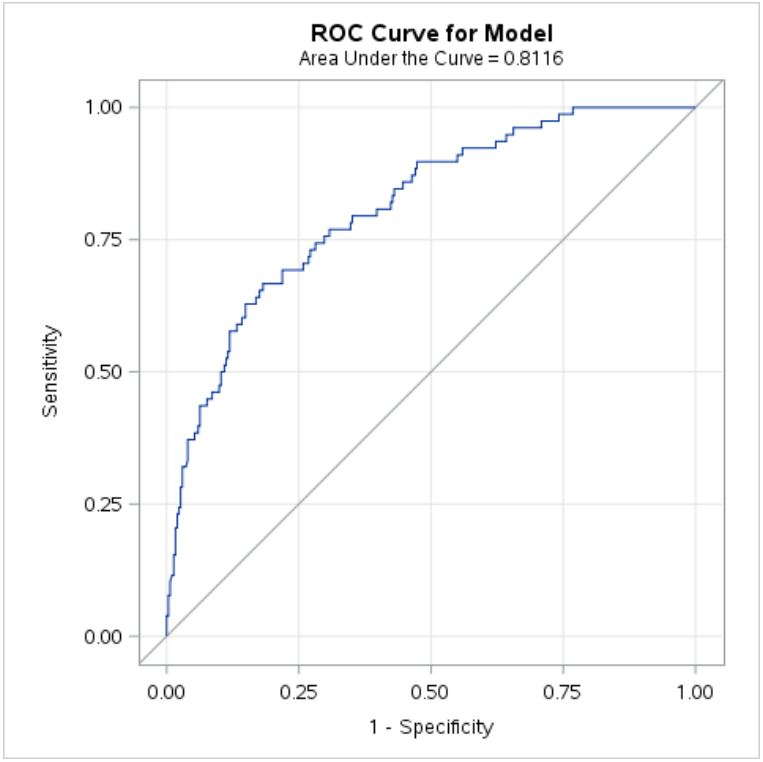
Figura 18 – Curva ROC para o modelo gerado na quarta imputação.



Fonte: Feita no SAS.

Modelo Logístico Ajustado - Imputação 5

Figura 19 – Curva ROC para o modelo gerado na quinta imputação.

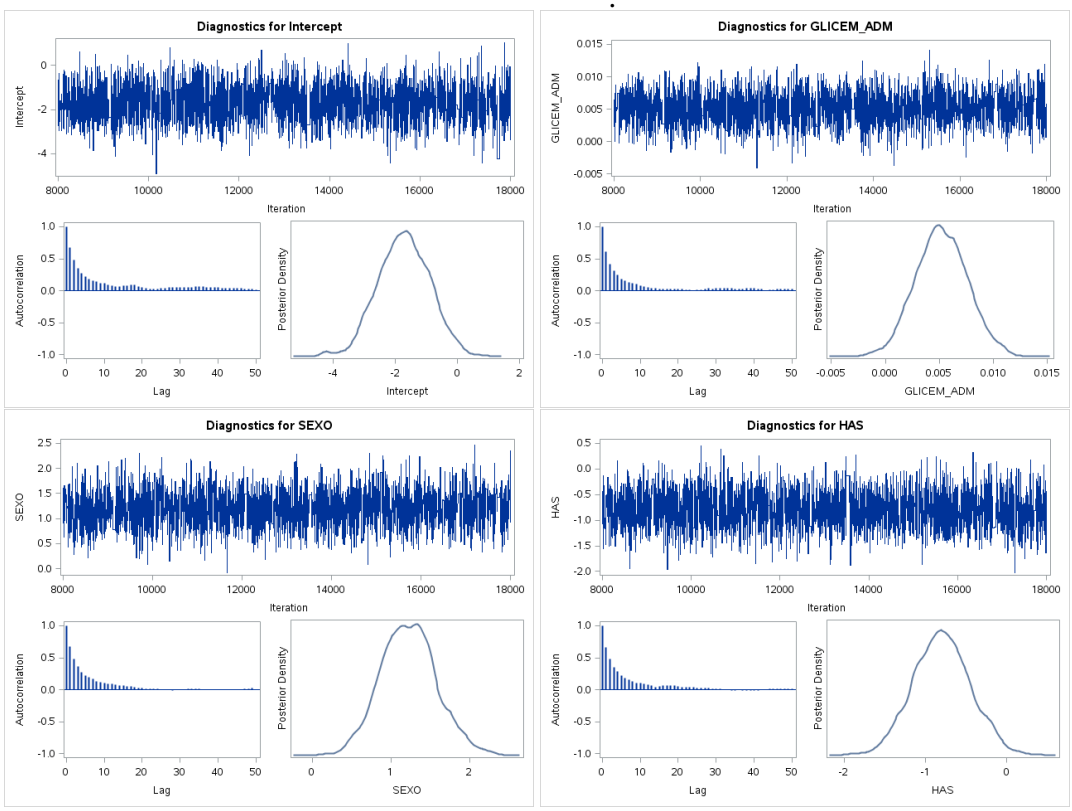


Fonte: Feita no SAS.

A seguir observa-se as cadeias de convergência das estimativas dos parâmetros estimados nas 5 imputações.

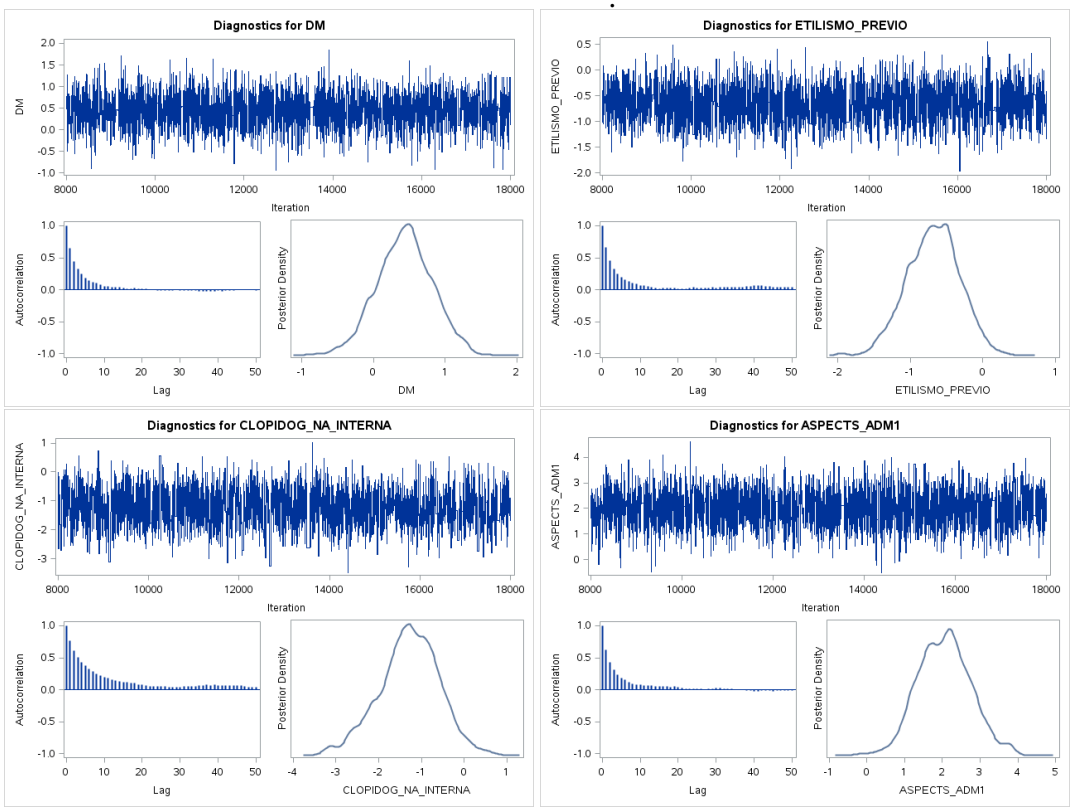
Diagnóstico de convergências dos parâmetros estimados. - Imputação 1

Figura 20 – Convergência dos parâmetros estimados pelo Modelo.



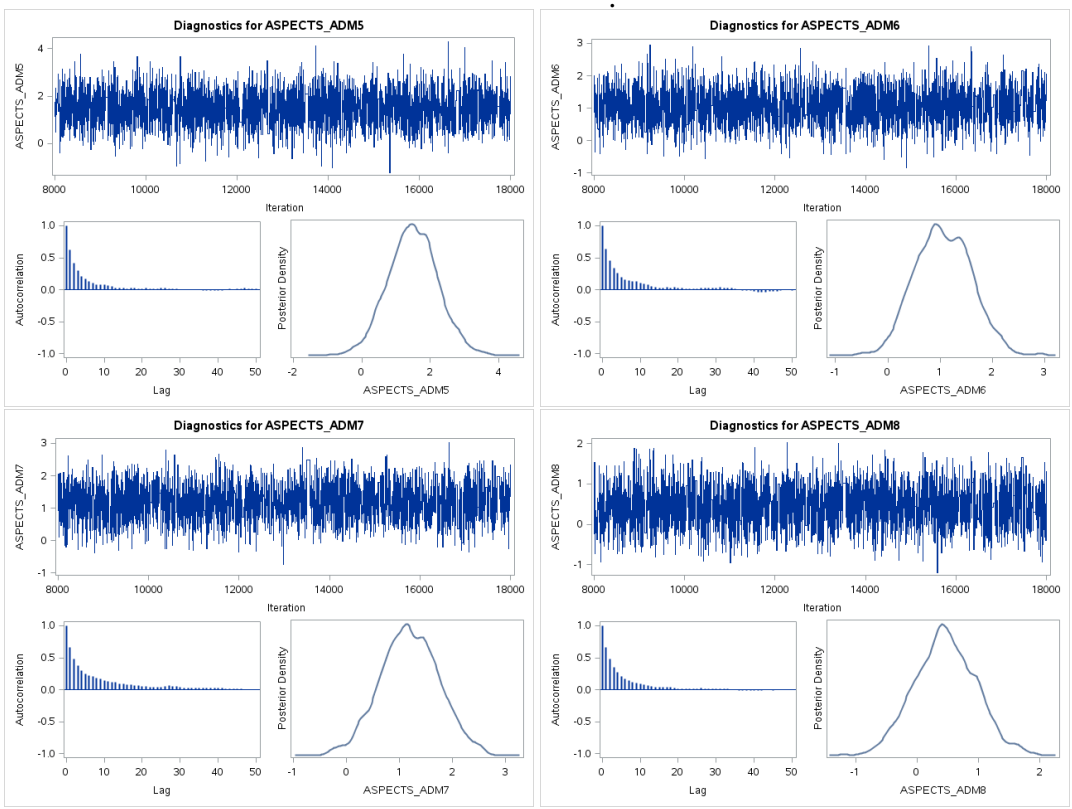
Fonte: Feita no SAS.

Figura 21 – Convergência dos parâmetros estimados pelo Modelo.



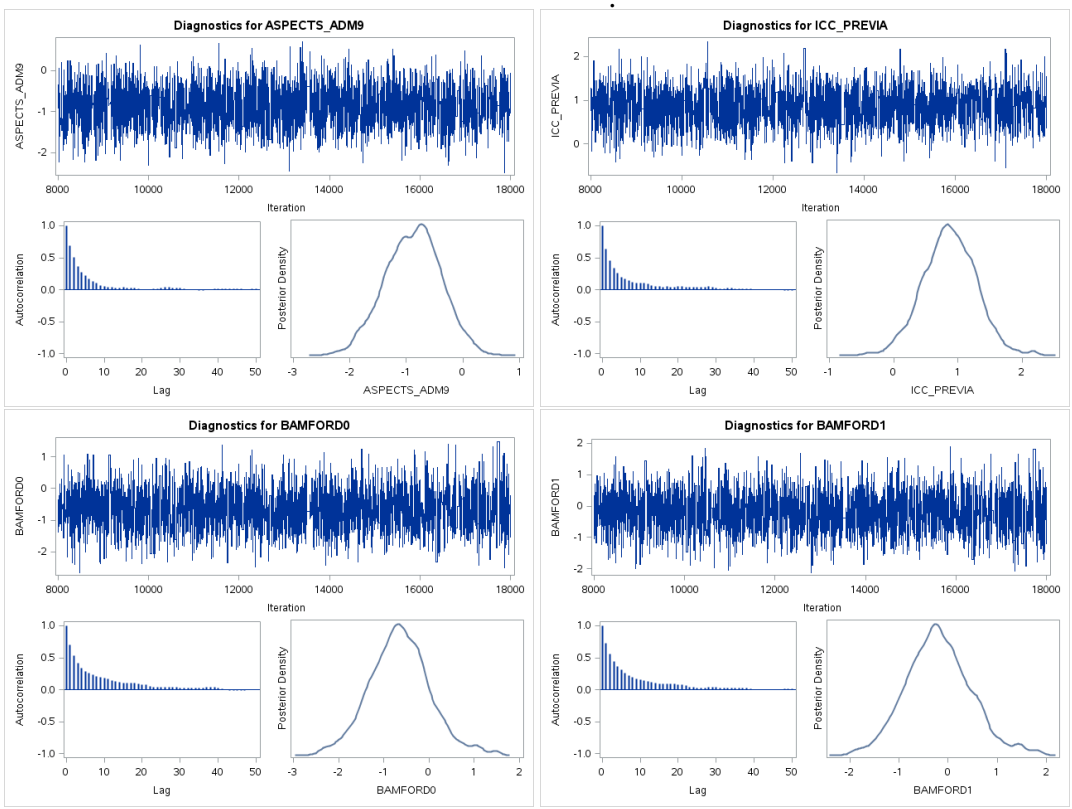
Fonte: Feita no SAS.

