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

Comparação de funções de ligação em modelos de regressão para respostas binárias com dados desbalanceados

Fabianna Akari Akinaga

Trabalho de Conclusão de Curso

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

Comparação de funções de ligação em modelos de regressão para respostas binárias com dados desbalanceados

Fabianna Akari Akinaga
Orientador: Gustavo Henrique de Araujo Pereira

Trabalho de Conclusão de Curso apresentado como parte dos requisitos para obtenção do título de Bacharel em Estatística.

São Carlos
Dezembro de 2025

Federal University of São Carlos
Exact and Technology Sciences Center
Department of Statistics

Comparison of Link Functions in Regression Models for Binary Responses with Imbalanced Data

Fabianna Akari Akinaga
Advisor: Gustavo Henrique de Araujo Pereira

Bachelors dissertation submitted to the Department of Statistics, Federal University of São Carlos - DEs-UFSCar, in partial fulfillment of the requirements for the degree of Bachelor in Statistics.

São Carlos
December 2025

Fabianna Akari Akinaga

Comparação de funções de ligação em modelos de regressão para respostas binárias com dados desbalanceados

Este exemplar corresponde à redação final do trabalho de conclusão de curso devidamente corrigido e defendido por Fabianna Akari Akinaga e aprovado pela banca examinadora.

Aprovado em 4 de dezembro de 2025

Banca Examinadora:

- Gustavo Henrique de Araujo Pereira (Orientador)
- Maria Sílvia de Assis Moura
- Michel Helcias Montoril

Dedico este trabalho à minha família e aos meus amigos, que estiveram ao meu lado durante toda a minha trajetória.

Agradecimentos

À minha família, deixo meu sincero agradecimento pelo apoio constante, pela confiança e pelo incentivo desde a decisão de mudar de cidade para cursar a graduação. A compreensão diante da distância, o apoio nos momentos de adaptação e o incentivo diário foram fundamentais para que eu conseguisse seguir em frente e concluir esta etapa tão importante.

Aos meus amigos, agradeço por terem sido apoio e acolhimento ao longo dessa trajetória. Em uma cidade nova, a convivência, as conversas e a presença de vocês tornaram o processo de adaptação mais leve e a rotina acadêmica mais divertida. O companheirismo e o incentivo de cada um fizeram toda a diferença ao longo da graduação.

Agradeço ao meu orientador, pela orientação, disponibilidade e contribuições ao longo do desenvolvimento deste trabalho.

Aos professores, agradeço pelos ensinamentos e pela contribuição para minha formação acadêmica.

Por fim, agradeço a todos que, direta ou indiretamente, contribuíram para o meu crescimento pessoal e profissional ao longo dessa trajetória.

Resumo

Este trabalho tem como objetivo comparar diferentes funções de ligação aplicadas a Modelos Lineares Generalizados (MLG) para variáveis resposta binárias, especialmente em contextos com dados desbalanceados. Foram avaliadas funções de ligação tradicionais, como Logito, Probito e Cloglog, bem como extensões generalizadas baseadas nas funções Potência e Potência Reversa. O interesse em conduzir este estudo surge das afirmações presentes na literatura de que ligações assimétricas tendem a apresentar desempenho preditivo superior ao de funções simétricas em cenários com forte desbalanceamento. Assim, buscou-se investigar, de forma sistemática, se essa vantagem realmente se confirma na prática. Para isso, foram conduzidos estudos de simulação de Monte Carlo e aplicações em vários conjuntos de dados reais, permitindo observar na prática como essas funções se comportam sob diferentes graus de desbalanceamento. As análises permitiram avaliar, em diversos contextos, se a flexibilidade introduzida pelos parâmetros adicionais dessas ligações resulta em ganhos relevantes de desempenho ou estabilidade. De maneira geral, os resultados mostraram que as diferentes funções de ligação, tradicionais e generalizadas, apresentaram valores da Área sob a Curva ROC (AUC) muito próximos tanto nas simulações quanto nas aplicações práticas, mantendo padrões semelhantes de desempenho mesmo diante de mudanças no grau de desbalanceamento, na distribuição das covariáveis ou no número de preditores.

Palavras-chave: *Área sob a curva ROC*; *classificação binária*; *dados desbalanceados*; *funções de ligação*; *simulação*; *modelos lineares generalizados.*

Abstract

This work aims to compare different link functions applied to Generalized Linear Models (GLMs) for binary response variables, especially in contexts with unbalanced data. Traditional link functions, such as Logit, Probit, and Cloglog, as well as generalized extensions based on Power and Reverse Power functions, were evaluated. The interest in conducting this study arises from the claims in the literature that asymmetric links tend to exhibit superior predictive performance compared to symmetric functions in scenarios with strong imbalance. Accordingly, this work sought to systematically investigate whether such an advantage is indeed observed in practice. To this end, Monte Carlo simulation studies and applications to various real datasets were conducted, allowing observation in practice of how these functions behave under different degrees of imbalance. The analyses allowed us to evaluate, in various contexts, whether the flexibility introduced by the additional parameters of these links results in relevant gains in performance or stability. In general, the results showed that the different link functions, both traditional and generalized, presented very similar Area Under the ROC Curve (AUC) values in both simulations and practical applications, maintaining similar performance patterns even when faced with changes in the degree of imbalance, the distribution of covariates, or the number of predictors.

Keywords: *Area under the ROC curve; binary classification; unbalanced data; link functions; simulation; generalized linear models..*

Lista de Figuras

Lista de Tabelas

Sumário

Capítulo 1

Introdução

Os Modelos Lineares Generalizados (MLG) são uma classe de modelos estatísticos que ampliam os modelos lineares tradicionais, permitindo a análise de uma variedade mais ampla de tipos de dados e distribuições de resposta. Introduzidos por Nelder e Wedderburn (1972), os MLG são projetados para modelar a relação entre uma variável resposta e um conjunto de variáveis preditoras, possibilitando que a variável resposta siga distribuições da família exponencial. Essa característica torna os MLG adequados para uma gama diversificada de dados, incluindo variáveis binárias, contagens e variáveis contínuas assimétricas.

Um exemplo notável de aplicação dos MLG é na análise de dados binários, como a presença ou ausência de uma doença em um grupo de indivíduos. Por exemplo, em um estudo sobre diabetes, a variável resposta pode indicar se um paciente é diabético (1) ou não (0), e o modelo pode ser ajustado utilizando uma distribuição binomial para estimar a probabilidade de diabetes com base em variáveis preditoras, como idade, índice de massa corporal e histórico familiar. Outro caso é a análise de dados de marketing, em que a variável resposta pode representar se um cliente comprou um produto (1) ou não (0). Nesse cenário, os MLG podem ser utilizados para modelar a probabilidade de compra em função de fatores como renda, idade e comportamento de compra anterior.

Esses modelos permitem que a relação entre a variável resposta e as variáveis preditoras seja modelada de forma flexível, utilizando diferentes funções de ligação para transformar a média da variável resposta em uma escala que pode ser linearmente modelada. A função de ligação Logito é a função canônica (Cordeiro *et al.*, 2024) para a distribuição binomial e é frequentemente utilizada em modelos de regressão para respostas binárias (McCullagh, 2019). Essa função é valorizada por sua interpretação intuitiva e suas propriedades

estatísticas favoráveis, como a capacidade de lidar com a heterocedasticidade e a não linearidade (Agresti, 2015).

No entanto, em contextos de dados desbalanceados, em que uma classe da resposta é significativamente mais frequente que a outra, a eficácia da função Logito pode ser questionada. Dados desbalanceados são aqueles em que a proporção entre as classes da variável dependente é desigual. Por exemplo, em um conjunto de dados em que 90% das observações pertencem à classe 0 e apenas 10% à classe 1, o modelo pode se tornar tendencioso em prever a classe majoritária, resultando em baixa sensibilidade para a classe minoritária (He e Garcia, 2009). Nesse contexto, a escolha da função de ligação e a abordagem de modelagem se tornam ainda mais críticas, pois podem impactar a capacidade do modelo de prever corretamente a classe minoritária.

A literatura sugere que funções de ligação assimétricas, como a Cloglog, podem ser mais adequadas para dados desbalanceados, pois são projetadas para lidar com a assimetria nas distribuições de probabilidade (Chen *et al.*, 1999). A função de ligação Logito e também a Probito, por serem simétricas, podem não ser a melhor opção para modelar respostas binárias para dados desbalanceados.

Embora a literatura sugira que funções de ligação assimétricas são mais adequadas para modelar dados binários desbalanceados (Huayanay *et al.*, 2019), teoricamente, isso não é necessariamente verdade. O motivo é que se o intercepto do modelo for, em módulo, grande, as funções de ligação simétricas podem, em princípio, ser adequadas para esse caso. A dúvida é se de fato, para dados simulados e reais (especialmente para estes últimos), as funções de ligação simétricas podem levar a um melhor ajuste de dados binários desbalanceados do que as funções de ligação assimétricas.

Nesse contexto, o objetivo principal deste trabalho é comparar diferentes funções de ligação simétricas e assimétricas em MLG com resposta binária e dados desbalanceados. Em específico, comparar o desempenho de funções de ligação tradicionais (Logito, Probito, Cloglog) e suas versões generalizadas: Potência e Reversa da Potência, nesse contexto de dados binários (Lemonte e Bazán, 2018). Para isso, aplicaremos estas metodologias a diversos bancos de dados reais e estudos de simulação em que os mesmos serão analisados através do *software* R (R Core Team, 2022).

Assim, este trabalho está organizado da seguinte maneira. No Capítulo 2, apresentamos a estrutura e os componentes de um modelo linear generalizado, com enfoque principal para dados binários, descrevendo as funções de ligação mais utilizadas e outras

propostas mais recentemente, procedimentos para estimação dos parâmetros pelo método da máxima verossimilhança e também análise de diagnóstico. No Capítulo 3 abordamos a implementação de toda a metodologia estudada e a aplicação a um conjunto de dados reais utilizando o software R. Em seguida, no Capítulo 4 apresentamos mais duas aplicações em conjuntos reais para comparar as funções de ligação em variados níveis de desbalanceamento. Já no Capítulo 5, realizamos estudos de simulação de Monte Carlo para comparar a performance das diferentes funções de ligação em diferentes cenários. Por fim, no Capítulo 6 temos uma síntese final sobre as conclusões que foram obtidas a cerca do objetivo em estudo com base na metodologia proposta.

Capítulo 2

Modelos Lineares Generalizados

Neste capítulo são apresentadas na Seção 2.1 a estrutura e os componentes de um modelo linear generalizado. Na Seção 2.2 é abordada a distribuição binomial apresentando sua densidade e logaritmo da função de verossimilhança em um MLG com resposta binária. Em seguida, a Seção 2.3 explora as funções de ligação mais comuns para modelos binários, além de descrever funções de ligação alternativas (Potência e Reversa da Potência das funções Logito, Probito e Cloglog). A Seção 2.4 discute a estimação dos parâmetros do modelo, enquanto a análise de diagnóstico é abordada na Seção 2.5. Por fim, a Seção 2.6 descreve como avaliar o poder preditivo do modelo utilizando a Curva ROC.

2.1 Especificação do Modelo

Os Modelos Lineares Generalizados (MLG) surgiram da necessidade de unificar diversos modelos estatísticos que, até 1972, eram tratados separadamente. Foi nesse ano que Nelder e Wedderburn (1972) propuseram a estrutura dos MLG, ampliando os modelos de regressão linear múltipla tradicional. A grande inovação dos MLG é que eles permitem o uso de uma variedade muito maior de distribuições para a variável resposta, desde que pertençam à família exponencial linear, e incorporam uma função de ligação para conectar a média da resposta aos preditores. Essa abordagem unificada engloba modelos como regressão linear clássica, regressão logística e regressão de Poisson em uma única estrutura teórica (McCullagh, 2019).

Sejam $y_1, y_2, \ldots, y_n$ variáveis aleatórias independentes. Assume-se que todos os y_i têm distribuição que pertence à família exponencial linear, na qual os parâmetros são θ_i, para $i = 1, \ldots, n$, e ϕ. A função de probabilidade é dada por:

$$f(y_i, \theta_i, \phi) = \exp\left[\frac{y_i\theta_i - b(\theta_i)}{\phi} + c(y_i, \phi)\right], \tag{2.1}$$

em que $b(\cdot)$ e $c(\cdot)$ são funções conhecidas, θ_i é o parâmetro canônico, que está diretamente ligado à média da distribuição, e ϕ é o parâmetro de dispersão, que controla a variabilidade dos dados e pode ser conhecido ou estimado. Dessa forma, temos que $E(y_i) = b'(\theta_i)$ e $Var(y_i) = \phi^{-1}b''(\theta_i)$.

A especificação do MLG se faz através de um modelo de ligação entre a média μ_i e as covariáveis, ou seja:

$$g(\mu_i) = \mathbf{x}_i^\top \boldsymbol{\beta}, \tag{2.2}$$

em que $\mu_i = E(y_i) = b'(\theta_i)$, $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_k)^\top$ são parâmetros desconhecidos a serem estimados e $\mathbf{x}_i = (1, x_{i1}, x_{i2}, \ldots, x_{ik})^\top$ são constantes que representam os valores das variáveis preditoras. A função g é uma função de ligação estritamente monótona e pelo menos duplamente diferenciável. Podemos representar também a Equação (2.2) como $g(\mu_i) = \eta_i$, em que $\eta_i = \mathbf{x}_i^\top \boldsymbol{\beta}$ é denominado preditor linear.

Em resumo, os MLG são caracterizados pela seguinte estrutura:

- **Componente Aleatória:** Representado por um conjunto de variáveis independentes $y_1, y_2, \ldots, y_n$ provenientes de uma mesma distribuição da família exponencial linear com parâmetros θ_i e ϕ, $i = 1, \ldots, n$.

- **Componente Sistemática:** Refere-se ao preditor linear, denotado por $\eta_i = \mathbf{x}_i^\top \boldsymbol{\beta}$. Ele é uma combinação linear do vetor de coeficientes $\boldsymbol{\beta}$ e das variáveis preditoras $\mathbf{x}_i$, sendo linear nos parâmetros.

- **Função de Ligação:** Relaciona o componente aleatório ao sistemático, sendo $g(\cdot)$ uma função estritamente monótona e duplamente diferenciável.

O processo de ajuste de um MLG envolve três etapas essenciais: escolher a distribuição adequada para a variável resposta, selecionar as variáveis preditoras mais relevantes e determinar a função de ligação ideal. Essas escolhas, feitas em conjunto, guiam a estimação dos parâmetros do modelo.

Neste trabalho, focaremos em um caso particular de MLG: a regressão para variável resposta binária. Nesses cenários, a variável dependente assume apenas dois resultados possíveis (por exemplo, sucesso/fracasso, presente/ausente), e o principal interesse

é modelar a probabilidade de ocorrência de um desses eventos em função das variáveis preditoras, um método amplamente utilizado, em diversas áreas do conhecimento (Paula, 2004).

2.2 Distribuição Binomial

A distribuição binomial modela o número de sucessos, Y, que ocorrem em uma série de n ensaios independentes. Cada um desses ensaios, conhecidos como ensaios de Bernoulli, possui apenas dois desfechos possíveis: sucesso ou fracasso. Se denotarmos a probabilidade de sucesso em um único ensaio como μ, a função de probabilidade da distribuição binomial é expressa como:

$$P(Y = y) = \binom{n}{y} \mu^y (1-\mu)^{n-y} I_{\{0,1,\ldots,n\}}(y). \tag{2.3}$$

Nessa expressão, $I_{\{0,1,\ldots,n\}}(y)$ é uma função indicadora, garantindo que y esteja dentro do conjunto que representa o número de sucessos. Para essa distribuição, a média e a variância de Y são, respectivamente, $E(Y) = n\mu$ e $Var(Y) = n\mu(1-\mu)$ (Magalhães e De Lima, 2011).

No contexto deste trabalho, o interesse reside especificamente em cenários em que a variável resposta é binária, o que corresponde a um único ensaio ($n = 1$) para cada unidade amostral. Esta situação particular da binomial é conhecida como distribuição de Bernoulli. Quando $Y \sim \text{Bernoulli}(\mu)$, sua função de probabilidade no formato da família exponencial linear é dada por:

$$P(Y = y) = \exp\left[y \log\left(\frac{\mu}{1-\mu} \right) + \log(1-\mu) \right] I_{\{0,1\}}(y). \tag{2.4}$$

Essa representação é crucial, pois facilita a integração da distribuição de Bernoulli na estrutura dos MLG, permitindo que sua probabilidade de sucesso seja relacionada aos preditores por meio de uma função de ligação.

2.3 Funções de Ligação para Dados Binários

A função de ligação é um componente crucial nos Modelos Lineares Generalizados (MLG), atuando como a ponte entre a média da variável resposta (μ_i) e o preditor li-

near ($\eta_i = \mathbf{x}_i^\top \boldsymbol{\beta}$). Já a função de ligação canônica é aquela que estabelece uma relação direta entre a média da variável resposta (μ_i) e o parâmetro canônico (θ_i) de uma distribuição pertencente à família exponencial, de forma que o preditor linear (η_i) seja igual ao parâmetro canônico (Agresti, 2015), ou seja,

$$\eta_i = g(\mu_i) = \theta_i.$$

No contexto de modelos com resposta binária, em que μ_i representa a probabilidade de sucesso (variando entre 0 e 1), a função de ligação tem o papel de transformar essa probabilidade para uma escala contínua e irrestrita $(-\infty, +\infty)$, em que a relação linear com as variáveis preditoras pode ser estabelecida (McCullagh, 2019). A escolha da função de ligação não apenas define a forma da curva de resposta do modelo, mas também influencia a interpretação dos coeficientes de regressão.

2.3.1 Funções de Ligação Tradicionais

Existem três funções de ligação que são classicamente empregadas para modelagem de dados binários, cada uma derivada de pressupostos subjacentes diferentes e oferecendo características distintas.

- **Função Logito (*logit*):** É a função de ligação canônica para a distribuição binomial, sendo simétrica em torno de $\mu = 0,5$. Sua popularidade advém da interpretabilidade dos coeficientes como log-razão de chances, ou, após exponenciação, como razão de chances (*odds ratio* em inglês). De acordo com Collett (1991), a função Logito é definida por:

$$\mathrm{g}(\mu_i) = \log\left(\frac{\mu_i}{1-\mu_i}\right) = \eta_i. \tag{2.5}$$

A partir dessa expressão, as chances de sucesso para a i-ésima observação podem ser obtidas exponenciando ambos os lados da equação:

$$\frac{\mu_i}{1-\mu_i} = \exp(\mathbf{x}_i^\top \boldsymbol{\beta}). \tag{2.6}$$

Essa razão compara a probabilidade de um evento ocorrer com a probabilidade de ele não ocorrer, que representa a chance de ocorrência de sucesso na variável resposta

para a i-ésima observação.

Para interpretar o modelo de regressão logística, utilizamos a razão de chances. Sendo assim, considere $x_{ij} = l$, fixando-se as outras variáveis, temos então que

$$\left.\frac{\mu_i}{1-\mu_i}\right|_{x_{ij}=l} = \exp(\beta_0 + \beta_1 x_{i1} + \cdots + l \times \beta_j + \cdots + \beta_k x_{ik}), \tag{2.7}$$

enquanto que, se $x_{ij} = l + 1$,

$$\left.\frac{\mu_i}{1-\mu_i}\right|_{x_{ij}=l+1} = \exp(\beta_0 + \beta_1 x_{i1} + \cdots + (l+1) \times \beta_j + \cdots + \beta_k x_{ik}). \tag{2.8}$$

A razão de chances (OR) para este caso é definida por (2.8) / (2.7). Portanto,

$$\text{OR} = \frac{\exp(\beta_0 + \beta_1 x_{i1} + \cdots + (l+1) \times \beta_j + \cdots + \beta_k x_{ik})}{\exp(\beta_0 + \beta_1 x_{i1} + \cdots + l \times \beta_j + \cdots + \beta_k x_{ik})} = \exp(\beta_j), \tag{2.9}$$

em que $i = 1, \ldots, n$.

Sendo assim, podemos interpretar a resultante $\exp(\beta_j)$ como o valor pelo qual é multiplicado a chance de ocorrência de sucesso na variável resposta quando x_{ij} tem acréscimo de uma unidade, mantendo as demais variáveis preditoras fixadas (Brolo, 2019).

- **Função Probito (*probit*):** Baseada na função de distribuição acumulada (FDA) da distribuição normal padrão (Φ), a função Probito também é simétrica. É frequentemente utilizada quando se assume que a ocorrência do evento é governada por um limiar em uma variável latente subjacente com distribuição normal (Collett, 1991). A função de ligação Probito é definida como:

$$g(\mu_i) = \Phi^{-1}(\mu_i) = \eta_i. \tag{2.10}$$

A função inversa é $\mu_i = \Phi(\eta_i)$.

- **Função Complemento Log-Log (*cloglog*):** Diferentemente do Logito e Probito, esta função de ligação é assimétrica. Ela é particularmente útil em situações em que a probabilidade de ocorrência de um evento aumenta de forma assimétrica, como em modelos de sobrevivência ou quando o evento é raro (Collett, 1991). A função

de ligação Complemento Log-Log é definida como:

$$g(\mu_i) = \log(-\log(1-\mu_i)) = \eta_i. \tag{2.11}$$

Sua função inversa é $\mu_i = 1 - \exp(-\exp(\eta_i))$.

2.3.2 Funções de Ligação Generalizadas: Potência e Potência Reversa

Para aumentar a flexibilidade das funções de ligação tradicionais e permitir que o modelo se ajuste a relações mais complexas e assimétricas em dados de resposta binária, Bazán *et al.* (2014) propôs transformações em funções de ligações usuais que incorporam um parâmetro de potência λ. A ideia subjacente a essas funções de ligação generalizadas é que o parâmetro λ (geralmente $\lambda > 0$) atua sobre a função de distribuição acumulada (CDF) subjacente $G(\eta_i)$, que forma a base das funções de ligação Logito, Probito, Cloglog, permitindo que a curva de resposta se ajuste a padrões não-lineares específicos.

As funções de ligação de potência são categorizadas em duas formas principais:

2.3.3 Funções de Ligação de Potência

Esta classe de funções de ligação assimétricas é definida pela transformação da CDF $G(\eta_i)$ por um parâmetro de potência λ. A probabilidade de sucesso μ_i é dada por:

$$\mu_i = F_P(\eta_i) = G(\eta_i)^{\lambda}, \tag{2.12}$$

em que $G(\eta_i)$ representa a CDF de uma distribuição simétrica subjacente (Logística para *logit*, Normal para *probit*) ou a CDF da distribuição Gumbel para Cloglog. Quando $\lambda = 1$, a transformação não altera $G(\eta_i)$, fazendo com que a função de ligação generalizada retorne à sua forma tradicional simétrica. Valores de $\lambda \neq 1$ alteram a forma da curva de resposta, introduzindo assimetria (Lemonte e Bazán, 2018).

2.3.4 Funções de Ligação de Potência Reversa

Esta é a segunda classe de funções de ligação assimétricas, em que a transformação envolve um parâmetro de potência λ aplicado à CDF do valor oposto do preditor linear

$G(-\eta_i)$, e então subtraída de 1. A probabilidade de sucesso μ_i é dada por:

$$\mu_i = F_{RP}(\eta_i) = 1 - G(-\eta_i)^{\lambda}. \tag{2.13}$$

Similarmente, se $\lambda = 1$, essa transformação retorna a função de ligação simétrica tradicional. O parâmetro $\lambda > 0$ em ambas as classes (2.12) e (2.13) caracteriza a assimetria das funções de ligação resultantes.