Figura 22 – Convergência dos parâmetros estimados pelo Modelo.



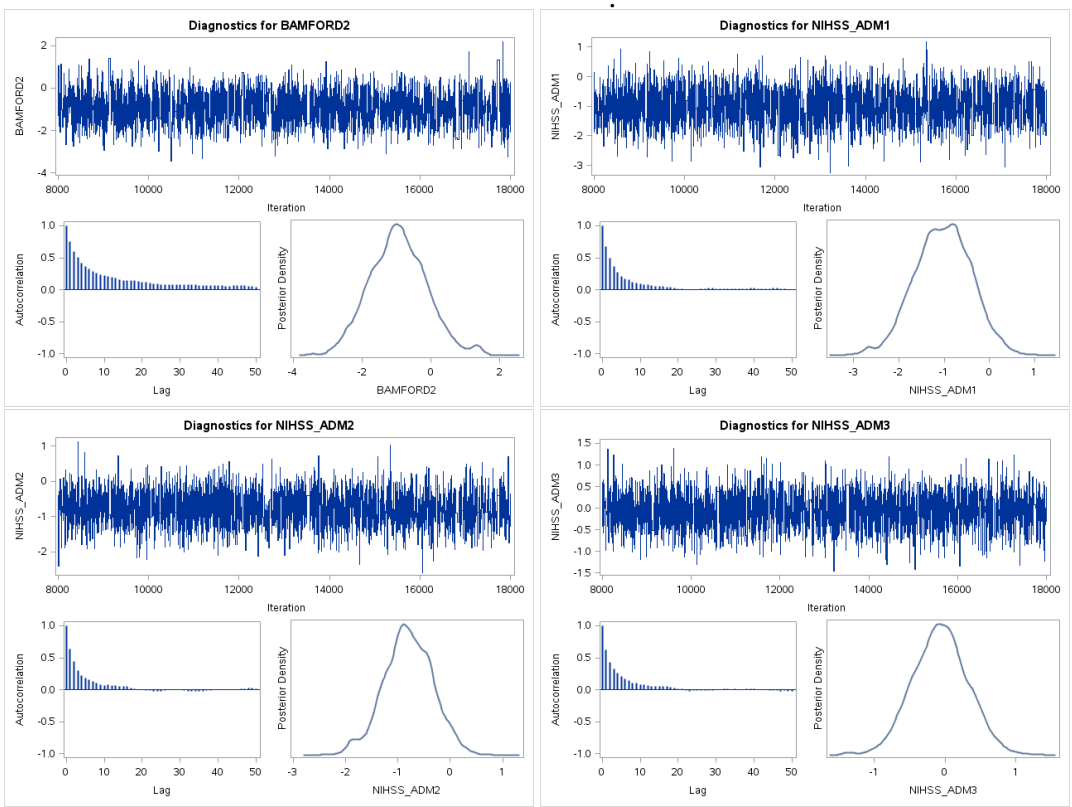
Fonte: Feita no SAS.

Figura 23 – Convergência dos parâmetros estimados pelo Modelo.



Fonte: Feita no SAS.

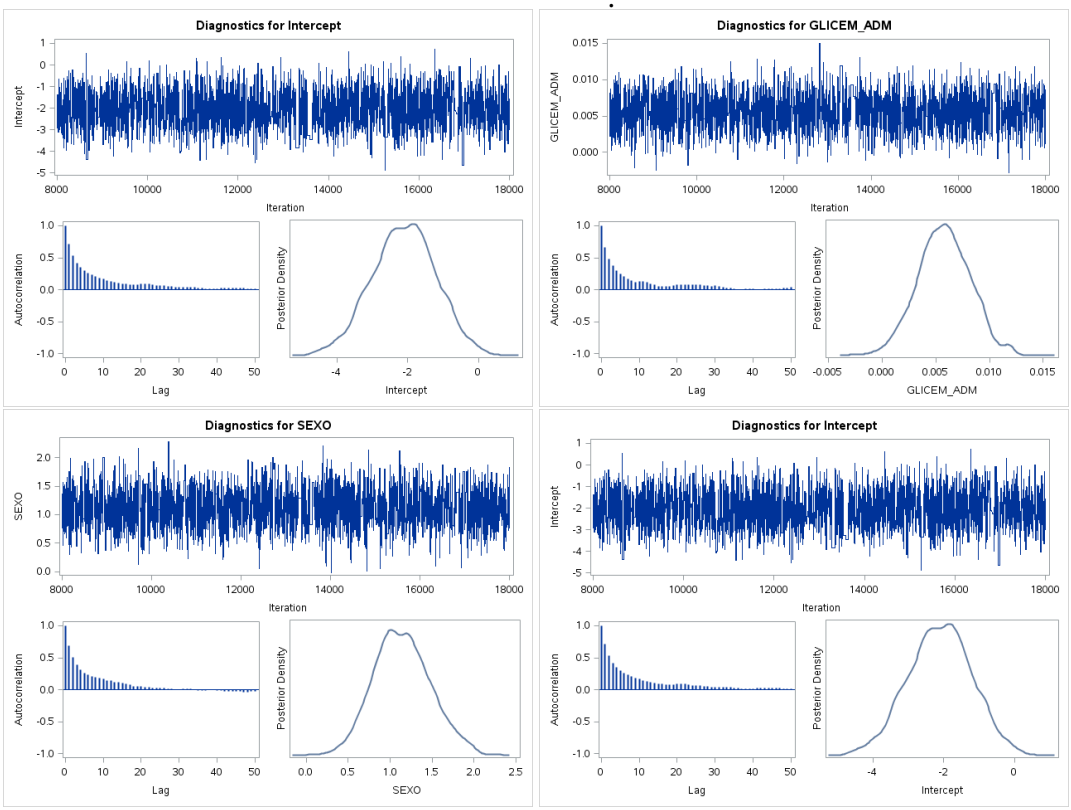
Figura 24 – Convergência dos parâmetros estimados pelo Modelo.



Fonte: Feita no SAS.

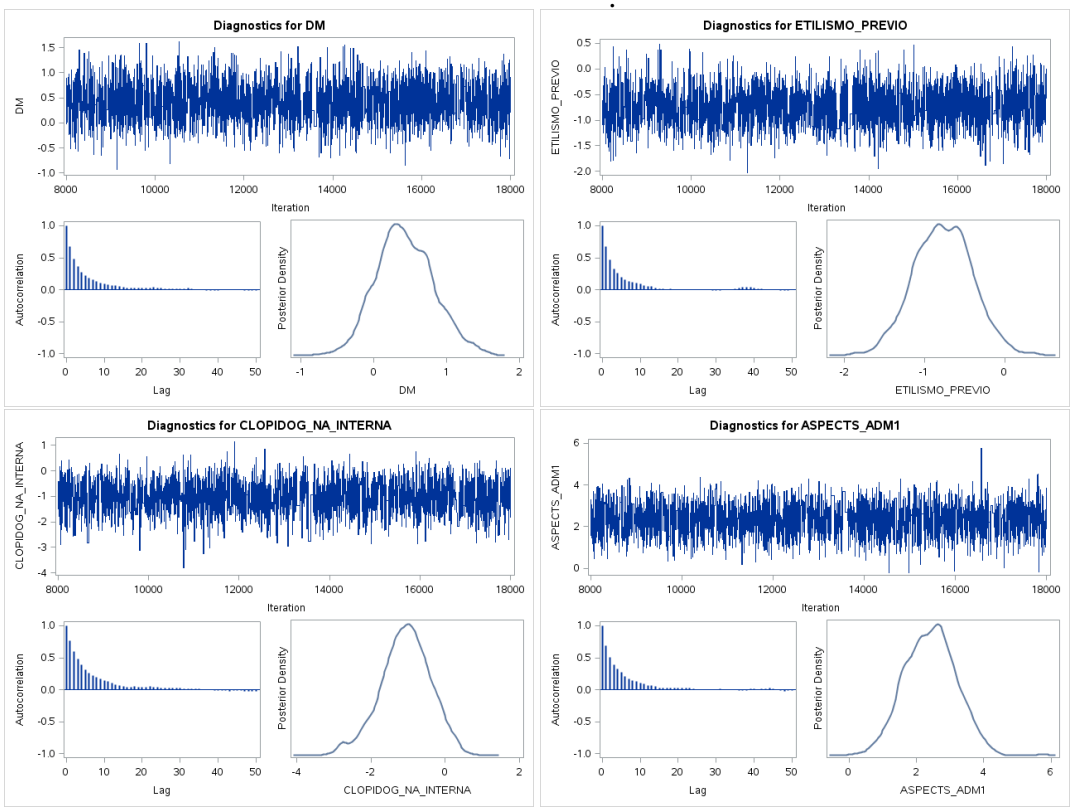
Diagnóstico de convergências dos parâmetros estimados. - Imputação 2

Figura 25 – Convergência dos parâmetros estimados pelo Modelo.



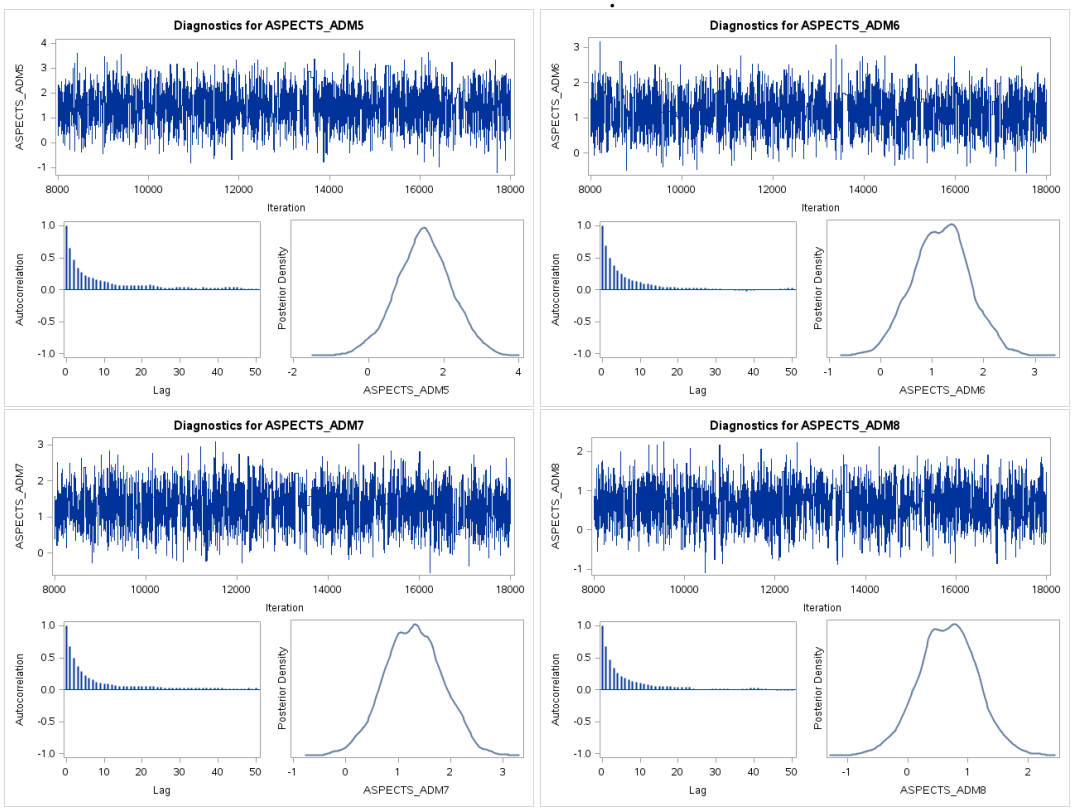
Fonte: Feita no SAS.

Figura 26 – Convergência dos parâmetros estimados pelo Modelo.