Ao combinar as funções de ligação tradicionais (Logito, Probito, Cloglog) com essas transformações de Potência e Potência Reversa, obtemos um conjunto expandido de funções de ligação capazes de modelar uma gama mais ampla de relações.

- **Logito Potência (*power logit*):** Aplica a transformação de potência à CDF Logística.

$$g(\mu_i, \lambda) = \log\left(\frac{\mu_i^{1/\lambda}}{1 - \mu_i^{1/\lambda}}\right) = \eta_i.$$

  Sua função inversa é: $\mu_i = \left(\frac{e^{\eta_i}}{1+e^{\eta_i}}\right)^{\lambda}$.

- **Logito Potência Reversa (*reverse power logit*):** Aplica a transformação reversa da potência à CDF Logística (negada).

$$g(\mu_i, \lambda) = -\log\left(\frac{(1-\mu_i)^{1/\lambda}}{1 - (1-\mu_i)^{1/\lambda}}\right) = \eta_i.$$

  Sua função inversa é: $\mu_i = 1 - \left(\frac{1}{1+e^{\eta_i}}\right)^{\lambda}$.

- **Probito Potência (*power probit*):** Aplica a transformação de potência à CDF Normal (Bazán *et al.*, 2014).

$$g(\mu_i, \lambda) = \Phi^{-1}(\mu_i^{1/\lambda}) = \eta_i.$$

  Sua função inversa é: $\mu_i = (\Phi(\eta_i))^{\lambda}$.

- **Probito Potência Reversa (*reverse power probit*):** Aplica a transformação reversa da potência à CDF Normal.

$$g(\mu_i, \lambda) = -\Phi^{-1}((1-\mu_i)^{1/\lambda}) = \eta_i.$$

Sua função inversa é: $\mu_i = 1 - (\Phi(-\eta_i))^{\lambda}$.

- **Cloglog Potência (*power cloglog*):** Aplica a transformação de potência à CDF Gumbel.

$$g(\mu_i, \lambda) = \log(-\log(1 - \mu_i^{1/\lambda})) = \eta_i.$$

Sua função inversa é: $\mu_i = \left(1 - e^{-e^{\eta_i}}\right)^{\lambda}$.

- **Cloglog Potência Reversa (*reverse power cloglog*):** Aplica a transformação reversa da potência à CDF Gumbel.

$$g(\mu_i, \lambda) = -\log(-\log(1 - (1 - \mu_i)^{1/\lambda})) = \eta_i.$$

Sua função inversa é: $\mu_i = 1 - \left(1 - e^{-e^{-\eta_i}}\right)^{\lambda}$.

A investigação do comportamento dessas funções de ligação generalizadas, com a estimação do parâmetro k, é fundamental para encontrar o modelo que melhor se ajusta aos dados, especialmente em situações em que as relações não-lineares ou assimetrias não são adequadamente capturadas pelas funções de ligação tradicionais, na visão da literatura (Bazán *et al.*, 2017).

2.4 Ajuste de um Modelo com Resposta Binária

A estimação do vetor de parâmetros $\boldsymbol{\beta}$ em Modelos Lineares Generalizados (MLG) é tradicionalmente realizada utilizando o método de máxima verossimilhança (Nelder e Wedderburn, 1972). Para simplificar os cálculos matemáticos, e devido ao fato de que ambas as abordagens conduzem ao mesmo estimador, é comum maximizar o logaritmo da função de verossimilhança em vez da própria função de verossimilhança.

Para modelos de regressão com dados de resposta binária, o logaritmo da função de verossimilhança para n observações independentes é expresso como:

$$l(\boldsymbol{\mu}; \mathbf{y}) = \sum_{i=1}^{n} \left\{ y_i \log\left(\frac{\mu_i}{1 - \mu_i}\right) + \log(1 - \mu_i) \right\}. \tag{2.14}$$

Ao tentar maximizar esta função de log-verossimilhança, derivando-a em relação aos parâmetros do modelo e igualando as derivadas a zero, percebe-se que o sistema de equações resultante não possui uma solução analítica explícita. Isso significa que não

há uma fórmula direta para obter os estimadores de $\boldsymbol{\beta}$.

Diante da ausência de uma solução analítica, a maximização é obtida por métodos de otimização numérica. A técnica de otimização mais amplamente empregada para MLG é o algoritmo de mínimos quadrados reponderados iterativos (McCullagh, 2019). Este método converge iterativamente para os estimadores de máxima verossimilhança, ajustando os pesos e os valores do preditor linear a cada passo até atingir a convergência.

2.4.1 Testes de Hipótese

Uma vez que os parâmetros do MLG são estimados, é possível realizar testes de hipótese para avaliar a significância estatística das variáveis preditoras ou de grupos de variáveis. Os testes mais comuns incluem:

- **Teste de Wald:** Este teste é o mais usado quando o interesse é avaliar a significância individual de cada coeficiente β_j, $j = 1, ..., p$. A hipótese nula é que $\beta_j = 0$, indicando que a variável preditora X_j não tem efeito significativo sobre a variável resposta. A estatística de Wald é calculada como $(\hat{\beta}_j/\widehat{SE(\hat{\beta}_j)})^2$, em que $\widehat{SE(\hat{\beta}_j)}$ é o erro padrão estimado de $\hat{\beta}_j$. Sob a hipótese nula, a estatística de Wald segue aproximadamente uma distribuição qui-quadrado com 1 grau de liberdade (Agresti, 2015).

- **Teste da Razão de Verossimilhanças (LR Test):** Este teste é utilizado para comparar a qualidade do ajuste de dois modelos aninhados (em que um é um subconjunto do outro). A estatística do teste é dada por $-2\log(L_0/L_1)$, em que L_0 é a verossimilhança do modelo nulo (menor) e L_1 é a verossimilhança do modelo alternativo (maior). Sob a hipótese nula de que o modelo mais simples é suficiente, a estatística segue aproximadamente uma distribuição qui-quadrado com graus de liberdade iguais à diferença no número de parâmetros entre os dois modelos (Dobson e Barnett, 2018). Esse teste é mais usado quando queremos testar mais de um parâmetro simultaneamente, o que é muito usado, por exemplo, quando temos covariáveis qualitativas com mais de dois níveis.

2.4.2 Análise de Diagnóstico

A análise de diagnóstico em Modelos Lineares Generalizados (MLG) é uma etapa crucial para avaliar a adequação do modelo ajustado aos dados. Embora não seja o

foco principal deste trabalho, é importante mencionar que ela envolve a verificação de suposições e a identificação de observações atípicas ou influentes (McCullagh, 2019).

Entre as ferramentas de diagnóstico, destacam-se a análise de resíduos, como os resíduos de Pearson e os resíduos deviance, que ajudam a identificar padrões não modelados, falta de ajuste ou heterocedasticidade. Além disso, gráficos de diagnóstico permitem visualizar a distribuição dos resíduos em relação aos preditores ou aos valores ajustados. A identificação de observações influentes, que podem distorcer as estimativas dos parâmetros, também é uma parte importante dessa análise. Métricas como a distância de Cook e a alavancagem são empregadas para detectar tais observações (Hosmer Jr *et al.*, 2013).

O objetivo dessa análise é assegurar que o modelo escolhido seja robusto e represente adequadamente a estrutura dos dados, fornecendo estimativas confiáveis e inferências válidas, mesmo que os detalhes dessas verificações não sejam aprofundados no presente trabalho. Maiores detalhes sobre esse assunto podem ser encontrados em Paula (2004).

2.5 Avaliação de Modelos para Dados Binários

A avaliação de modelos para dados binários requer métricas específicas que considerem a natureza classificatória da variável resposta. Ao invés de apenas avaliar a capacidade de previsão contínua, o foco recai sobre a capacidade do modelo de classificar corretamente as observações nas suas respectivas categorias.

2.5.1 Ponto de Corte e Matriz de Confusão

No caso binário, para classificar as observações, é necessário definir um ponto de corte (*threshold*) na probabilidade estimada pelo modelo. Se a probabilidade prevista for maior ou igual ao ponto de corte, a observação é classificada como 1; caso contrário, como 0. A escolha do ponto de corte é crucial, pois ela afeta o desempenho do modelo. Um ponto de corte comumente usado é 0,5, mas nem sempre é o ideal, especialmente em dados desbalanceados (Zou *et al.*, 2016).

A partir do ponto de corte, a performance do modelo pode ser resumida em uma matriz de confusão, que é uma tabela que compara as classes reais com as classes previstas pelo modelo:

Em que:

Tabela 2.1: Matriz de Confusão

	Previsto Positivo (1)	**Previsto Negativo (0)**
Real Positivo (1)	Verdadeiros Positivos (VP)	Falsos Negativos (FN)
Real Negativo (0)	Falsos Positivos (FP)	Verdadeiros Negativos (VN)

- **Verdadeiros Positivos (VP):** O modelo previu corretamente a classe positiva.
- **Falsos Negativos (FN):** O modelo previu erroneamente a classe negativa quando a real era positiva (erro tipo II).
- **Falsos Positivos (FP):** O modelo previu erroneamente a classe positiva quando a real era negativa (erro tipo I).
- **Verdadeiros Negativos (VN):** O modelo previu corretamente a classe negativa.

Note que cada uma das quatro caselas da matriz de confusão trazem, entre as n observações da amostra, a quantidade delas que está na respectiva situação.

2.5.2 Métricas de Desempenho Baseadas na Matriz de Confusão

Com base na matriz de confusão, várias métricas podem ser calculadas para avaliar o desempenho do modelo:

- **Acurácia:** Proporção de classificações corretas em relação ao total de observações.

$$Acurácia = \frac{VP + VN}{VP + VN + FP + FN}. \tag{2.15}$$

  A acurácia pode ser enganosa em dados desbalanceados, pois um modelo pode atingir alta acurácia simplesmente prevendo a classe majoritária (He e Garcia, 2009).

- **Precisão:** Proporção de verdadeiros positivos em relação ao total de previsões positivas.

$$Precisão = \frac{VP}{VP + FP}. \tag{2.16}$$

  A precisão é importante quando o custo de um falso positivo é alto (Davis e Goadrich, 2006).

- **Revocação/Sensibilidade:** Proporção de verdadeiros positivos em relação ao total de observações positivas reais.

$$Sensibilidade = \frac{VP}{VP + FN}. \tag{2.17}$$

A sensibilidade é crucial quando o custo de um falso negativo é alto (ex: diagnóstico de doenças).

- **Especificidade:** Proporção de verdadeiros negativos em relação ao total de observações negativas reais.

$$Especificidade = \frac{VN}{VN + FP}. \tag{2.18}$$

- **F1-Score:** Média harmônica da precisão e sensibilidade, que fornece uma medida equilibrada quando há um desbalanceamento entre as classes.

$$F1 - Score = 2 \times \frac{Precisão \times Sensibilidade}{Precisão + Sensibilidade}. \tag{2.19}$$

2.5.3 Área Sob a Curva ROC (AUC)

A curva ROC é uma das medidas mais importantes para verificar o desempenho de qualquer modelo de classificação. Quanto maior for a área sob a curva ROC, melhor é o poder preditivo do modelo utilizado (Fawcett, 2006). O gráfico da Curva ROC exibe a taxa de verdadeiros positivos (sensibilidade) versus a taxa de falsos positivos (1 - especificidade) para diferentes pontos de corte, ou seja, a Área Sob a Curva ROC (AUC) é uma métrica que resume o desempenho do modelo em todos os possíveis pontos de corte (Hanley e McNeil, 1982).