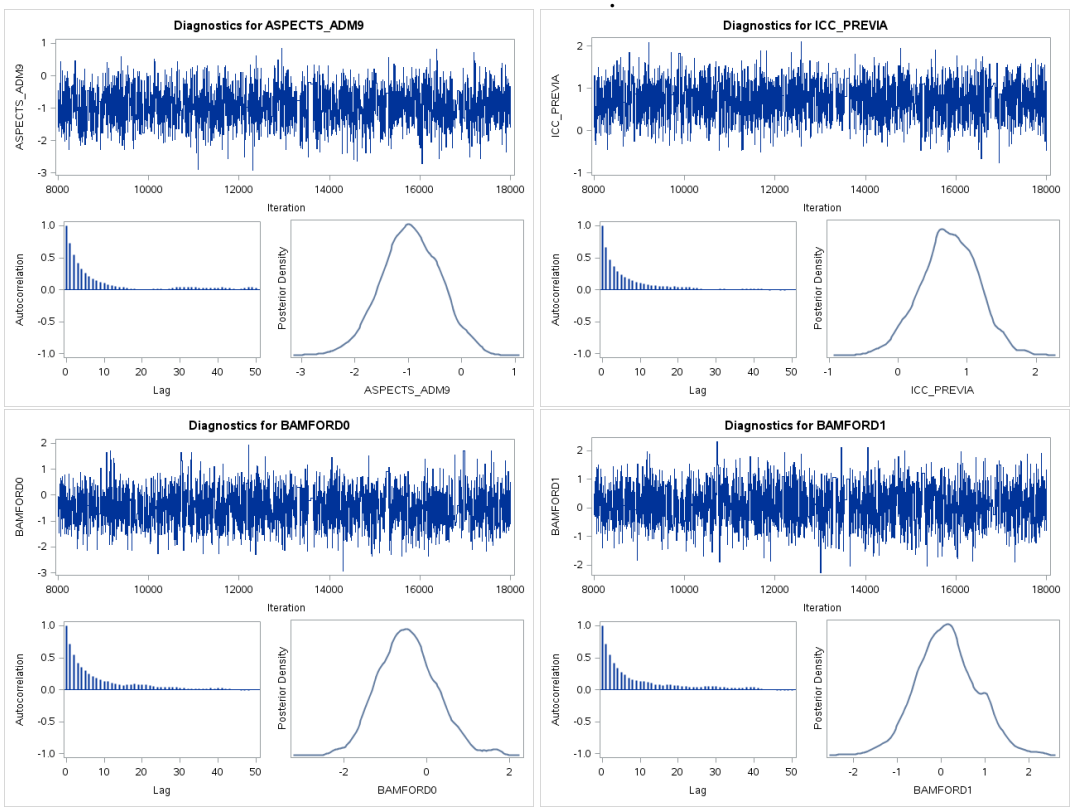
Fonte: Feita no SAS.

Figura 27 – Convergência dos parâmetros estimados pelo Modelo.



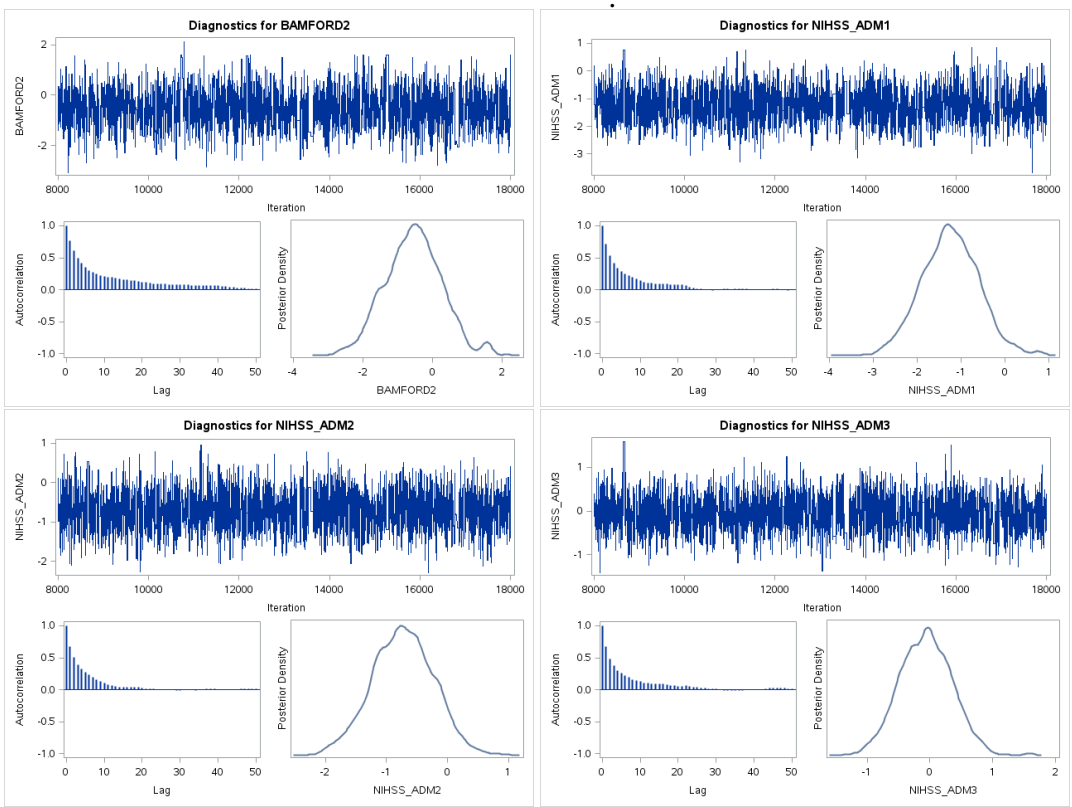
Fonte: Feita no SAS.

Figura 28 – Convergência dos parâmetros estimados pelo Modelo.



Fonte: Feita no SAS.

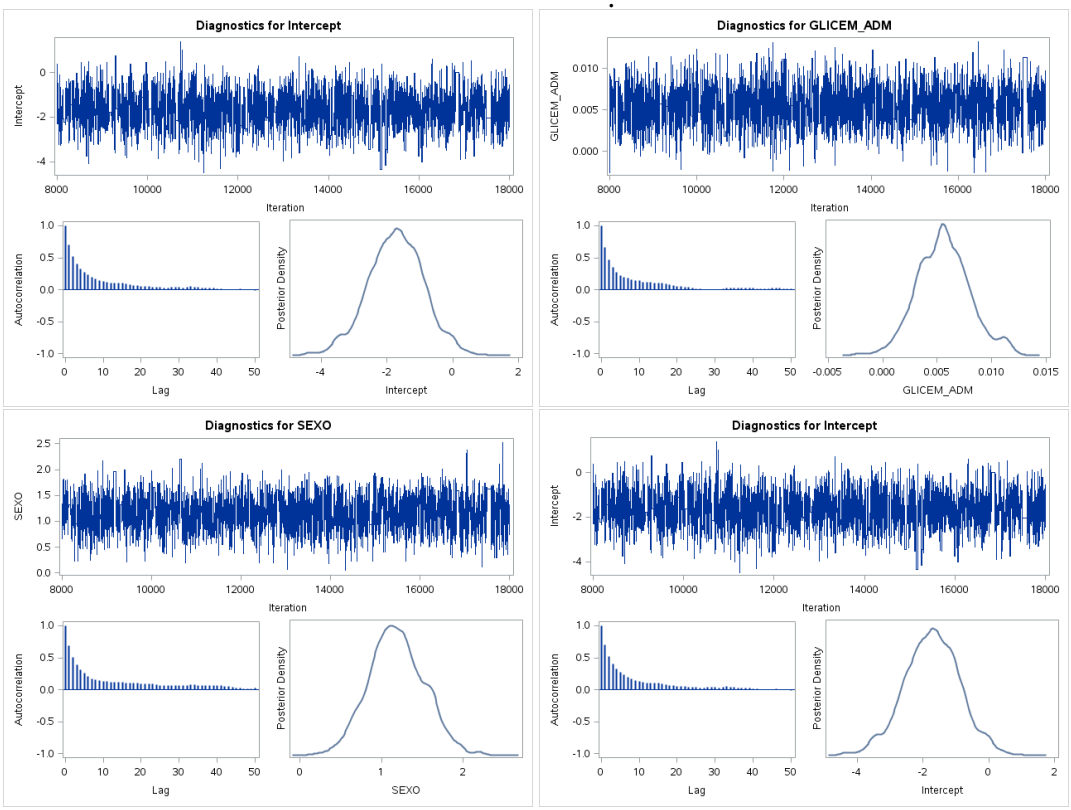
Figura 29 – Convergência dos parâmetros estimados pelo Modelo.



Fonte: Feita no SAS.

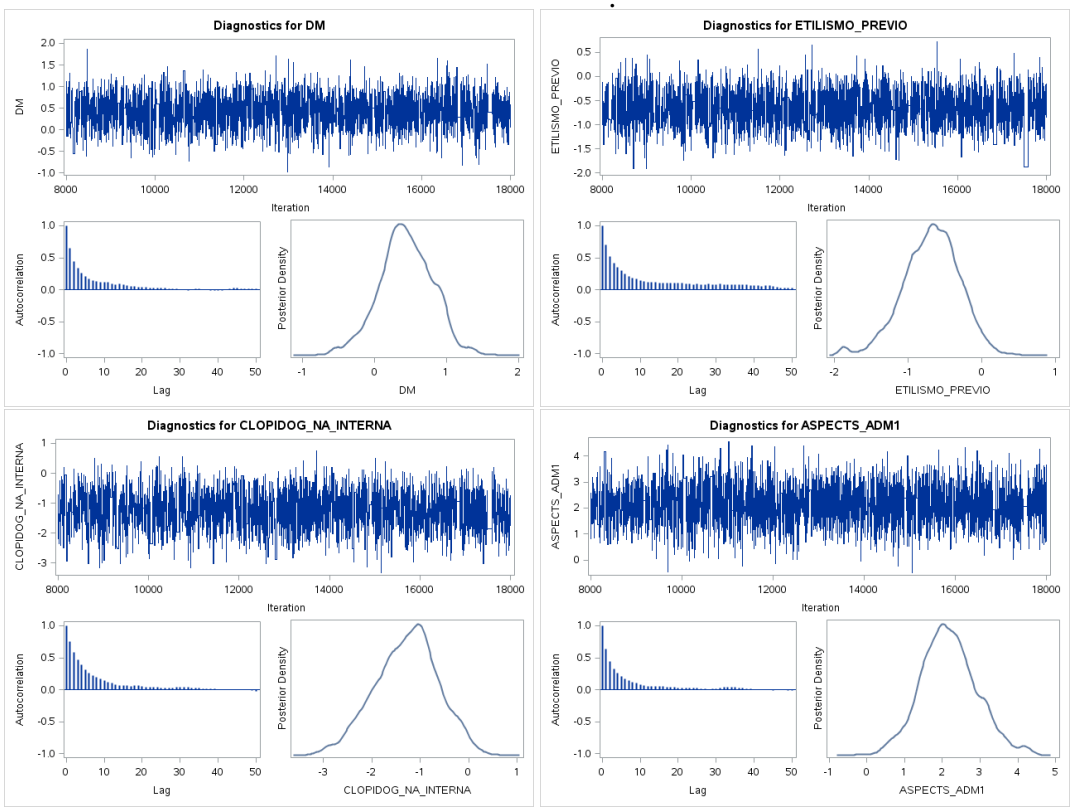
Diagnóstico de convergências dos parâmetros estimados. - Imputação 3

Figura 30 – Convergência dos parâmetros estimados pelo Modelo.



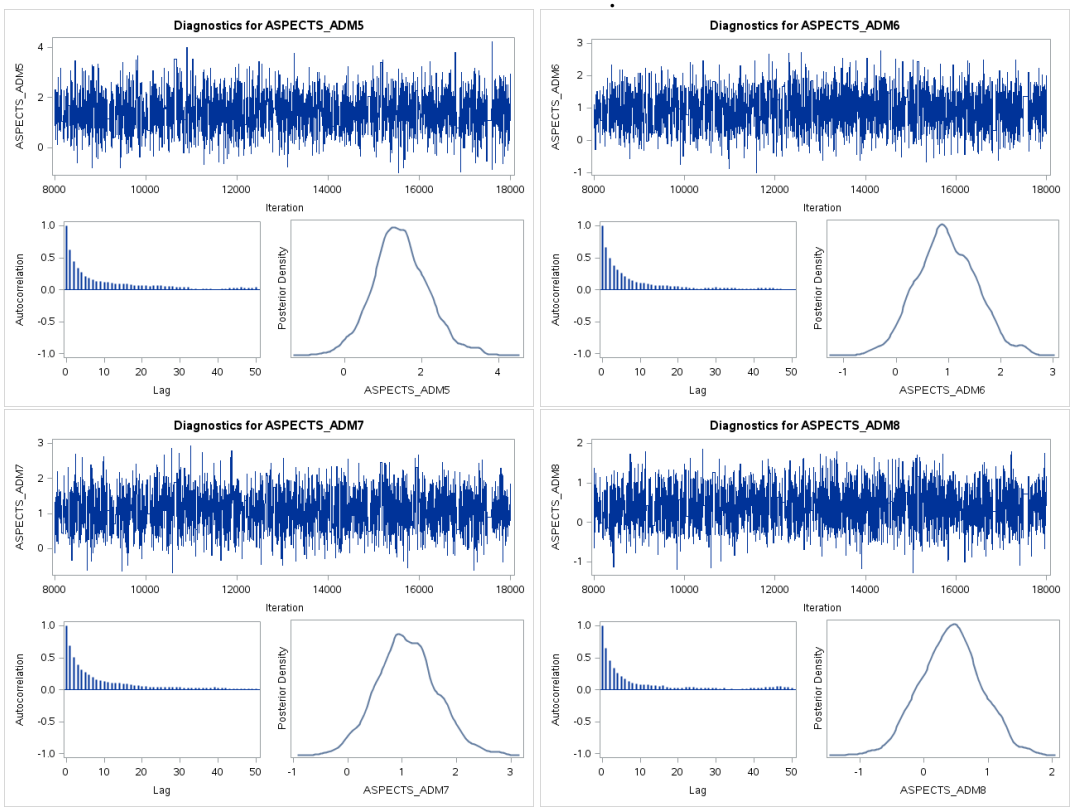
Fonte: Feita no SAS.