A principal vantagem da AUC é que ela não depende da escolha de um ponto de corte específico, o que a torna uma métrica robusta para comparar o desempenho geral de modelos, especialmente em cenários com dados desbalanceados (Hosmer Jr *et al.*, 2013). Um valor de AUC de 0,5 indica um desempenho aleatório (o modelo não é melhor que o acaso), enquanto um valor de 1,0 indica um classificador perfeito. Por essa razão, a AUC será a métrica principal utilizada neste trabalho para comparar as diferentes funções de ligação em dados binários desbalanceados, fornecendo uma avaliação mais abrangente da capacidade discriminatória dos modelos.

Capítulo 3

Aplicação em Dados Reais

Neste capítulo discutiremos a implementação da metodologia adotada e a aplicação a um banco de dados real.

3.1 Banco de dados

Para a comparação inicial das funções de ligação discutidas foi utilizado o conjunto de dados Indian Liver Patient Dataset (ILPD), obtido da base de dados *UCI Machine Learning Repository* (Ramana e Venkateswarlu, 2022). Este conjunto contém informações clínicas de 583 pacientes da região de Andhra Pradesh, Índia, com o objetivo de prever a presença ou ausência de doença hepática. As variáveis preditoras incluem características demográficas e exames bioquímicos hepáticos, enquanto a variável resposta, denominada *Selector*, é binária, indicando a presença (1) ou ausência (0) de doença hepática. As variáveis envolvidas neste estudo estão descritas na Tabela 3.1:

Tabela 3.1: Descrição das variáveis do conjunto de dados

Nome da Variável	**Descrição**
Age	Idade do paciente.
Gender	Gênero do paciente.
TB	Nível total de bilirrubina no sangue.
DB	Nível de bilirrubina direta no sangue.
Alkphos	Nível de fosfatase alcalina.
Sgpt	Nível da enzima alanina aminotransferase.
Sgot	Nível da enzima aspartato aminotransferase.
TP	Concentração total de proteínas no sangue.
ALB	Nível de albumina no sangue.
A/G Ratio	Razão entre albumina e globulina.
Selector	Presença ou ausência de doença hepática (variável resposta).

3.2 Análise Descritiva

Nesta seção apresentaremos uma análise descritiva dos dados descritos na Seção 3.1. Durante a etapa inicial de tratamento dos dados, foi identificado que a variável referente à razão entre albumina e globulina apresentava quatro observações com valores ausentes. Considerando o pequeno percentual de casos com dados faltantes e seguindo as recomendações de Salgado *et al.* (2016), optou-se por excluir essas observações por meio de remoção por lista completa (*listwise deletion*). A base de análise passou a conter, portanto, 579 observações completas.

A distribuição da variável resposta apresentou um desbalanceamento expressivo entre as classes. Do total de pacientes, 412 (71,2%) apresentaram diagnóstico de doença hepática, enquanto apenas 167 (28,8%) não apresentaram a condição. A Tabela 3.2 apresenta as frequências absolutas e relativas das categorias da variável resposta.

Tabela 3.2: Distribuição de frequências da variável resposta

Classe	**Frequência**	**Porcentagem**
Com Doença Hepática (1)	412	71,2%
Sem Doença Hepática (0)	167	28,8%

Essa característica é ilustrada visualmente na Figura 3.1, que mostra um gráfico de barras destacando a predominância da classe positiva (pacientes com doença hepática).

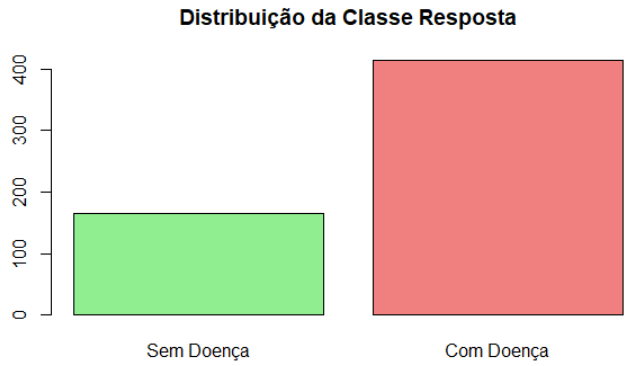


Figura 3.1: Frequência absoluta das classes da variável resposta

Em relação ao gênero, as Figuras 3.2 e 3.3 mostram que a presença da doença é ligeiramente mais comum entre os homens (73,58%) do que entre as mulheres (65%). Consequentemente, a proporção de indivíduos sem a doença é um pouco maior entre as mulheres (35%) em comparação aos homens (26,42%). Embora a diferença entre os

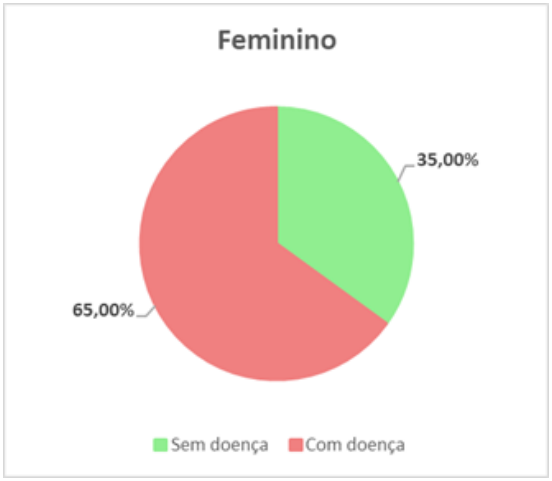


Figura 3.2: Distribuição percentual de mulheres com e sem a doença

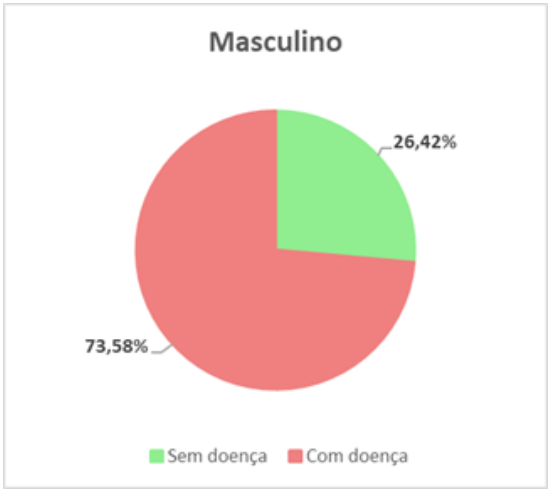


Figura 3.3: Distribuição percentual de homens com e sem a doença

gêneros não seja muito expressiva, os dados indicam uma possível associação entre gênero e ocorrência da doença, sugerindo que esse fator pode influenciar, ainda que de forma discreta, sua distribuição.

A análise univariada das variáveis numéricas está sumarizada na Tabela 3.3. Foram calculadas medidas de tendência central e dispersão, como média, mediana, desvio padrão, além dos valores mínimo e máximo observados para cada variável.

Tabela 3.3: Estatísticas descritivas das variáveis contínuas

Variável	**Média**	**DP**	**Mínimo**	**1º Q**	**Mediana**	**3º Q**	**Máximo**
Age	44,78	16,22	4,00	33,00	45,00	58,00	90,00
TB	3,32	6,23	0,40	0,80	1,00	2,60	75,00
DB	1,49	2,82	0,10	0,20	0,30	1,30	19,70
Alkphos	291,40	243,56	63,00	175,50	208,00	298,00	2110,00
Sgpt	81,13	183,18	10,00	23,00	35,00	61,00	2000,00
Sgot	110,40	289,85	10,00	25,00	42,00	87,00	4929,00
TP	6,48	1,08	2,70	5,80	6,60	7,20	9,60
ALB	3,14	0,79	0,90	2,60	3,10	3,80	5,50
A/G Ratio	0,95	0,32	0,30	0,70	0,93	1,10	2,80

Ao analisar o perfil dos dados, observa-se que as variáveis Bilirrubina Total (TB), Bilirrubina Direta (DB), Sgpt, Sgot e Fosfatase Alcalina (Alkphos) apresentam distribuições acentuadamente assimétricas à direita, com uma concentração de valores extremos no grupo com doença. Conforme ilustrado nos boxplots apresentados no Apêndice A, essa característica é evidenciada pela presença de muitos outliers, com valores máximos que chegam a 75,00 para TB e 4929,00 para Sgot, que são muito distantes da média, o que sugere que, em alguns pacientes, a condição clínica pode se manifestar de forma bastante

acentuada. Optou-se por manter esses valores na análise, pois eles podem ser indicativos de estados clínicos mais severos.

As concentrações de Bilirrubina Total (TB) e Direta (DB) mostram-se visivelmente mais elevadas no grupo com doença, indicando que essas variáveis são indicadores importantes para a identificação da condição clínica. O mesmo padrão de elevação é observado nas enzimas Sgpt, Sgot e Alkphos. Embora as medianas gerais dessas enzimas sejam baixas, os boxplots revelam que quase a totalidade dos valores muito altos pertence ao grupo enfermo, o que sugere que a presença da doença tende a elevar drasticamente esses níveis em comparação ao grupo saudável.

Em contrapartida, as variáveis ligadas às proteínas, como a Albumina (ALB) e a Razão A/G, apresentam um comportamento inverso. No grupo com doença, os valores tendem a ser menores, com a mediana da Albumina fixada em 3,10. Os boxplots demonstram que, enquanto o grupo saudável mantém níveis proteicos mais altos e estáveis, o grupo enfermo apresenta uma queda nesses índices, o que pode ser um sinal clínico da progressão da doença.

No caso da Idade (Age) e das Proteínas Totais (TP), nota-se uma maior sobreposição entre os grupos, o que indica uma capacidade limitada de distinção quando essas variáveis são analisadas isoladamente, embora a idade média dos pacientes doentes seja ligeiramente superior (44,78 anos). Por outro lado, a inspeção dos boxplots sugere que as variáveis relacionadas à Bilirrubina Total (TB) e à Bilirrubina Direta (DB) exibem padrões mais distintos entre pacientes com e sem doença, sobretudo em termos de dispersão e assimetria, sinalizando a condição clínica por meio de valores elevados. De forma complementar, a Albumina (ALB) e a Razão Albumina/Globulina (A/G Ratio) apresentam deslocamentos mais sutis em suas distribuições, indicando a presença da patologia por meio de valores reduzidos. No contexto da análise gráfica realizada, esse conjunto de métricas apresenta indícios relevantes para a distinção entre os grupos.

Tabela 3.4: Correlação entre a variável resposta e as demais covariáveis

Variável	**Correlação**
DB	0,2463
TB	0,2202
Alkphos	0,1834
A/G Ratio	- 0,1631
Sgpt	0,1631
ALB	- 0,1598
Sgot	0,1518
Age	0,1332
TP	- 0,0336

Calculamos também a correlação de Pearson entre a variável resposta e as covariáveis. A correlação de Pearson foi idealizada inicialmente para medir associação entre duas variáveis quantitativas. No entanto, ela também é usada quando uma das variáveis é binária e nesse caso ela é denominada correlação ponto-bisserial (Tate, 1954). Pela Tabela 3.4 observamos que as correlações entre a variável resposta e as covariáveis são, em sua maioria, fracas. As variáveis DB (0,2463) e TB (0,2202) apresentam as maiores correlações positivas, indicando que maiores níveis desses indicadores estão associados a uma maior probabilidade de ocorrência da doença. Outras variáveis, como Alkphos (0,1834), Sgpt (0,1631) e Sgot (0,1518), também apresentam correlações positivas, embora mais fracas. Por outro lado, as variáveis A/G Ratio (-0,1631), ALB (-0,1598) e TP (-0,0336) apresentam correlações negativas, sugerindo uma possível associação inversa com o desfecho. Embora os valores sejam relativamente baixos (todas abaixo de 0,25), esses resultados ajudam a identificar possíveis covariáveis relevantes, especialmente DB e TB, que exibem associação um pouco mais expressiva com a variável resposta.