Figura 31 – Convergência dos parâmetros estimados pelo Modelo.



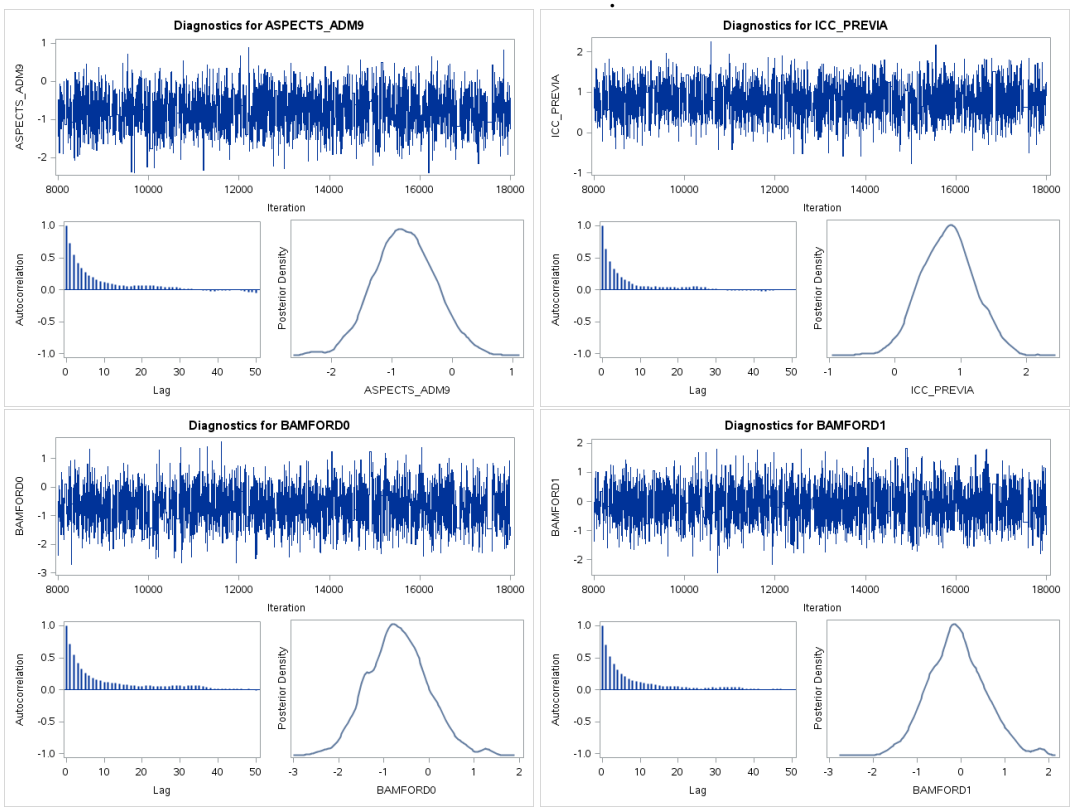
Fonte: Feita no SAS.

Figura 32 – Convergência dos parâmetros estimados pelo Modelo.



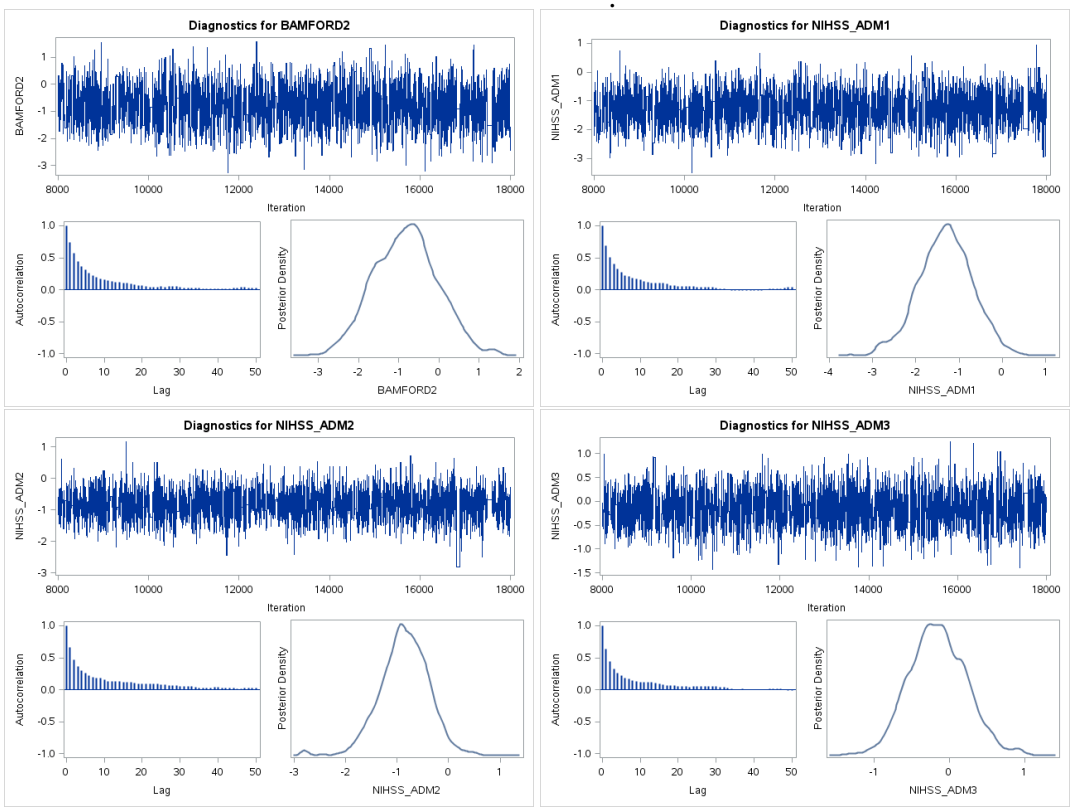
Fonte: Feita no SAS.

Figura 33 – Convergência dos parâmetros estimados pelo Modelo.



Fonte: Feita no SAS.

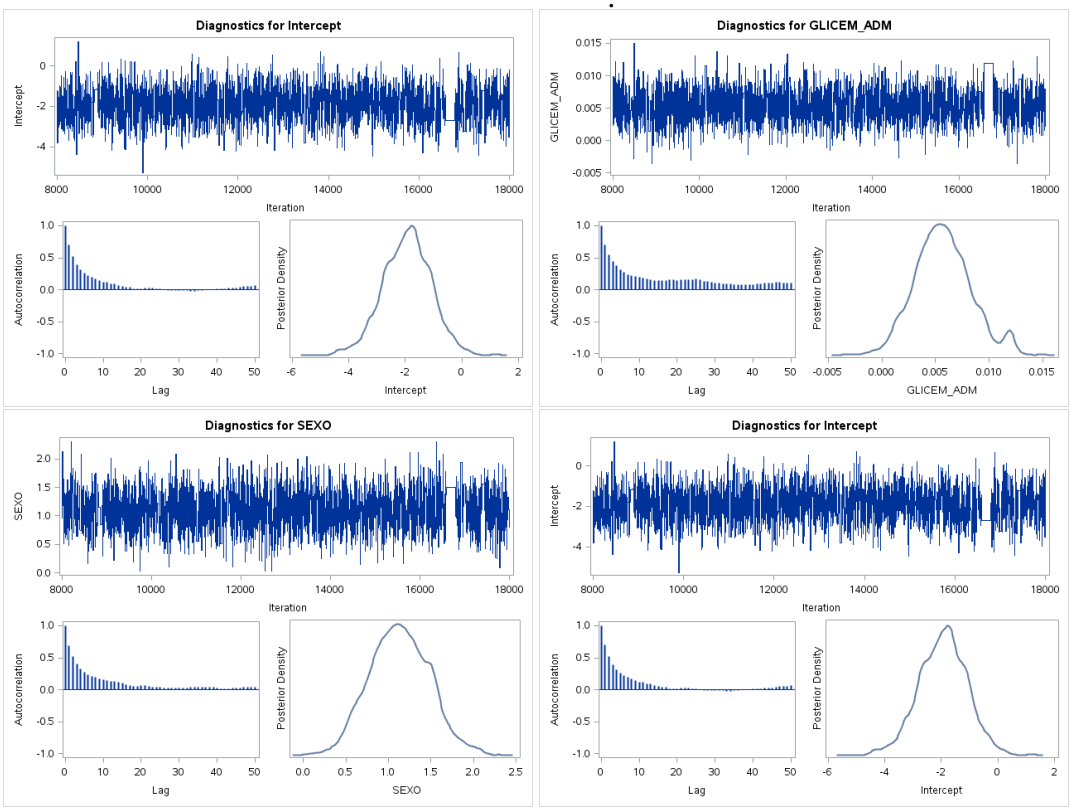
Figura 34 – Convergência dos parâmetros estimados pelo Modelo.



Fonte: Feita no SAS.

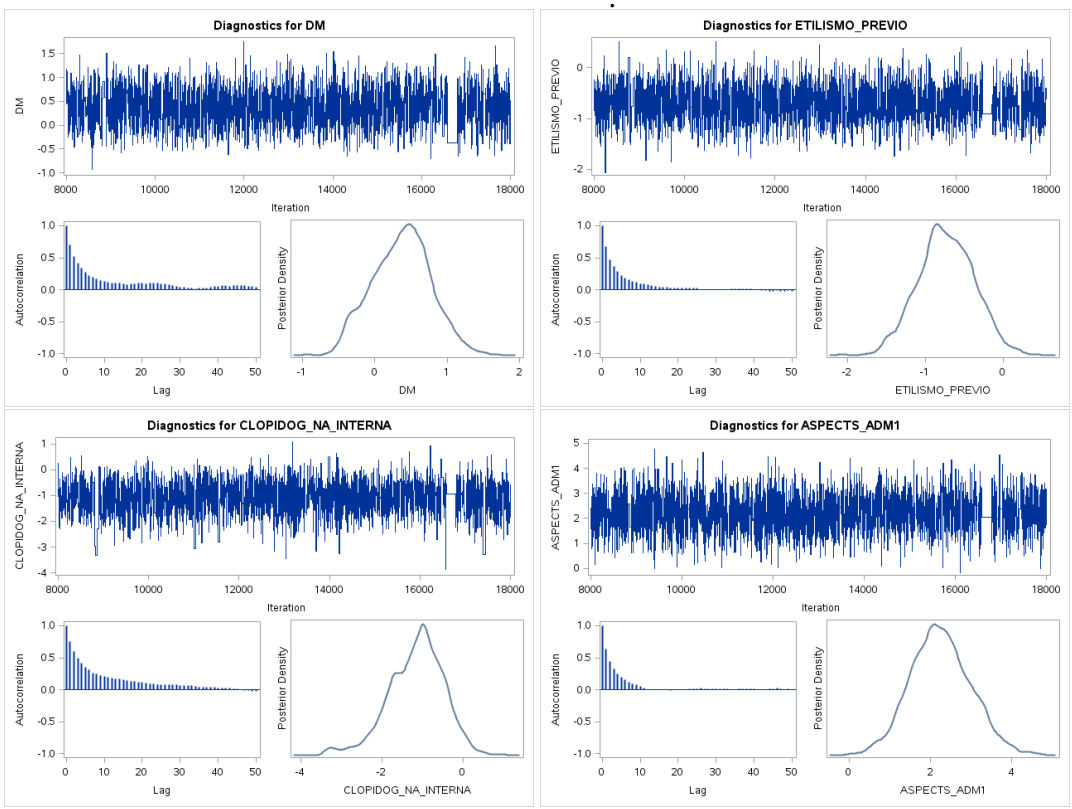
Diagnóstico de convergências dos parâmetros estimados. - Imputação 4

Figura 35 – Convergência dos parâmetros estimados pelo Modelo.



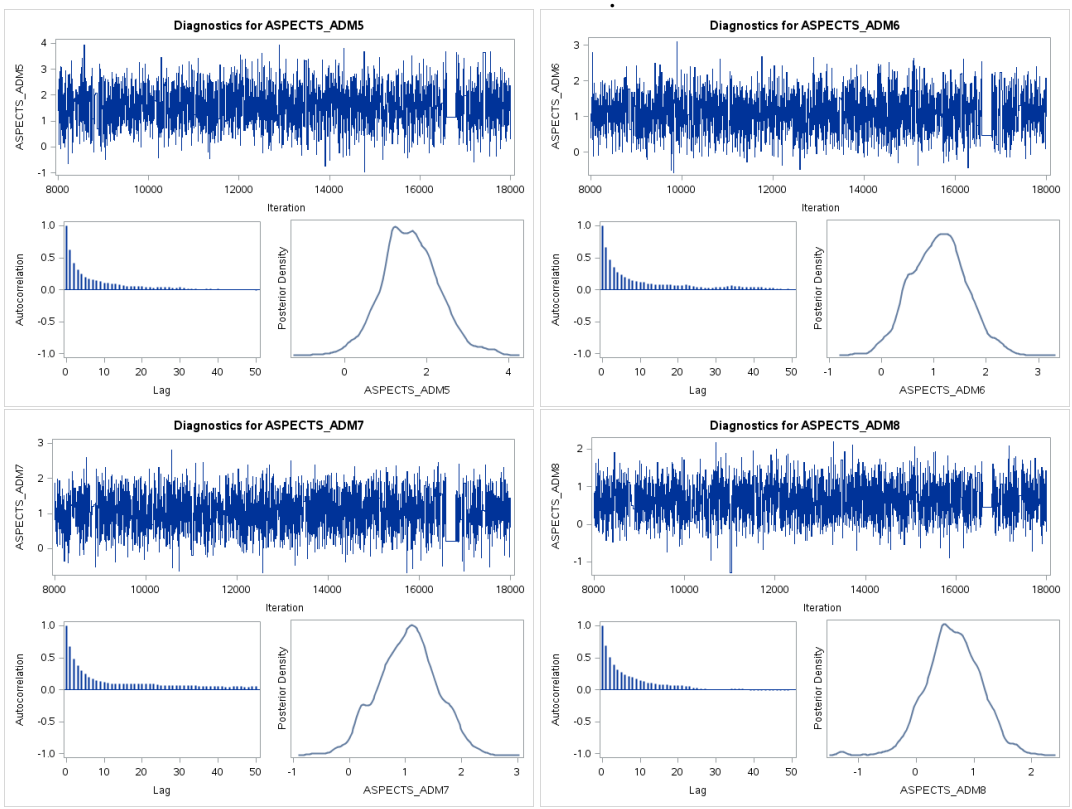
Fonte: Feita no SAS.

Figura 36 – Convergência dos parâmetros estimados pelo Modelo.



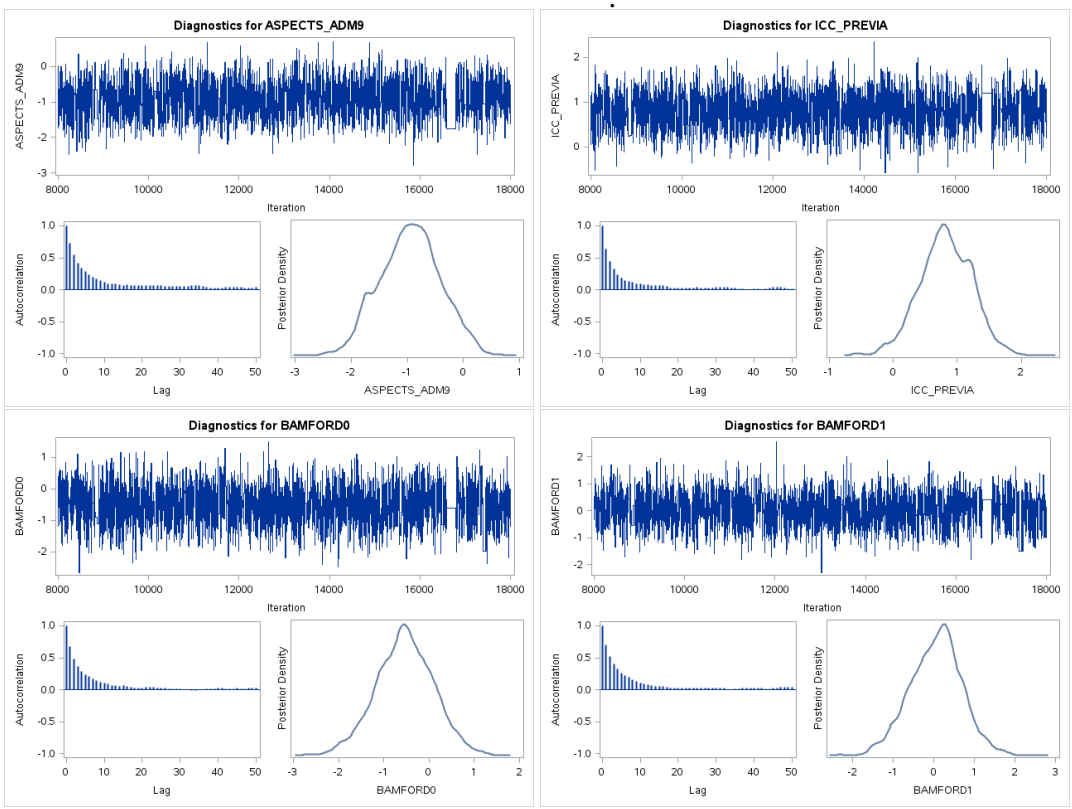
Fonte: Feita no SAS.

Figura 37 – Convergência dos parâmetros estimados pelo Modelo.



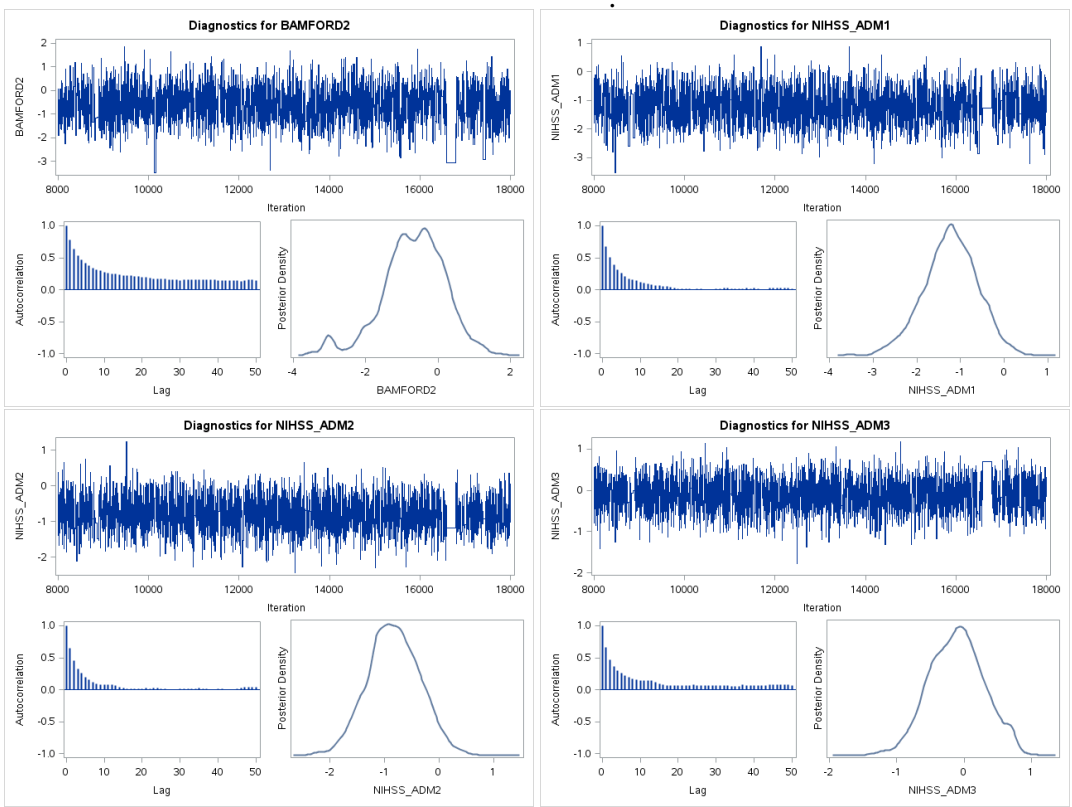
Fonte: Feita no SAS.

Figura 38 – Convergência dos parâmetros estimados pelo Modelo.



Fonte: Feita no SAS.

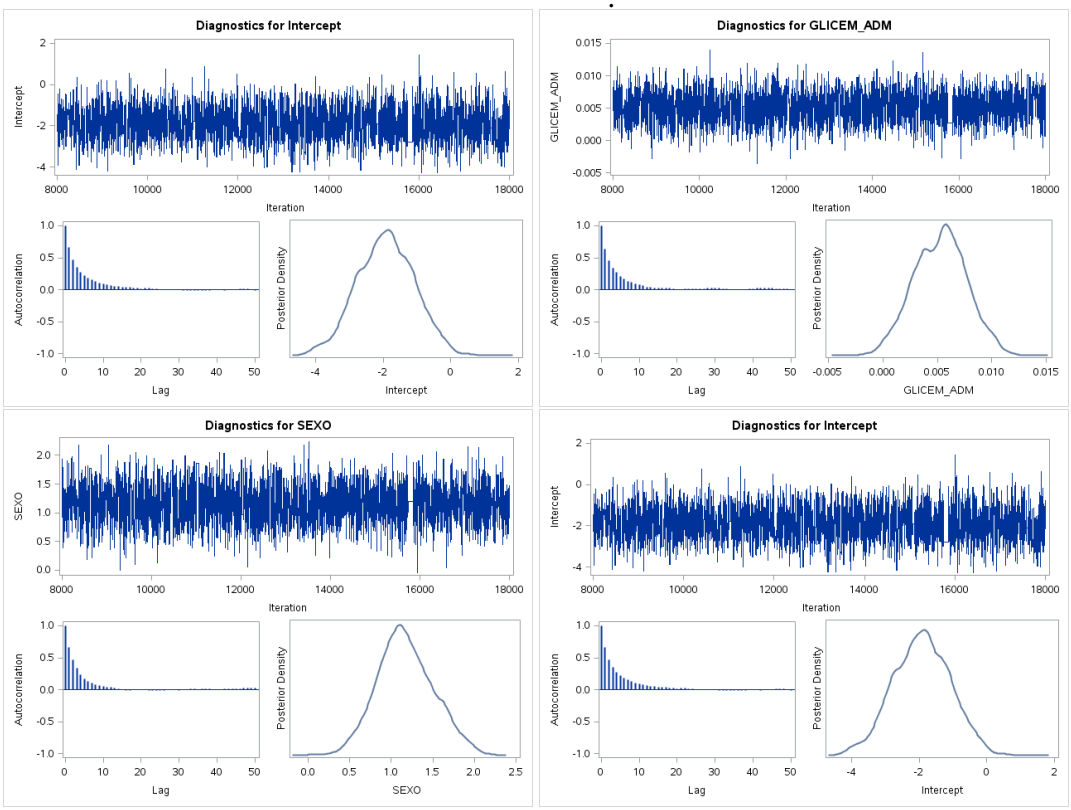
Figura 39 – Convergência dos parâmetros estimados pelo Modelo.



Fonte: Feita no SAS.

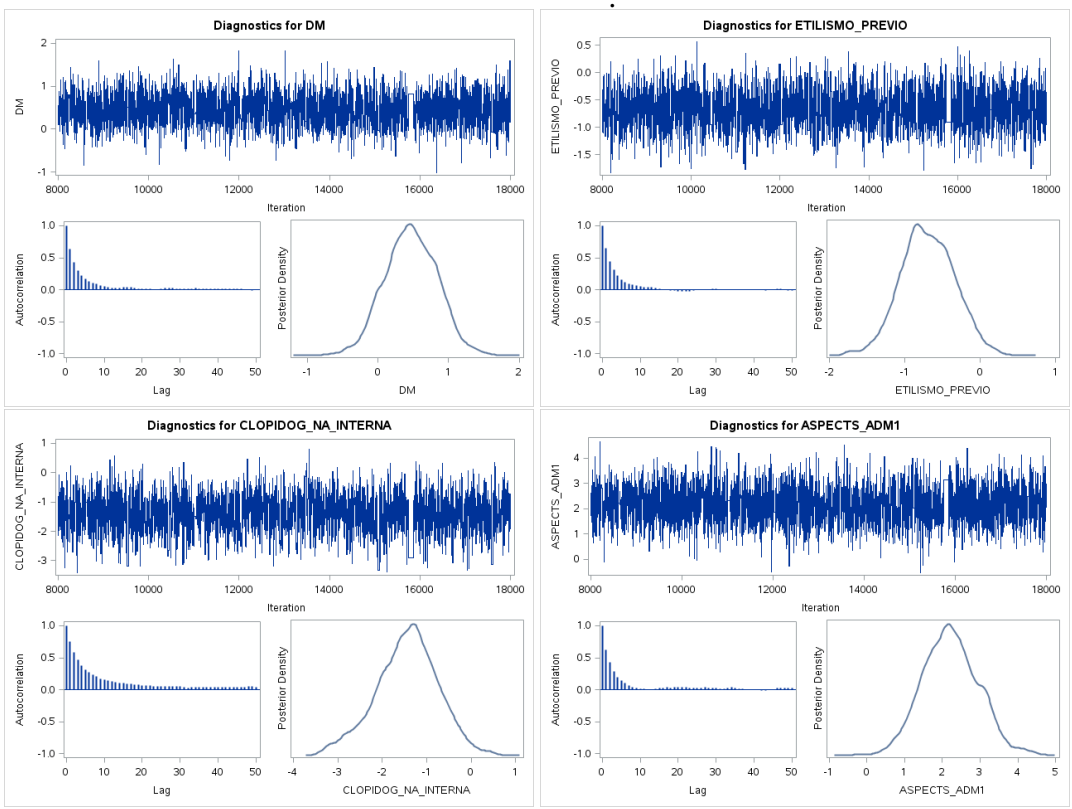
Diagnóstico de convergências dos parâmetros estimados. - Imputação 5

Figura 40 – Convergência dos parâmetros estimados pelo Modelo.