A análise das relações entre as covariáveis numéricas foi conduzida através do cálculo da matriz de correlação de Pearson. A Figura 3.4 apresenta o resultado desta análise, representado em formato de mapa de calor (*heatmap*). Nota-se uma correlação muito forte ($r = 0,87$) entre o total de bilirrubina (TB) e a bilirrubina direta (DB), além de uma correlação elevada ($r = 0,79$) entre as enzimas hepáticas Sgpt e Sgot. Também há correlação alta entre a albumina (ALB) e a razão A/G Ratio ($r = 0,69$), além de forte associação ($r = 0,78$) entre a ALB e TP. Tais correlações indicam a possibilidade de multicolinearidade, aspecto que foi considerado na etapa de seleção de variáveis preditoras por meio do método *stepwise*.

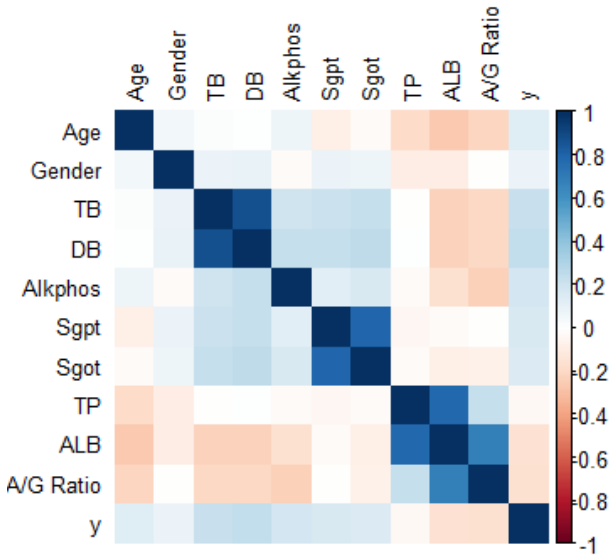


Figura 3.4: Correlação entre variáveis

3.3 Modelagem

Nesta seção apresentaremos os resultados do ajuste dos modelos lineares generalizados com distribuição da família binomial. A seleção de variáveis foi realizada através do método *Stepwise*.

Para ajuste e seleção de modelos, optamos por utilizar todas as funções de ligações propostas para realização de comparação entre elas.

3.3.1 Seleção de variáveis

Para a comparação das funções de ligação, foram considerados cinco modelos:

- **Modelo 1**: Inclui as variáveis do melhor ajuste, selecionadas pelo método *stepwise* em cada função de ligação testada.
- **Modelo 2**: Incluiu as variáveis selecionadas pelo método *stepwise* quando foram usadas as seguintes funções de ligação: Logito, Logito Potência, Probito Potência, Cloglog Potência, Logito Potência Reversa e Probito Potência Reversa. Resultando nas seguintes variáveis explicativas: **Age**, **DB**, **Alkphos**, **Sgpt**, **ALB**, **A/G Ratio** e **TP**.
- **Modelo 3**: Incluiu as variáveis selecionadas pelo método *stepwise* quando foram

usadas as seguintes funções de ligação: Probito e Cloglog, contendo as seguintes variáveis explicativas: **Age**, **DB**, **Sgpt**, **ALB**, **A/G Ratio** e **TP**.

- **Modelo 4**: Identificou entre as covariáveis do Modelo 3 aquelas que eram fortemente correlacionadas. Em seguida, de cada par fortemente correlacionado, excluiu-se a covariável menos correlacionada com a resposta, buscando evitar problemas de multicolinearidade. Foram selecionadas as seguintes variáveis explicativas: **Age**, **DB**, **Sgpt** e **TP**.

- **Modelo 5**: Foi construído a partir da remoção inicial das variáveis mais correlacionadas, seguida da aplicação do método *stepwise* sobre o conjunto reduzido, o que acabou gerando alguns modelos candidatos. Entre eles, observou-se que variáveis como **Age**, **DB**, **Sgpt** e **A/G Ratio** apareceram com mais frequência, indicando maior consistência na contribuição dessas covariáveis ao ajuste.

3.3.2 Ajuste do modelo

As funções de ligação Potência e Potência Reversa possuem um parâmetro λ que deve ser escolhido. Para essas funções de ligação, consideramos diferentes valores do parâmetro λ, variando de 0,1 a 200. A métrica utilizada para comparação foi a Área sob a Curva ROC (AUC), por ser adequada para avaliação de desempenho preditivo em situações de desbalanceamento entre as classes. Para cada função de ligação, escolhemos o valor de λ que levava a um melhor AUC.

Tabela 3.5: Área sob a curva ROC para os diferentes modelos e funções de ligação

	Modelos				
Função de Ligação	**1**	**2**	**3**	**4**	**5**
Logito	0,7679	0,7679	0,7599	0,7617	0,7615
Probito	0,7595	0,7653	0,7595	0,7593	0,7587
Cloglog	0,7587	0,7626	0,7587	0,7559	0,7546
Logito Potência ($\lambda = 80$)	0,7690	0,7690	0,7604	0,7633	0,7629
Probito Potência ($\lambda = 160$)	0,7680	0,7680	0,7578	0,7611	0,7606
Cloglog Potência ($\lambda = 180$)	0,7658	0,7658	0,7529	0,7577	0,7572
Logito Potência Reversa ($\lambda = 0,2$)	**0,7704**	**0,7704**	0,7604	0,7634	0,7633
Probito Potência Reversa ($\lambda = 0,1$)	0,7663	0,7663	0,7551	0,7591	0,7583
Cloglog Potência Reversa ($\lambda = 0,8$)	–	0,7697	0,7604	0,7634	0,7630

Após a análise dos resultados (Tabela 3.5), foi identificado que a função **Logito Potência Reversa**, com λ = 0,2, apresentou o melhor desempenho preditivo (AUC

= 0,7703), superando as funções tradicionais. Outros modelos com bom desempenho incluíram a função Cloglog Potência Reversa com $\lambda = 0{,}8$ (AUC = 0,7697) e a Logito Potência com $\lambda = 7$ (AUC = 0,7690). Ressalta-se, entretanto, que a função de ligação Cloglog Potência Reversa não apresentou valor de AUC para o Modelo 1, devido a problemas no algoritmo de estimação durante a execução desse modelo. Dessa forma, o modelo final adotado utiliza a função de ligação Logito Potência Reversa com $\lambda = 0{,}2$, aplicada ao conjunto de covariáveis selecionadas no Modelo 2.

Portanto o modelo final é:

$$-\log\left(\frac{(1-\mu)^{1/\lambda}}{1-(1-\mu)^{1/\lambda}}\right)$$

$$= \beta_0 + \beta_1 \text{Age}_i + \beta_2 \text{DB}_i + \beta_3 \text{Alkphos}_i + \beta_4 \text{Sgpt}_i + \beta_5 \text{ALB}_i + \beta_6 \text{A/G Ratio}_i + \beta_7 \text{TP}_i. \quad (3.1)$$

3.3.3 Interpretação dos sinais dos coeficientes

A Tabela 3.6 resume os resultados do modelo final com a função de ligação Logito Potência Reversa. Todas as covariáveis do modelo final são estatisticamente significativas ao nível de 5%, com exceção da variável Alkphos, cujo valor-p (0,05377) está muito próximo do limiar. Ao contrário do que ocorre quando a função Logito é usada, para a função de ligação Logito Potência Reversa, as estimativas dos parâmetros não podem ser interpretadas. No entanto, é possível interpretar os sinais das estimativas, como faremos a seguir.

Tabela 3.6: Resultados dos coeficientes do modelo final com função Logito Potência Reversa ($\lambda = 0{,}2$)

Variável	**Estimativa**	**Erro Padrão**	**z-valor**	**valor-p**
(Intercept)	-10,425	4,218	-2,472	0,0134
Age	0,057	0,019	3,047	0,0023
DB	1,931	0,660	2,925	0,0034
Alkphos	0,006	0,003	1,929	0,0538
Sgpt	0,052	0,015	3,547	0,0004
ALB	3,487	1,237	2,820	0,0048
A/G Ratio	-6,565	2,396	-2,740	0,0061
TP	7,913	3,615	2,189	0,0286

A variável Idade (Age) apresentou coeficiente positivo (0,057), indicando que o aumento da idade está associado a uma maior probabilidade de ter a doença hepática. O mesmo ocorre com as variáveis nível de bilirrubina direta no sangue (DB), nível da en-

zima alanina aminotransferase (Sgpt), nível de albumina (ALB) e a concentração total de proteínas no sangue (TP), que também possuem coeficientes positivos, sugerindo uma relação direta com a resposta, ou seja, aumenta a probabilidade de doença hepática.

Por outro lado, a variável que representa a razão entre albumina e globulina (A/G Ratio) apresentou coeficiente negativo (-6,56), indicando que maiores valores dessa razão estão associados a menor probabilidade de doença hepática. Esse achado pode ter relevância clínica e merece atenção especial.

A variável Nível de fosfatase alcalina (Alkphos), apesar de apresentar uma tendência positiva, teve significância marginal ($p = 0{,}05377$), e por isso sua inclusão no modelo pode ser discutida dependendo do critério adotado para significância estatística.

Ao compararmos os sinais dos coeficientes do modelo com os resultados da análise descritiva, observa-se que a maioria está em conformidade com as correlações previamente verificadas entre as covariáveis e a variável resposta. No entanto, duas exceções importantes merecem destaque: as variáveis ALB (nível de albumina) e TP (proteína total) apresentaram correlações negativas com a presença de doença hepática na análise descritiva, mas seus coeficientes estimados no modelo foram positivos. Essa discrepância pode estar associada à presença de efeitos de confusão ou multicolinearidade entre as variáveis, o que pode comprometer a interpretação direta dos sinais. Se o objetivo da modelagem for essencialmente preditivo, o modelo final pode ser mantido, uma vez que captura adequadamente a relação global entre as covariáveis e a resposta. Contudo, se o foco for interpretativo, ou seja, compreender os efeitos individuais das covariáveis, seria conveniente reespecificar o modelo, removendo uma ou mais covariáveis até que os sinais das estimativas estejam coerentes com os padrões observados na análise descritiva.

3.3.4 Discussão dos Resultados

De modo geral, os valores de AUC foram muito próximos entre as funções testadas, com diferenças pequenas e pouco significativas do ponto de vista prático. Isso sugere que todas as funções avaliadas apresentam desempenho preditivo semelhante neste conjunto de dados, mesmo com variações no parâmetro λ.

Adicionalmente, observou-se que as funções de ligação simétricas, como o Logito e o Probito, também apresentaram resultados satisfatórios, mesmo diante do desbalanceamento das classes. Esses resultados reforçam a robustez dessas funções tradicionais na modelagem de dados binários desbalanceados.

Embora a função de ligação Logito Potência Reversa com $\lambda = 0{,}2$ tenha apresentado o melhor desempenho preditivo entre os modelos avaliados, a função Logito apresenta a importante vantagem de permitir a interpretação direta das estimativas dos parâmetros. Ademais, sua AUC foi apenas ligeiramente inferior à do modelo final. Essa característica torna o Logito uma alternativa particularmente útil em contextos em que a compreensão dos efeitos das covariáveis é tão relevante quanto a capacidade preditiva do modelo.

Capítulo 4

Outras Aplicações

Nesta seção são apresentadas novas aplicações da metodologia proposta a dois conjuntos de dados adicionais: o *Default of Credit Card Clients*, obtido do repositório *UCI Machine Learning Repository* (Yeh, 2009), e o *Premature Birth Dataset*, disponibilizado no artigo original em que foi estudado (Galo *et al.*, 2023). O objetivo, assim como no capítulo anterior, é avaliar o desempenho de diferentes funções de ligação em Modelos Lineares Generalizados (MLG) com resposta binária, e em níveis diferentes de desbalanceamento entre as classes.

4.1 Banco de dados e metodologia

Os dois bancos apresentam características distintas, tanto em tamanho quanto no nível de desbalanceamento da variável resposta. O conjunto de crédito reúne informações financeiras e de histórico de pagamento de clientes de cartão de crédito em Taiwan, com o objetivo de analisar e prever a probabilidade de inadimplência. Já o conjunto de prematuridade contém registros clínicos maternos e variáveis relativas às condições gestacionais, utilizados para modelar a ocorrência de parto prematuro. A Tabela 4.1 apresenta um resumo das principais características de ambos os conjuntos de dados.