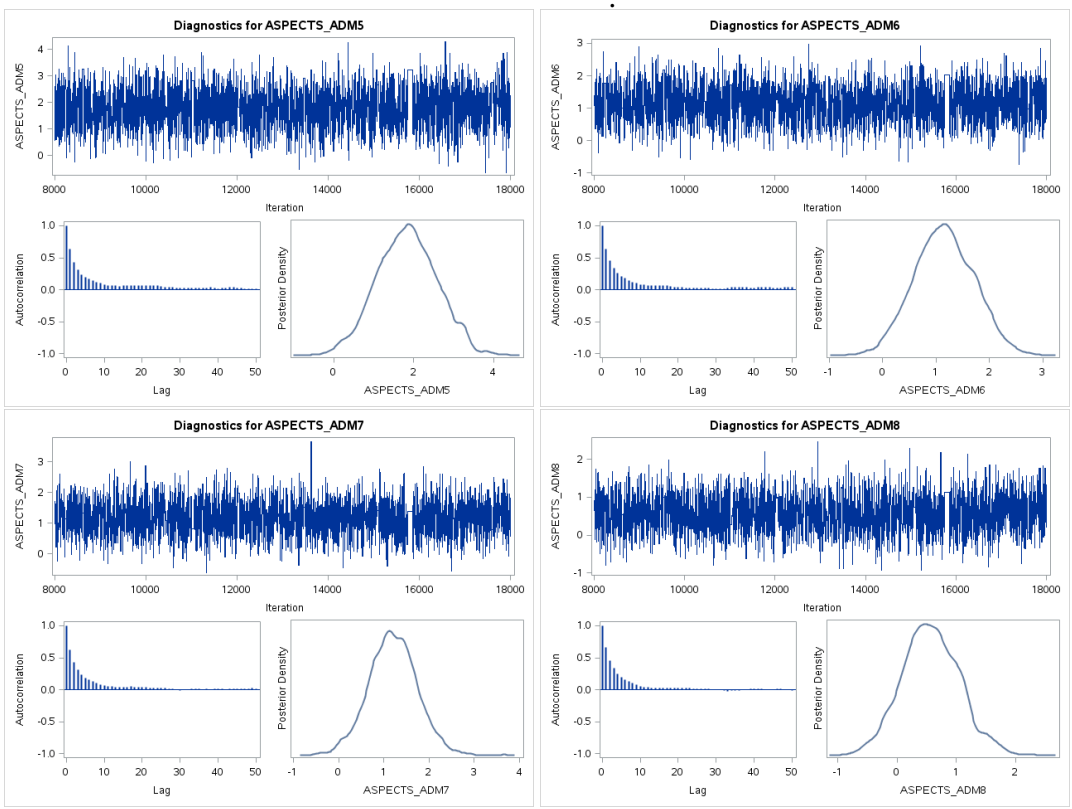
Fonte: Feita no SAS.

Figura 41 – Convergência dos parâmetros estimados pelo Modelo.



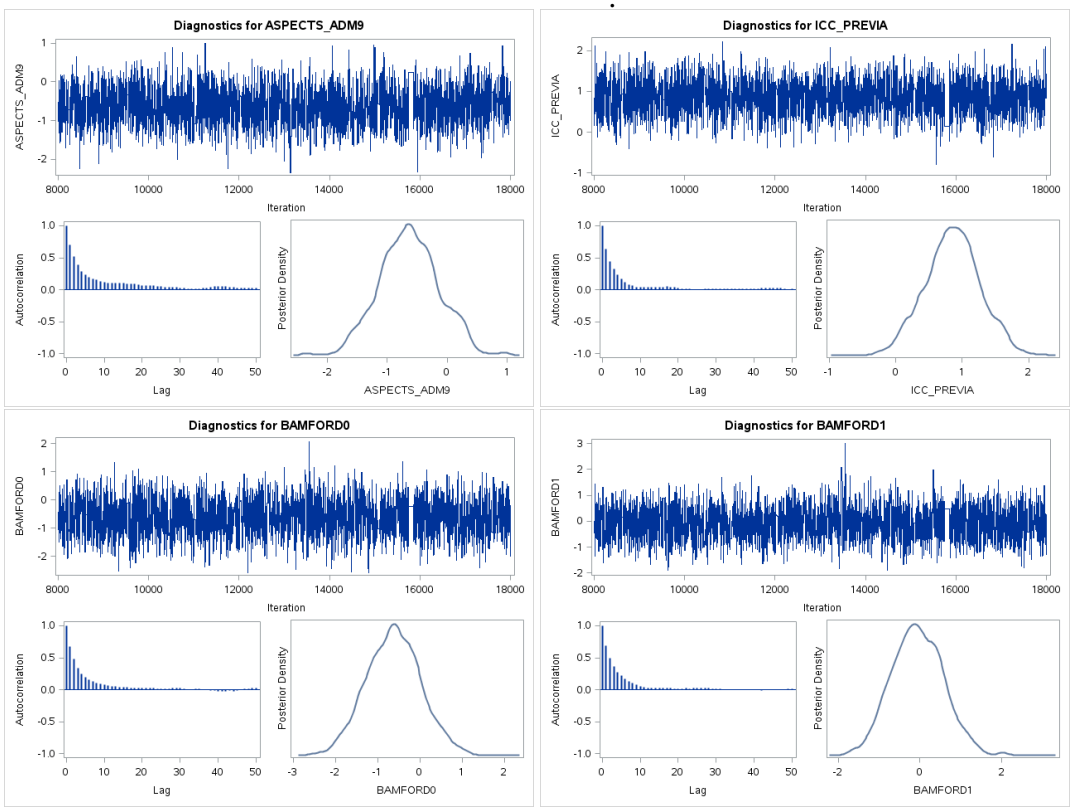
Fonte: Feita no SAS.

Figura 42 – Convergência dos parâmetros estimados pelo Modelo.



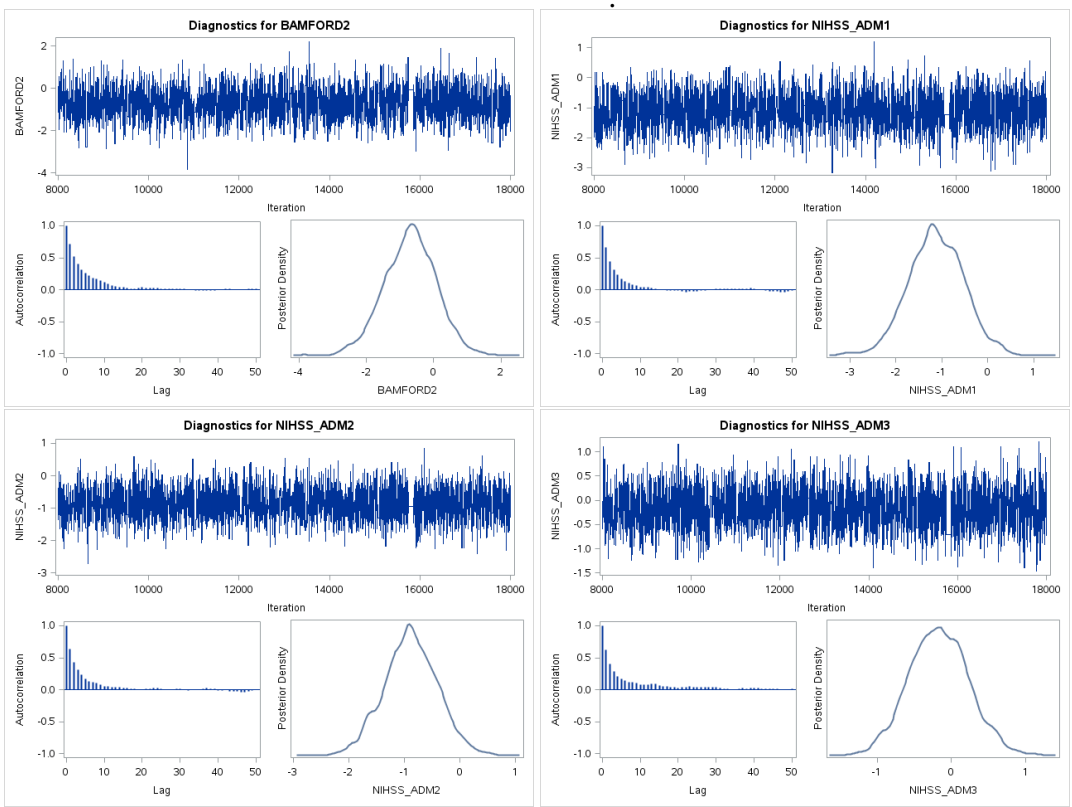
Fonte: Feita no SAS.

Figura 43 – Convergência dos parâmetros estimados pelo Modelo.



Fonte: Feita no SAS.

Figura 44 – Convergência dos parâmetros estimados pelo Modelo.

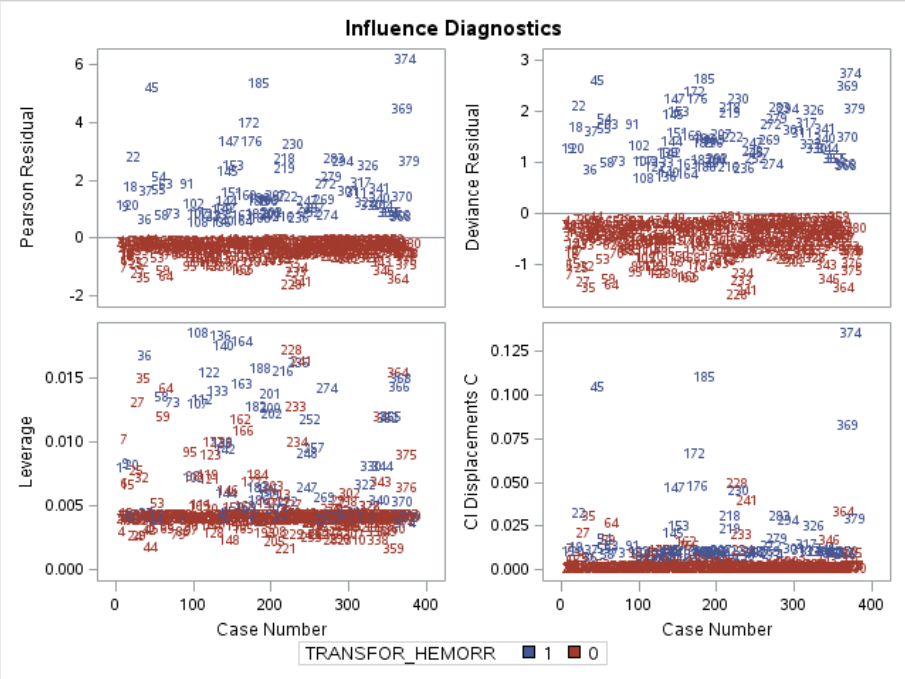


Fonte: Feita no SAS.

A seguir observa-se a análise de sensibilidade do modelo ajustado para os parâmetros estimados das 5 imputações.

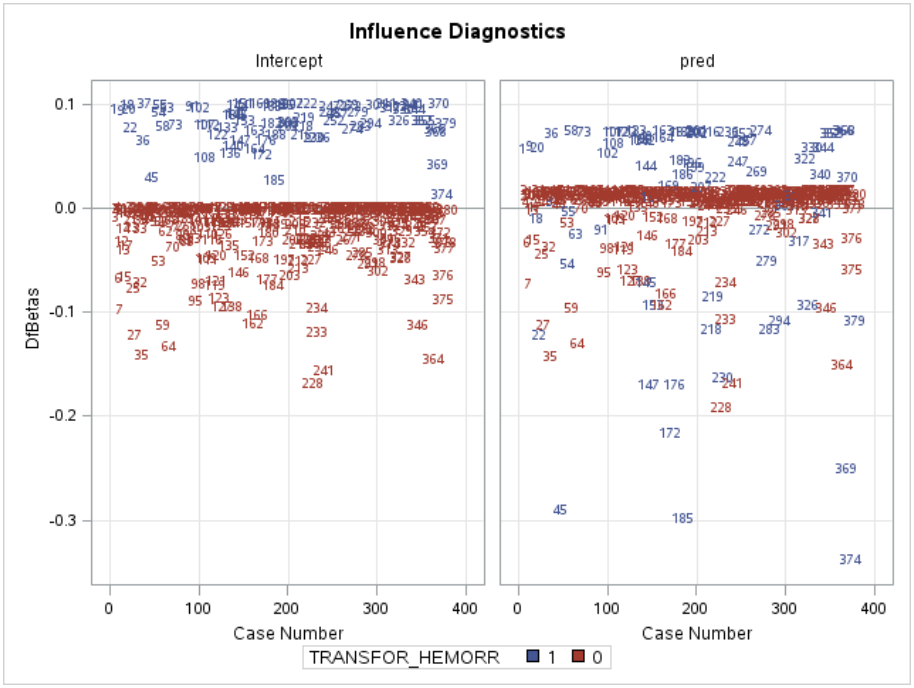
Diagnóstico do modelo ajustado. - Imputação 1

Figura 45 – Diagnóstico de observações influentes.



Fonte: Feita no SAS.

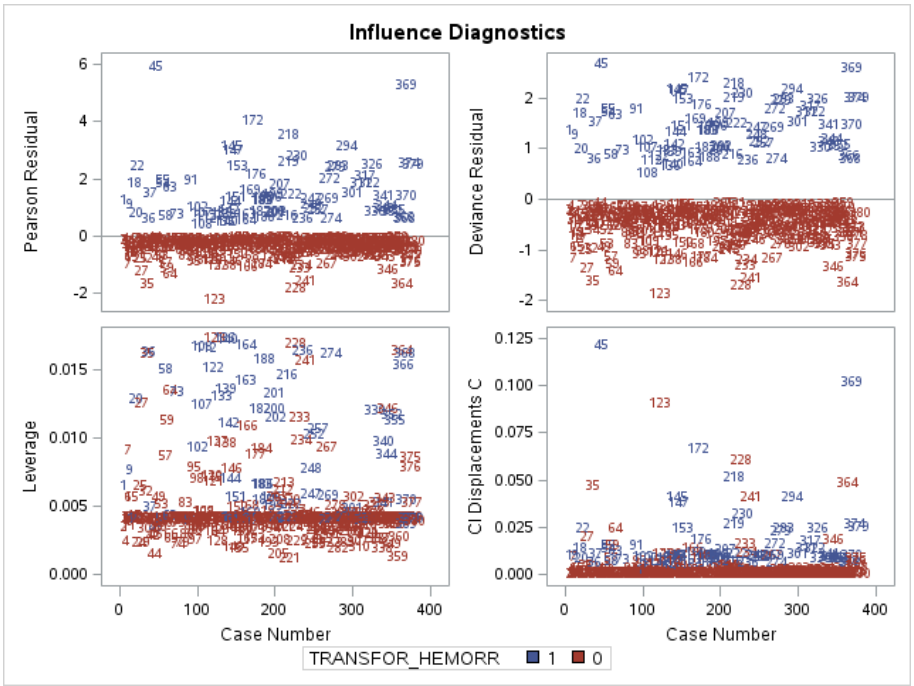
Figura 46 – Diagnóstico de observações influentes.



Fonte: Feita no SAS.

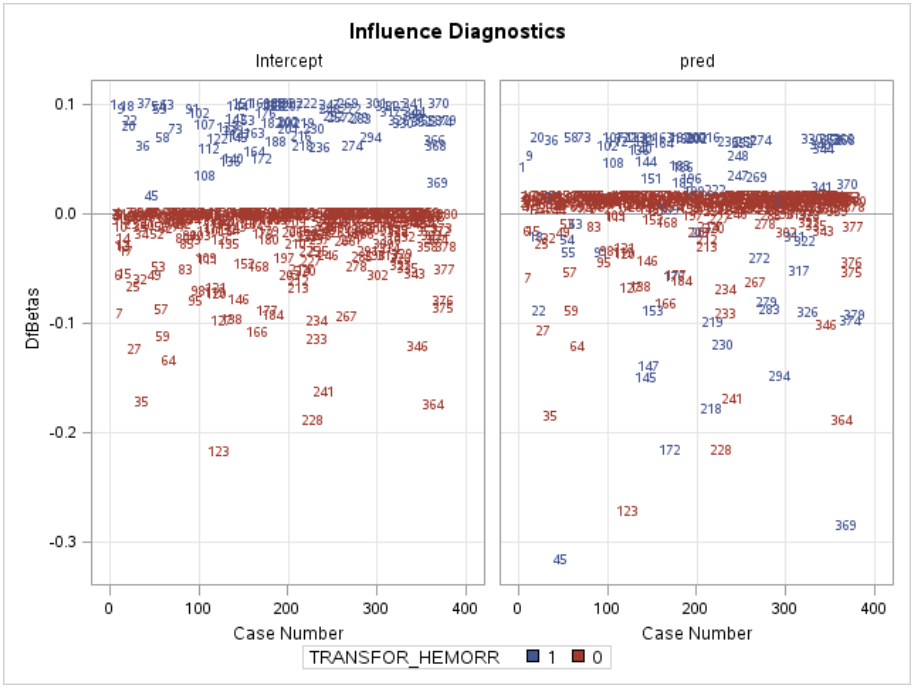
Diagnóstico do modelo ajustado. - Imputação 2

Figura 47 – Diagnóstico de observações influentes.



Fonte: Feita no SAS.

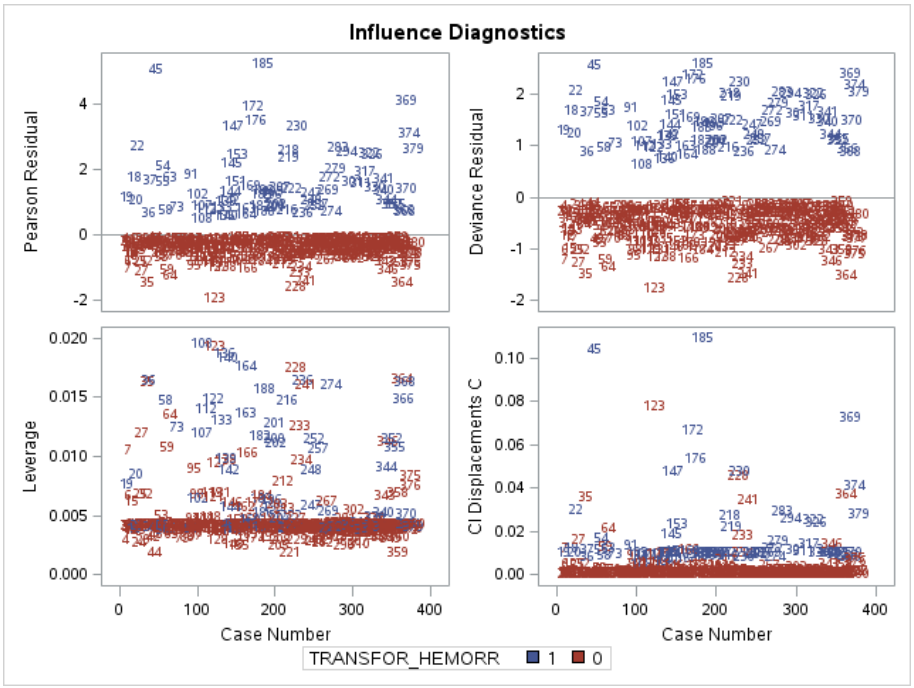
Figura 48 – Diagnóstico de observações influentes.



Fonte: Feita no SAS.

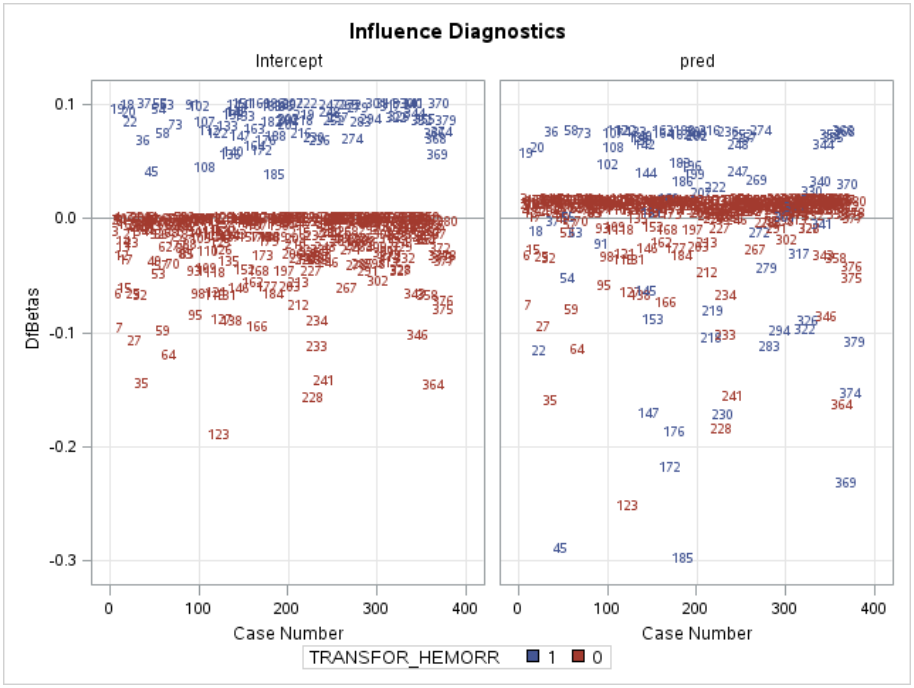
Diagnóstico do modelo ajustado. - Imputação 3

Figura 49 – Diagnóstico de observações influentes.



Fonte: Feita no SAS.

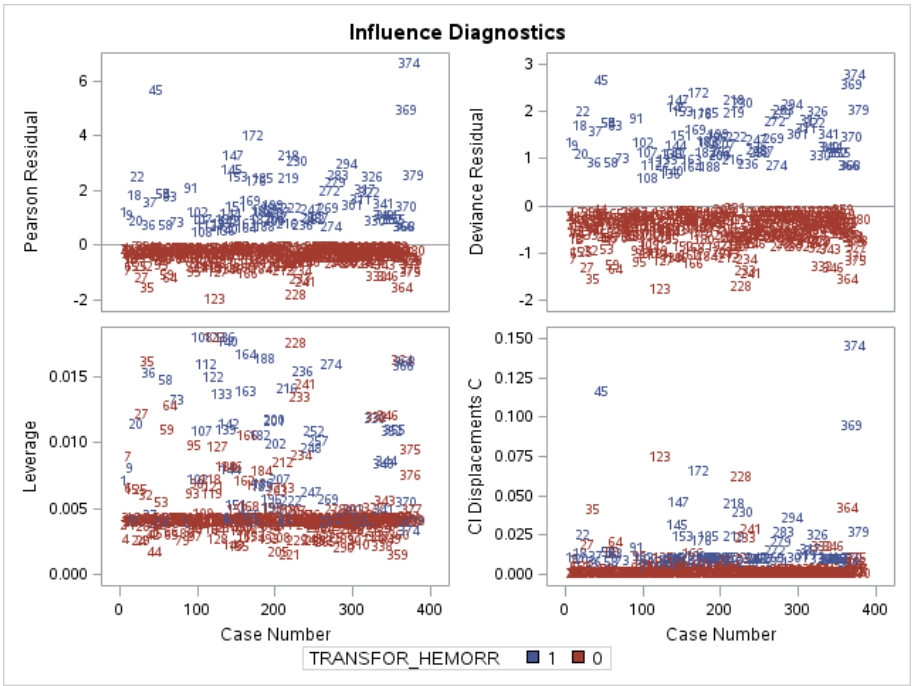
Figura 50 – Diagnóstico de observações influentes.



Fonte: Feita no SAS.

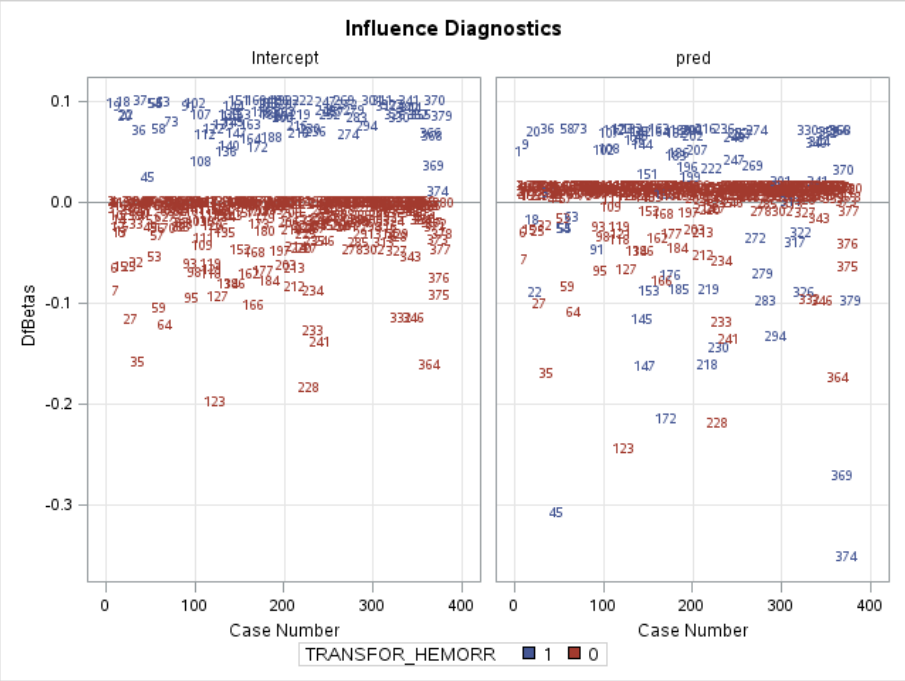
Diagnóstico do modelo ajustado. - Imputação 4

Figura 51 – Diagnóstico de observações influentes.