Tabela 4.1: Resumo dos bancos de dados utilizados

Conjunto de Dados	**Observações**	**Variáveis**	**Desbalanceamento**
Credit Card	30.000	23	78%
Premature Birth	5.060	15	88%

A análise inicial foi conduzida utilizando a amostra completa de cada conjunto de

dados, de modo a identificar diretamente as funções de ligação com melhor desempenho. No caso do *Default of Credit Card Clients*, todos os ajustes foram realizados dessa forma, pois a divisão dos dados em treino e teste traria pouca vantagem prática, já que a amostra é grande e a complexidade dos modelos considerados neste trabalho não é muito diferente. Além disso, isso aumentaria consideravelmente o custo computacional, especialmente devido ao grande número de modelos avaliados nas funções generalizadas.

Já para o *Premature Birth Dataset*, foi realizada uma segunda etapa de análise, motivada principalmente pelo alto grau de desbalanceamento da variável resposta. Para avaliar a estabilidade preditiva dos modelos sob esse cenário, os dados foram particionados em treino (50%), validação (25%) e teste (25%). Esse procedimento foi repetido dez vezes, para reduzir a influência da semente aleatória nos resultados. As funções tradicionais foram ajustadas no conjunto de treino e avaliadas diretamente no teste. Para as funções Potência e Reversa da Potência, o parâmetro λ foi selecionado com base no melhor desempenho na validação, e o modelo final, reestimado com esse λ, teve sua AUC obtida no conjunto de teste.

A seleção de variáveis seguiu a mesma estratégia adotada no Capítulo 3, sendo realizada por meio do método *stepwise* guiado pelo critério de Akaike (AIC). Optou-se por não utilizar o método LASSO, uma vez que sua implementação no software `R` não permite incorporar funções de ligação personalizadas, o que inviabilizaria a comparação uniforme entre todos os modelos. O uso do *stepwise* assegurou maior flexibilidade e permitiu aplicar o mesmo procedimento de seleção a todas as funções de ligação. O desempenho final dos modelos foi comparado por meio da AUC, apropriada para cenários de desbalanceamento.

4.2 Resultados

Os resultados obtidos para ambos os conjuntos de dados são apresentados na Tabela 4.2, que resume os valores de AUC e os correspondentes valores de λ para cada função de ligação considerada. Essa primeira análise, conduzida sobre a base completa de cada conjunto de dados, permite observar como as diferentes funções se comportam quando avaliadas sob as mesmas condições de ajuste e seleção de variáveis.

A partir desses resultados, nota-se que as funções tradicionais alcançam desempenhos bastante semelhantes entre si e algumas funções generalizadas produzem resultados ligeiramente superiores, tanto no *Premature Birth Dataset* quanto no *Default of Credit Card*

Tabela 4.2: Área sob a curva ROC para cada função de ligação

	Premature Birth		**Credit Card**	
Função de Ligação	λ	AUC	λ	AUC
Logito	–	0,8307	–	0,7241
Probito	–	0,8311	–	0,7233
Cloglog	–	0,8287	–	0,7259
Logito Potência	2,0	0,8311	0,1	0,7257
Probito Potência	0,7	0,8313	0,2	0,7289
Cloglog Potência	11,0	0,8313	0,9	0,7261
Logito Potência Reversa	0,4	0,8311	110,0	0,7259
Probito Potência Reversa	1,4	0,8313	40,0	0,7234
Cloglog Potência Reversa	21,0	0,8314	120,0	0,7222

Clients. Essas diferenças, no entanto, são pequenas e não modificam de forma relevante a relação de desempenho entre as funções. Além disso, observa-se que os valores de λ que proporcionam melhor ajuste diferem de forma considerável entre as duas bases, indicando que o comportamento ótimo das funções generalizadas é sensível às características específicas de cada conjunto de dados.

Tabela 4.3: Média e desvio-padrão do AUC no conjunto de teste

Função de Ligação	**Média**	**DP**
Logito	0,812	0,0110
Probito	0,811	0,0109
Cloglog	0,810	0,0106
Logito Potência	0,814	0,0112
Probito Potência	0,811	0,0115
Cloglog Potência	0,812	0,0105
Logito Potência Reversa	0,813	0,0112
Probito Potência Reversa	0,812	0,0113
Cloglog Potência Reversa	0,812	0,0111

Além da análise com a amostra completa, o *Premature Birth Dataset* passou por uma avaliação adicional, baseada na divisão em conjuntos de treino, validação e teste, repetida dez vezes. A Tabela 4.3 apresenta a média e o desvio-padrão do AUC no conjunto de teste, permitindo observar a estabilidade dos resultados ao longo das repetições. Nota-se que, mesmo nesse cenário mais rigoroso, as diferenças entre as funções de ligação permanecem pequenas, reforçando a proximidade de desempenho entre as abordagens tradicionais e generalizadas.

De maneira geral, os resultados indicam que as funções de ligação tradicionais e suas

versões generalizadas apresentam desempenhos bastante próximos. Embora as funções Potência e Reversa da Potência permitam ajustes ligeiramente superiores em alguns casos, esses ganhos são discretos e não alteram substancialmente o padrão de desempenho observado entre as diferentes funções.

Capítulo 5

Estudos de Simulção

Neste capítulo, apresentamos a implementação do estudo de simulação de Monte Carlo para avaliar o desempenho de diferentes funções de ligação em cenários com variados níveis de desbalanceamento da variável resposta.

5.1 Estrutura da Simulação

O estudo de simulação foi conduzido utilizando 1000 réplicas de Monte Carlo e um tamanho amostral igual a 500. Em cada réplica, geraram-se covariáveis X_1 e X_2 (ou mais, dependendo do cenário), e a variável resposta y_i foi simulada a partir de uma distribuição Bernoulli com probabilidade p_i. O modelo considerado é o seguinte:

$$y_i \sim \text{Bernoulli}(p_i), \quad g(p_i) = \eta_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i},$$

em que $g(\cdot)$ representa a função de ligação utilizada para gerar os dados. Inicialmente, a simulação foi conduzida tomando a ligação Logito como geradora do processo, e todo o procedimento de ajuste e comparação das funções de ligação foi repetido posteriormente para o caso em que a função de ligação Cloglog foi utilizada como geradora. Note que consideramos uma função de ligação simétrica e uma assimétrica para gerar os dados nesses estudos de simulação.

As funções de ligação que foram ajustadas aos dados simulados incluíram as tradicionais (Logito, Probito e Cloglog) e suas versões generalizadas, obtidas por meio das transformações Potência e Potência Reversa. Para essas versões, avaliou-se um grid de valores do parâmetro λ variando entre 0,1 e 50, de forma a selecionar o valor que maxi-

mizasse o desempenho do modelo em cada réplica.

Além disso, o desempenho dos modelos foi avaliado por meio da Área sob a Curva ROC (AUC), utilizada para medir a capacidade discriminatória do modelo. Além disso, considerou-se cenários com diferentes níveis de desbalanceamento.

5.2 Cenários Avaliados

Foram considerados quatro cenários distintos, construídos para avaliar o desempenho das diferentes funções de ligação sob variações nos parâmetros do modelo e na quantidade e distribuições das covariáveis. As simulações foram estruturadas de forma a abranger situações com diferentes níveis de desbalanceamento da variável resposta, distintos graus de complexidade do modelo e diferentes distribuições para a variável resposta. A Tabela 5.1 descreve as principais características de cada cenário.

Tabela 5.1: Resumo dos cenários de simulação

Cenário	β	**Covariáveis**	**Prop. 1's**
1	(2,5; 2; -2)	$X_1, X_2 \sim U(0,1)$	0,90
2	(1; 2; -2)	$X_1, X_2 \sim U(0,1)$	0,70
3	(2,5; 2; -2)	$X_1 \sim N(0.5, 1/12)$	0,90
		$X_2 \sim \text{Gama}(0.5, 1/12)$	
4	(2,5; 2; -2;...; 2; -2)	$X_1, ..., X_{10} \sim U(0,1)$	0,92

Quando o processo de simulação foi repetido utilizando a função Cloglog como ligação geradora, foi necessário reajustar o intercepto para que os níveis de desbalanceamento se mantivessem comparáveis aos obtidos com a ligação Logito. Para os Cenários 1 e 2, utilizaram-se $\beta_0 = 1,3$ e $\beta_0 = 0,3$, respectivamente. Já nos Cenários 3 e 4, adotou-se o mesmo valor de β_0 empregado no Cenário 1 ($\beta_0 = 1,3$), assegurando proporções de 1's semelhantes às pretendidas inicialmente.

5.3 Resultados Obtidos

Os resultados obtidos são apresentados separadamente para os cenários gerados a partir da ligação Logito e para os gerados a partir da Cloglog. Em ambos os casos, as tabelas reúnem as médias de AUC e os respectivos desvios-padrão das 1000 réplicas, permitindo comparar diretamente o comportamento das funções tradicionais e generalizadas.

5.3.1 Cenários Gerados pela Função de Ligação Logito

De modo geral, os resultados indicaram alta consistência entre as funções de ligação testadas, tanto nas versões tradicionais quanto nas estendidas. A Tabela 5.2 apresenta as médias e desvios-padrão do AUC para os quatro cenários analisados.

Tabela 5.2: Resultados de AUC médio e desvio-padrão com função de ligação Logito

Função de Ligação	**Cenário 1**		**Cenário 2**	
	Média	**DP**	**Média**	**DP**
Logito	0,7189	0,0378	0,7127	0,0251
Probito	0,7190	0,0378	0,7127	0,0251
Cloglog	0,7190	0,0378	0,7127	0,0251
Logito Potência	0,7194	0,0377	0,7129	0,0251
Probito Potência	0,7193	0,0377	0,7129	0,0251
Cloglog Potência	0,7190	0,0378	0,7127	0,0251
Logito Potência Reversa	0,7194	0,0377	0,7129	0,0251
Probito Potência Reversa	0,7192	0,0378	0,7128	0,0251
Cloglog Potência Reversa	0,7190	0,0378	0,7127	0,0251
Função de Ligação	**Cenário 3**		**Cenário 4**	
	Média	**DP**	**Média**	**DP**
Logito	0,7242	0,0387	0,8761	0,0289
Probito	0,7244	0,0387	0,8773	0,0289
Cloglog	0,7245	0,0388	0,8778	0,0290
Logito Potência	0,7252	0,0387	0,8790	0,0287
Probito Potência	0,7249	0,0387	0,8782	0,0288
Cloglog Potência	0,7245	0,0388	0,8770	0,0286
Logito Potência Reversa	0,7251	0,0387	0,8785	0,0288
Probito Potência Reversa	0,7248	0,0387	0,8782	0,0289
Cloglog Potência Reversa	0,7245	0,0388	0,8770	0,0286

Nos quatro cenários analisados, observou-se que as diferenças entre as funções de ligação foram mínimas, com desempenhos praticamente equivalentes em termos de AUC. Note que, para cada método, o AUC não é muito diferente nos cenários 1, 2 e 3.

Nos quatro cenários, as funções de ligação Logito Potência e Logito Potência Reversa foram as que apresentam maior AUC em média. No entanto, a diferença do desempenho destas para as funções de ligação simétricas são bem pequenas.

No cenário 4, o aumento no número de covariáveis resultou em maior AUC médio, como esperado, mas as diferenças entre funções de ligação continuaram mínimas, reforçando a robustez das funções simétricas.