Fonte: Feita no SAS.

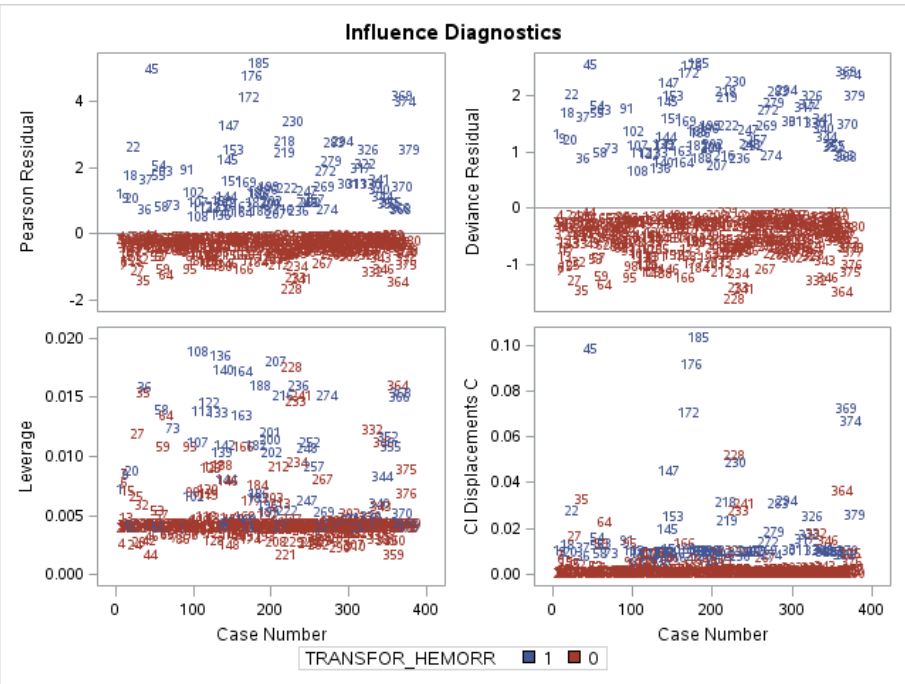
Figura 52 – Diagnóstico de observações influentes.



Fonte: Feita no SAS.

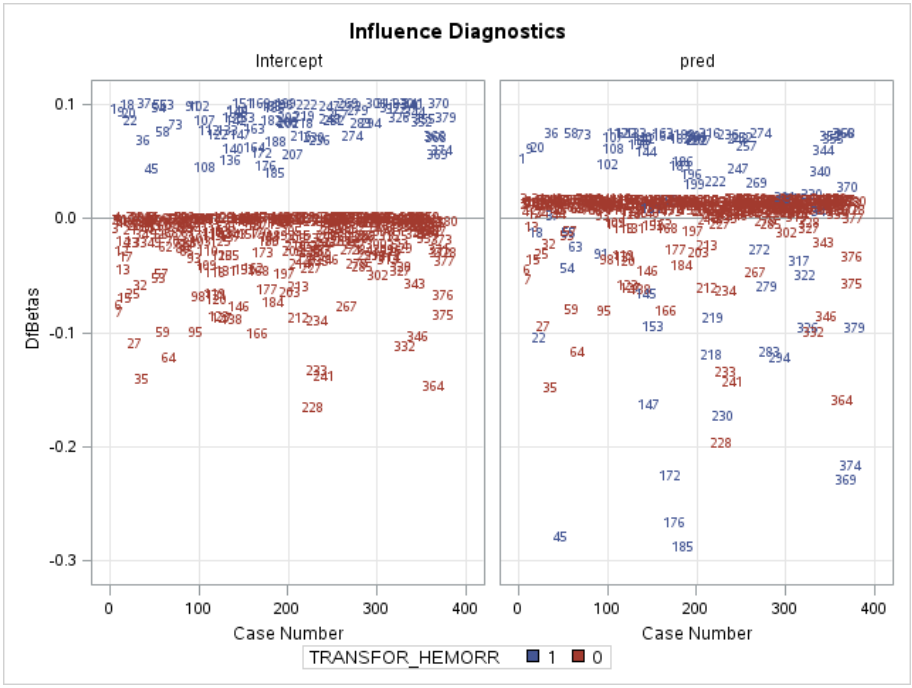
Diagnóstico do modelo ajustado. - Imputação 5

Figura 53 – Diagnóstico de observações influentes.



Fonte: Feita no SAS.

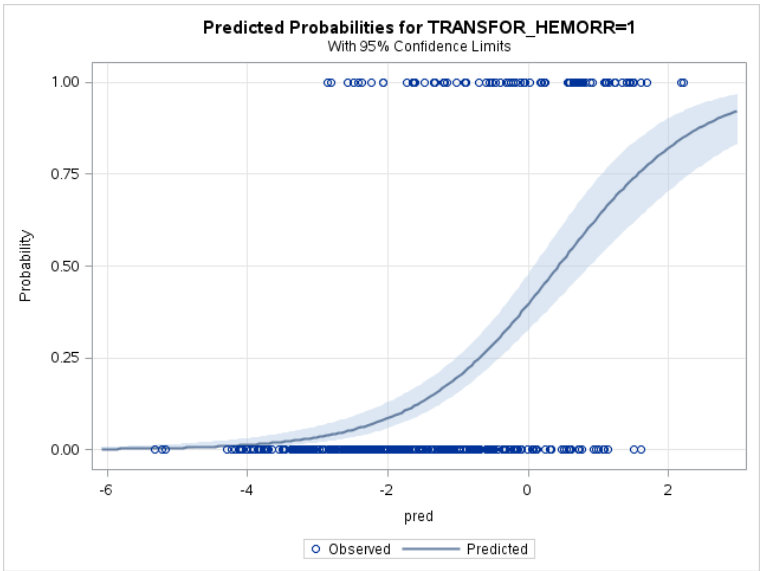
Figura 54 – Diagnóstico de observações influentes.



Fonte: Feita no SAS.

Predição do modelo ajustado - Imputação 1

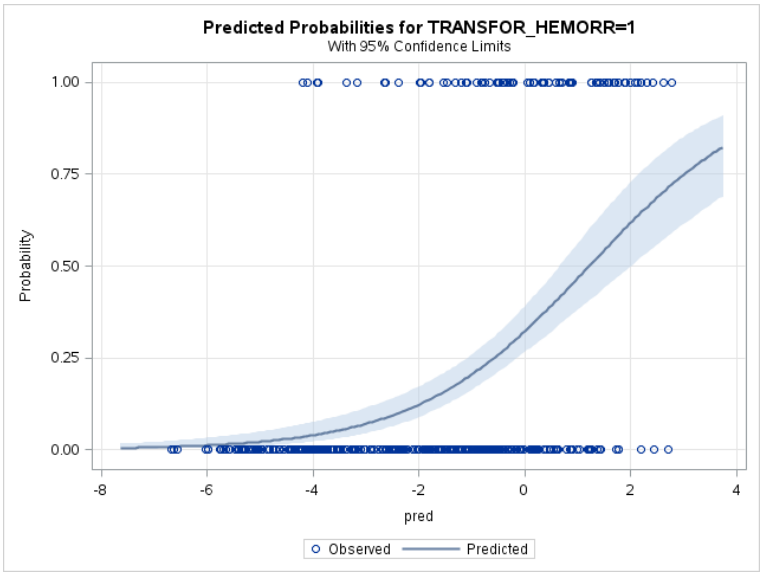
Figura 55 – Predição para o modelo gerado na primeira imputação.



Fonte: Feita no SAS.

Predição do modelo ajustado - Imputação 2

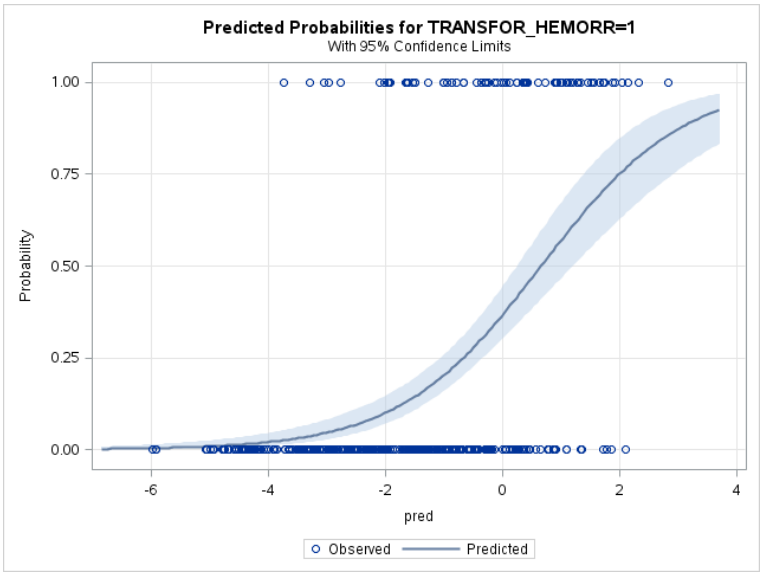
Figura 56 – Predição para o modelo gerado na segunda imputação.



Fonte: Feita no SAS.

Predição do modelo ajustado - Imputação 3

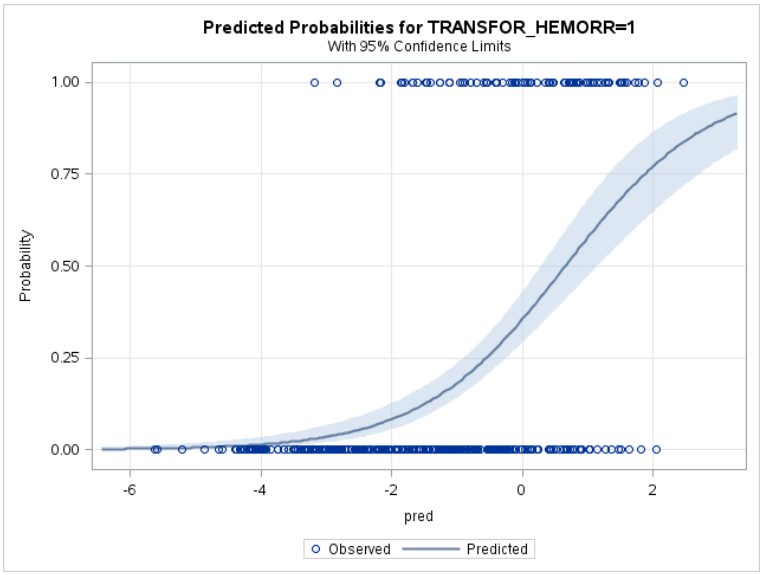
Figura 57 – Predição para o modelo gerado na terceira imputação.



Fonte: Feita no SAS.

Predição do modelo ajustado - Imputação 4

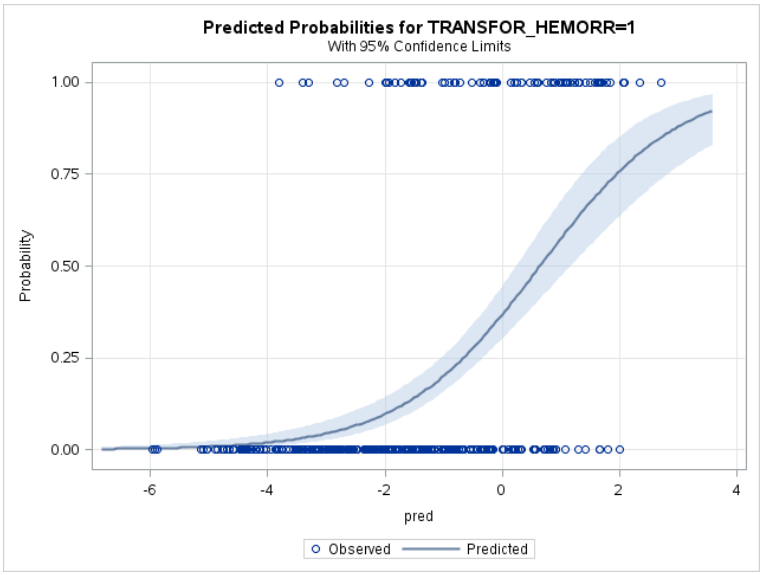
Figura 58 – Predição para o modelo gerado na quarta imputação.



Fonte: Feita no SAS.

Predição do modelo ajustado - Imputação 5

Figura 59 – Predição para o modelo gerado na quinta imputação.



Fonte: Feita no SAS.