5.3.2 Cenários Gerados pela Função de Ligação Cloglog

Os resultados obtidos quando os dados foram gerados pela função de ligação Cloglog seguem o mesmo padrão observado na simulação baseada na Logito. Conforme apresentado na Tabela 5.3, todas as funções de ligação, tanto tradicionais quanto generalizadas, exibiram desempenhos extremamente próximos nos quatro cenários avaliados, com diferenças de AUC sempre muito pequenas. Mesmo diante da assimetria inerente à Cloglog, os modelos ajustados com Logito e Probito apresentaram desempenho praticamente equivalente ao da função geradora dos dados, indicando que a escolha da ligação não altera de forma relevante a capacidade discriminatória dos modelos nesses contextos.

Tabela 5.3: Resultados de AUC médio e desvio-padrão com função de ligação Cloglog

Função de Ligação	**Cenário 1**		**Cenário 2**	
	Média	**DP**	**Média**	**DP**
Logito	0,8804	0,02112	0,8130	0,0196
Probito	0,8805	0,02114	0,8130	0,0196
Cloglog	0,8804	0,0211	0,8130	0,0196
Logito Potência	0,8808	0,0211	0,8132	0,0196
Probito Potência	0,8808	0,0211	0,8131	0,0196
Cloglog Potência	0,8804	0,0211	0,8130	0,0196
Logito Potência Reversa	0,8808	0,0211	0,8132	0,0196
Probito Potência Reversa	0,8807	0,0211	0,8131	0,0196
Cloglog Potência Reversa	0,8804	0,0211	0,8130	0,0196
Função de Ligação	**Cenário 3**		**Cenário 4**	
	Média	**DP**	**Média**	**DP**
Logito	0,8962	0,0205	0,9646	0,0095
Probito	0,8964	0,0204	0,9650	0,0094
Cloglog	0,8967	0,0204	0,9654	0,0094
Logito Potência	0,8970	0,0204	0,9657	0,0094
Probito Potência	0,8968	0,0204	0,9651	0,0093
Cloglog Potência	0,8961	0,0201	0,9623	0,0103
Logito Potência Reversa	0,8970	0,0204	0,9655	0,0094
Probito Potência Reversa	0,8968	0,0204	0,9651	0,0094
Cloglog Potência Reversa	0,8961	0,0201	0,9623	0,0103

Nos cenários 1 e 2, em que foi necessário ajustar o valor do intercepto para manter níveis comparáveis de desbalanceamento, os valores de AUC situaram-se entre 0,81 e 0,88, acompanhados de desvios-padrão reduzidos, o que reforça a estabilidade dos resultados independentemente da função utilizada. As versões com parâmetro λ, tanto Potência

quanto Potência Reversa, apresentaram pequenas melhorias pontuais, porém sem impacto significativo no desempenho global.

Nos cenários 3 e 4, caracterizados por covariáveis mais assimétricas ou por maior número de preditores, os valores de AUC aumentaram de maneira natural. Ainda assim, a relação entre as funções permaneceu essencialmente inalterada. Todas convergiram para desempenhos praticamente idênticos, novamente sem vantagem expressiva para as funções generalizadas.

Novamente, as funções Logito Potência e Logito Potência Reversa apresentaram resultados, em geral, ligeiramente superiores às demais. Porém, as diferenças são bem pequenas em relação às demais funções de ligação.

Capítulo 6

Considerações Finais

Este trabalho buscou compreender como diferentes funções de ligação, tradicionais e generalizadas, se comportam na modelagem de respostas binárias em cenários marcados por forte desbalanceamento. A partir de aplicações em dados reais e estudos de simulação, comparou-se o desempenho das ligações Logito, Probito e Complemento Log-log com suas extensões das famílias Potência e Potência Reversa, que introduzem um parâmetro extra para ajustar a assimetria da curva de resposta. O objetivo central foi avaliar se essa flexibilidade adicional realmente se traduz em ganhos preditivos relevantes em relação às funções clássicas. Também se buscava verificar se realmente em dados desbalanceados o desempenho é melhor se for usada uma função de ligação assimétrica, como afirmam alguns trabalhos da literatura estatística.

Os resultados obtidos, tanto nas simulações quanto nas aplicações práticas, mostraram que as funções de ligação simétricas, como Logito e Probito, são tão eficazes quanto as funções assimétricas, mesmo em situações com forte desbalanceamento ou covariáveis assimétricas. Embora algumas versões generalizadas, como a Logito Potência Reversa, tenham apresentado pequenas melhorias em termos de AUC, esses ganhos não foram expressivos e não alteram a conclusão geral de que as ligações tradicionais oferecem desempenho estável e confiável. Essa evidência reforça que, se o foco principal do estudo for a interpretabilidade dos coeficientes, as funções de ligação simétricas, especialmente a Logito, continuam sendo uma escolha sólida, pois oferecem um bom equilíbrio entre o ajuste, a clareza e a facilidade de implementação.

Os resultados deste estudo também abrem espaço para novas investigações. Uma possibilidade é aplicar esse tipo de análise a outras variáveis resposta, como contagens ou tempos até evento, para verificar se o uso de funções de ligação mais flexíveis produz

efeitos parecidos com os observados em dados binários. Outro caminho interessante seria desenvolver algoritmos que permitam usar métodos de penalização, como o LASSO, junto com funções de ligação personalizadas que possuem parâmetros extras, algo que ainda não é oferecido pelas ferramentas atuais de forma direta. Criar esse tipo de algoritmo poderia tornar a modelagem mais robusta, eficiente e adequada para bases de dados maiores e mais complexas.

De forma geral, este trabalho mostra que não existe uma função de ligação que seja a melhor em todos os cenários. O desempenho dos modelos depende muito das características dos dados, e pequenas diferenças entre as ligações podem fazer diferença em algumas aplicações. Ainda assim, as funções tradicionais continuam sendo opções confiáveis por serem estáveis e mais fáceis de interpretar, enquanto as versões generalizadas surgem como alternativas válidas quando se busca mais flexibilidade. Particularmente, a função de ligação Logito, que é a mais interpretável delas, mostrou-se bastante competitiva. Em todos os cenários de simulação e nas 3 aplicações, a função Logito, mesmo sendo simétrica, obteve AUC próximo das funções de ligação assimétricas mais flexível, sob diferentes graus de desbalanceamento dos dados. De qualquer forma, o estudo de diferentes funções de ligação continua sendo um tema relevante e com muitos caminhos para avanços futuros, tanto na prática quanto no campo teórico.

Referências Bibliográficas

Agresti, A. (2015). *Foundations of linear and generalized linear models*. John Wiley & Sons.

Bazán, J., Torres-Avilés, F., Suzuki, A. e Uribe-Opazo, M. A. (2017). Power and reversal power links for binary regressions: an application for motor insurance policyholders. *Appl Stoch Models Bus Ind*, **33**(1), 22–34.

Bazán, J. L., Romeo, J. S. e Rodrigues, J. (2014). Bayesian skew-probit regression for binary response data.

Brolo, C. L. (2019). Comparação da performance do lasso e do método da máxima verossimilhança com seleção de variáveis em modelos de regressão para dados binários.

Chen, M.-H., Dey, D. K. e Shao, Q.-M. (1999). A new skewed link model for dichotomous quantal response data. *Journal of the American Statistical Association*, **94**(448), 1172–1186.

Collett, D. (1991). *Modelling binary data*. Chapman and Hall/CRC, London.

Cordeiro, G. M., Demétrio, C. G. B. e Moral, R. A. (2024). *Modelos lineares generalizados e aplicações*. Blucher, São Paulo, SP, first edition.

Davis, J. e Goadrich, M. (2006). The relationship between precision-recall and roc curves. Em *Proceedings of the 23rd international conference on Machine learning*, páginas 233–240.

Dobson, A. J. e Barnett, A. G. (2018). *An Introduction to Generalized Linear Models*. Chapman and Hall/CRC, Boca Raton, FL, fourth edition.

Fawcett, T. (2006). An introduction to roc analysis. *Pattern recognition letters*, **27**(8), 861–874.

Galo, R., Rossi, R. M., Alves, D. C. e de Oliveira, R. R. (2023). Bayesian binary regression using power and power reverse link functions: an application to premature birth data. *Brazilian Journal of Biometrics*, **41**(2), 131–143.

Hanley, J. A. e McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (roc) curve. *Radiology*, **143**(1), 29–36.

He, H. e Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on knowledge and data engineering*, **21**(9), 1263–1284.

Hosmer Jr, D. W., Lemeshow, S. e Sturdivant, R. X. (2013). *Applied logistic regression*. John Wiley & Sons, third edition.

Huayanay, A., Bazan, J. L., Cancho, V. G. e Dey, D. K. (2019). Performance of asymmetric links and correction methods for imbalanced data in binary regression. *Journal of Statistical Computation and Simulation*, **89**(9), 1694–1714.

Lemonte, A. J. e Bazán, J. L. (2018). New links for binary regression: an application to coca cultivation in peru. *Test*, **27**, 597–617.

Magalhães, M. N. e De Lima, A. C. P. (2011). *Noções de Probabilidade e Estatística*. Edusp, São Paulo.

McCullagh, P. (2019). *Generalized linear models*. Routledge.

Nelder, J. A. e Wedderburn, R. W. (1972). Generalized linear models. *Journal of the Royal Statistical Society Series A: Statistics in Society*, **135**(3), 370–384.

Paula, G. A. (2004). *Modelos lineares generalizados com inferência bayesiana*. Editora da Universidade de São Paulo (EDUSP), São Paulo.

R Core Team (2022). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Ramana, B. e Venkateswarlu, N. (2022). ILPD (Indian Liver Patient Dataset). UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C5D02C.

Salgado, C. M., Azevedo, C., Proença, H. e Vieira, S. M. (2016). *Missing Data*, páginas 143–162. Springer International Publishing, Cham. ISBN 978-3-319-43742-2.

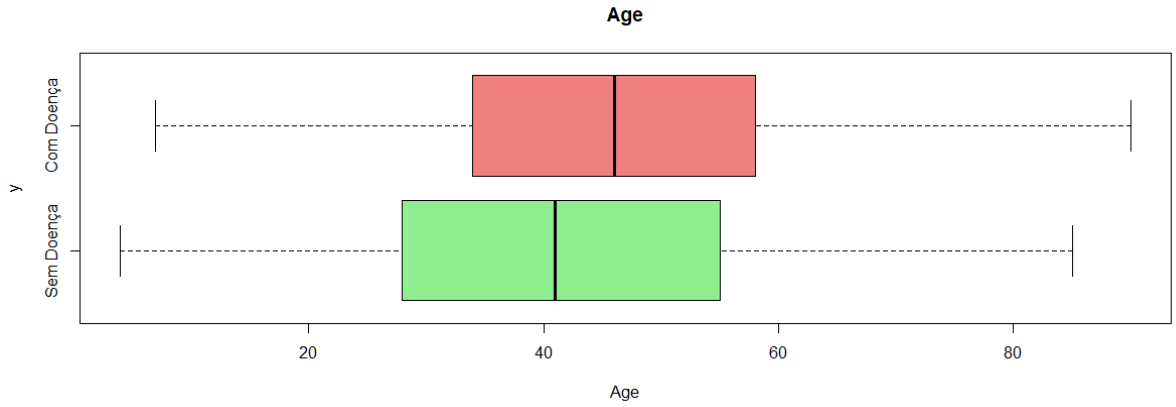
Tate, R. F. (1954). Correlation between a discrete and a continuous variable. point-biserial correlation. *The Annals of mathematical statistics*, **25**(3), 603–607.

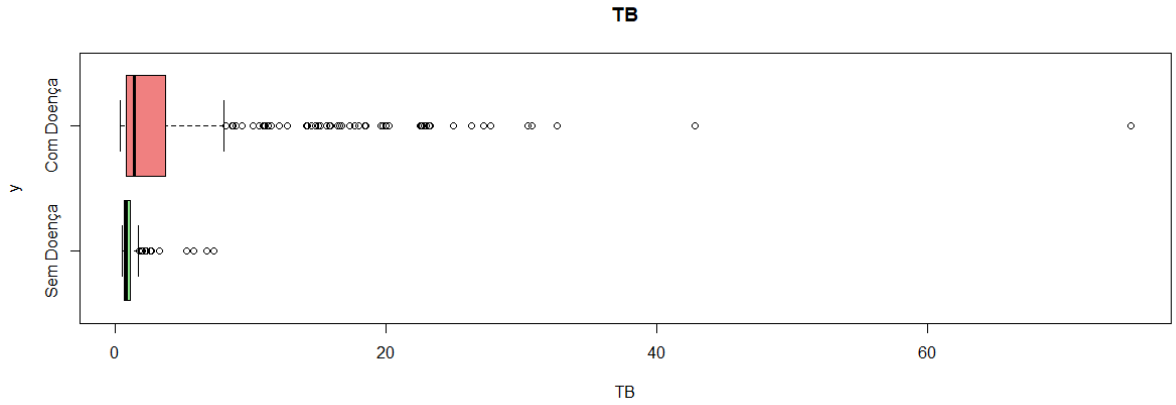
Yeh, I.-C. (2009). Default of Credit Card Clients. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C55S3H.

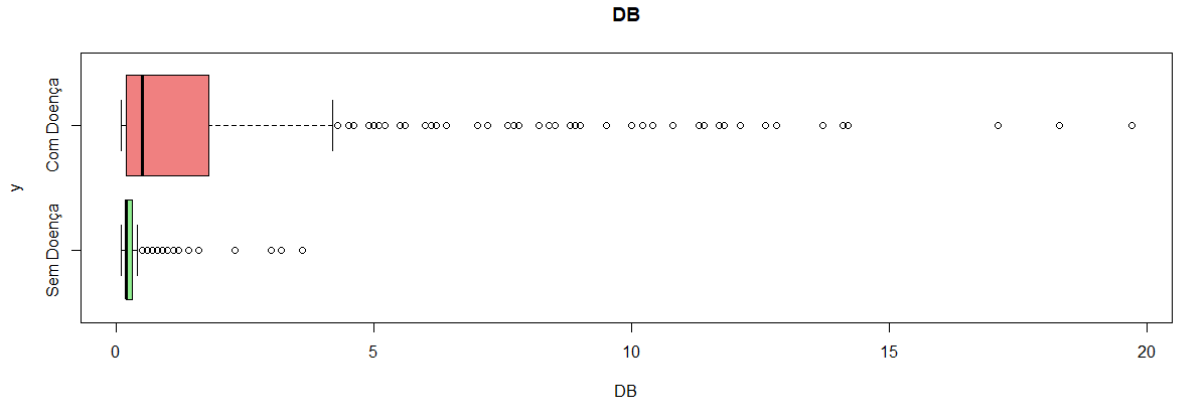
Zou, Q., Xie, S., Lin, Z., Wu, M. e Ju, Y. (2016). Finding the best classification threshold in imbalanced classification. *Big Data Research*, **5**, 2–8.

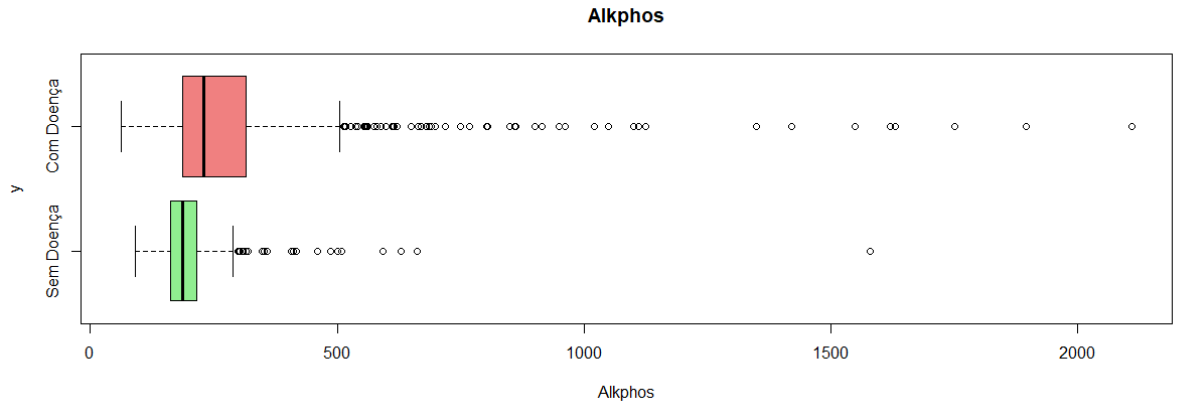
Apêndice A

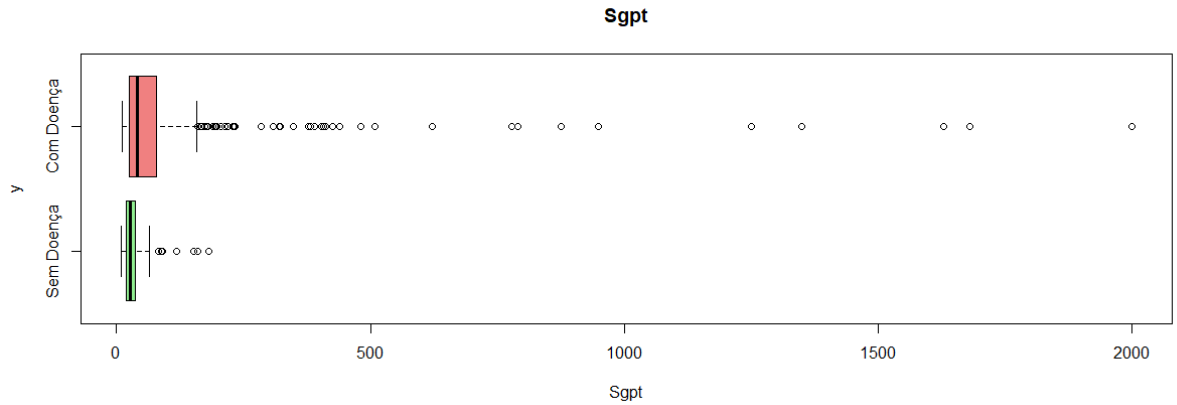
Gráficos adicionais

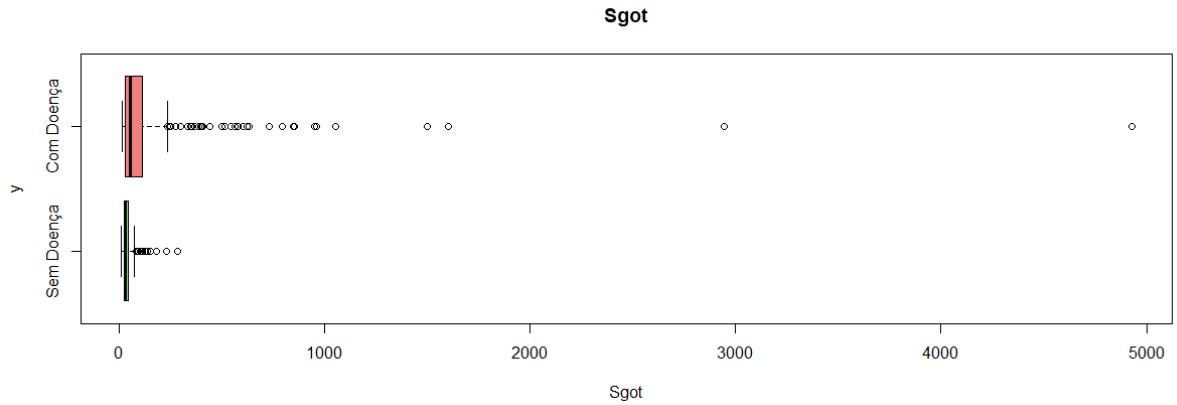





DB
y
Com Doença
Sem Doença
0
5
10
15
20
DB


Alkphos
y
Com Doença
Sem Doença
0
500
1000
1500
2000
Alkphos


Sgpt
y
Com Doença
Sem Doença
0
500
1000
1500
2000
Sgpt


Sgot
y
Com Doença
Sem Doença
0
1000
2000
3000
4000
5000
Sgot

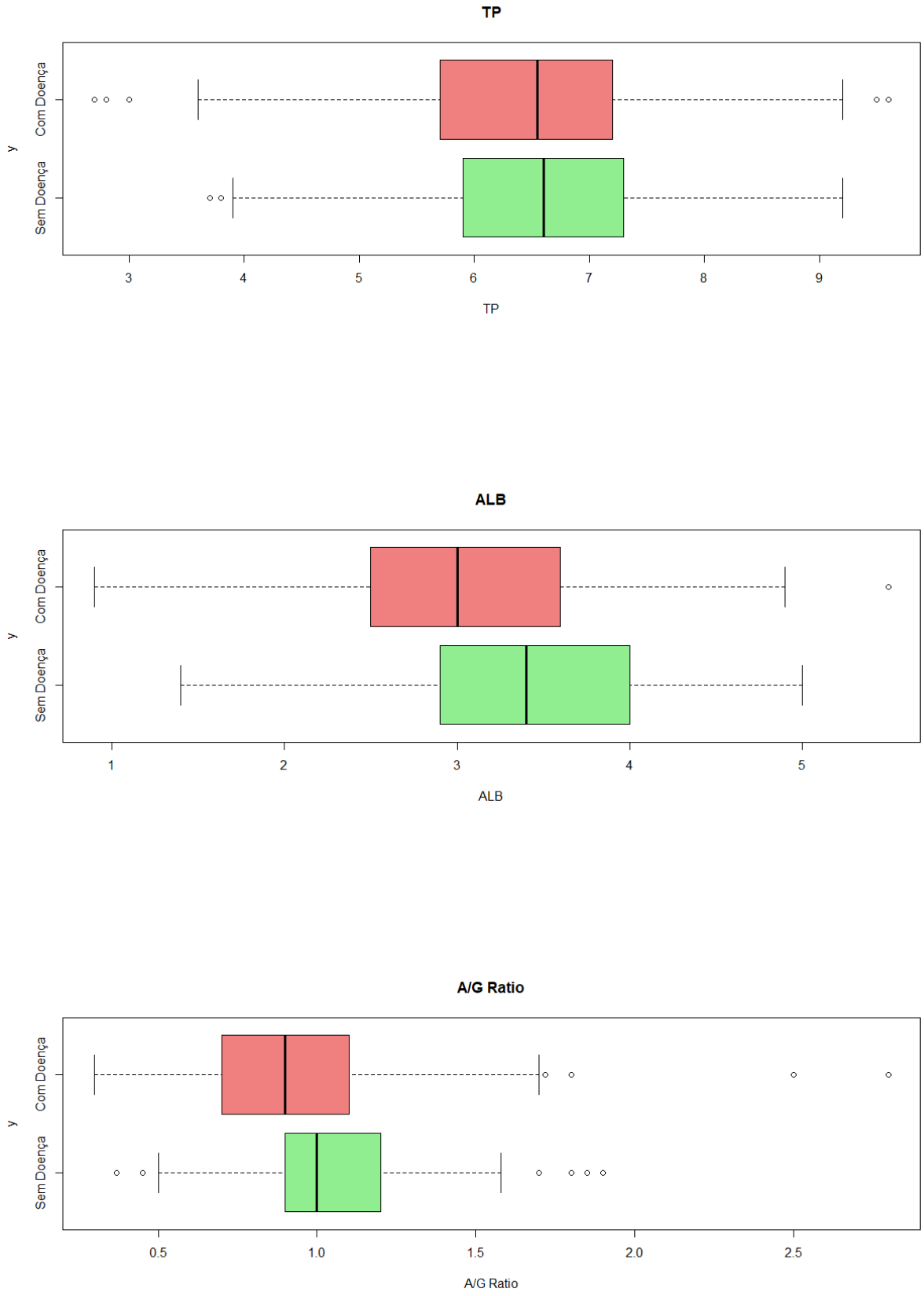


Figura A.1: Boxplots das variáveis numéricas