UNIVERSIDADE FEDERAL DE JUIZ DE FORA
INSTITUTO DE CIÊNCIAS EXATAS
DEPARTAMENTO DE ESTATÍSTICA

MARIA JÚLIA NEVES GREGÓRIO

REGRESSÃO LOGÍSTICA:

sob o ponto de vista inferencial e preditivista

Juiz de Fora
2021

MARIA JÚLIA NEVES GREGÓRIO

REGRESSÃO LOGÍSTICA:

sob o ponto de vista inferencial e preditivista

Trabalho de Conclusão de Curso apresentado ao Curso de Estatística da Universidade Federal de Juiz de Fora como requisito parcial para obtenção do título de Bacharel em Estatística

Orientador: Prof. Dr. Tiago Maia Magalhães

Coorientador: Prof. Dr. Márcio Augusto Diniz

Juiz de Fora

2021

Ficha catalográfica elaborada através do programa de geração
automática da Biblioteca Universitária da UFJF,
com os dados fornecidos pelo(a) autor(a)

Neves Gregório, Maria Júlia.
Regressão Logística : sob o ponto de vista inferencial e preditivista / Maria Júlia Neves Gregório. -- 2021.
40 f. : il.

Orientador: Tiago Maia Magalhães
Coorientador: Márcio Augusto Diniz
Trabalho de Conclusão de Curso (graduação) - Universidade Federal de Juiz de Fora, Instituto de Ciências Exatas, 2021.

1. Regressão Logística. 2. Estimação. 3. Predição. 4. Lasso. 5. Ridge. I. Maia Magalhães, Tiago, orient. II. Augusto Diniz, Márcio, coorient. III. Título.

MARIA JÚLIA NEVES GREGÓRIO

REGRESSÃO LOGÍSTICA:

Sob o ponto de vista inferencial e preditivista

Trabalho de Conclusão de Curso apresentado ao Curso de Estatística da Universidade Federal de Juiz de Fora, como requisito parcial para obtenção do título de Bacharel em Estatística

Aprovada em:

BANCA EXAMINADORA

_____

Prof. Dr. Tiago Maia Magalhães - Orientador
Universidade Federal de Juiz de Fora

_____

Prof. Dra. Camila Borelli Zeller
Universidade Federal de Juiz de Fora

_____

Prof. Dr. Clécio da Silva Ferreira
Universidade Federal de Juiz de Fora

Resumo

Ao construir um modelo, nós temos a opção de duas perspectivas, uma inferencial, que possui enfoque maior na interpretação das variáveis explicativas e a outra que prioriza a previsão de novas observações. Este trabalho apresenta as ferramentas de ambas pespectivas para o modelo de Regressão Logística. Para isso será apresentado a estimação por máxima verossimilhança, testes para ajuste do modelo, para o primeiro caso e, artifícios usados para a melhora da previsão, como Validação Cruzada e penalizações como Lasso e Ridge para o segundo caso.

Palavras-chave: Regressão Logística, Máxima verossimilhança, Predição, Lasso, Ridge.

Abstract

When building a model, we have the option of two perspectives, an inferential one, which has a greater focus on the interpretation of explanatory variables and the other that prioritizes the prediction of new observations. This work intends to present tools from both perspectives for the Logistic Regression model. For this, the estimation by maximum likelihood will be presented, tests to adjust the model, for the first case, and devices used to improve the forecast, such as Cross Validation and penalties such as Lasso and Ridge for the second case.

Keywords: Logistic Regression, maximum likelihood, prediction, Lasso, Ridge.

Agradecimentos

Gostaria de agradecer primeiramente a Deus, aos meus pais, a minha irmã. Aos meus tios, primos e afilhados. Fizeram parte da minha base e são meu suporte.

Aos amigos de início de faculdade, que trilharam outros caminhos e, estão sempre presente nos meus pensamentos.

A Isabela, amiga que encontrei nessa jornada, a qual compartilhamos experiências e se manteve presente ao longo dela.

Aos meus amigos de longa data, Carol, Jordana e Mattheus, pelo apoio e companheirismo de sempre.

Aos professores do Departamento de Estatística que contribuíram para meu aprendizado, em especial ao meu orientador, Tiago Maia Magalhães pelo exemplo, incentivo e paciência.

E a todas pessoas que passaram pela minha vida nesse período e que contribuíram direta ou indiretamente com essa conquista.

Conteúdo

Lista de Figuras

Lista de Tabelas

1 Introdução

Os primeiros conceitos de Regressão foram propostos por Galton (1889), aplicados na Antropometria, em que o objetivo era equacionar as relações de dependência entre a altura dos pais (variável explicativa) e a altura dos filhos (variável de interesse ou variável resposta ou desfecho). Porém esses conceitos podem ser aplicados em qualquer contexto, como por exemplo, a necessidade de uma empresa em analisar os fatores (variáveis explicativas) que podem interferir nas vendas (variável resposta) ou na predição de mortalidade de uma pessoa ao contrair uma doença, baseada nas suas características, como idade, sexo e doenças pré existentes. Por via de regra, sempre teremos uma equação, em que uma função de ligação será responsável por relacionar um parâmetro da variável resposta com as variáveis explicativas. A versatilidade dos modelos de regressão faz com que esta área esteja em constante crescimento.

A ideia inicial dos primeiros modelos de regressão contava com a normalidade da variável resposta e uma associação linear da média com variáveis preditoras, como nem sempre estas suposições são atendidas, é necessário pensar em estratégias como a transformação dos dados. Contudo, a mesma muitas vezes acaba dificultando a análise dos dados, pois estes estarão com uma nova unidade e, por vezes, não são capazes de corrigir certos pontos. Uma outra solução para o problema seria o uso dos Modelos Lineares Generalizados (MLG), que são uma extensão dos modelos de regressão e, permitem a utilização de outras distribuições para a variável resposta. Neste trabalho, nós estamos interessados em situações em que a variável resposta é dicotomica, isto é, admite apenas dois resultados (0/1, não/sim, falso/verdadeiro). Este é um objetivo muito comum para análises de crediting scoring, onde é aprovado ou não um pedido de crédito segundo o risco de inadimplência do cliente, ou em ensaios clínicos em que pesquisas investigativas com o interesse de analisar a ação de medicamentos como a eficácia do mesmo, por exemplo.

Nós podemos pensar em um modelo de regressão sob duas perspectivas específicas: preditiva (prever variáveis de interesse), ou inferencial (foco maior na interpretação e na associação das variáveis). Dentro de cada objetivo existem duas abordagens possíveis:

A da Estatística, na qual os resultados da previsão limitados e a da Aprendizagem de Máquina, na qual os algoritmos não se prendem a suposições limitantes, como a necessidade do tamanho da amostra ser maior que o número de variáveis.

Neste trabalho abordaremos o Modelo de Regressão Logística, quando os dados de resposta são categóricos. No Capítulo 2, nós apresentaremos os conceitos de Regressão Logística sob a perspectiva da Estatística Clássica, já o Capítulo 3, vemos modificações que auxiliam na produção de bons resultados, quanto a predição. Nos Capítulos 4 e 5 veremos alguns conceitos aplicados em simulações e base de dados reais.

2 Regressão Logística

O Modelo de Regressão Logística é aconselhado quando as variáveis dependentes, Y, são categóricas, mais especificamente binárias, que são comumente denominadas de sucesso (quando $Y = 1$) e fracasso (para $Y = 0$). Inicialmente nós podemos pensar em uma estrutura que relacione um parâmetro π (proporção de sucesso) a uma função linear de p covariáveis. No entanto, nós sabemos que uma proporção deve estar contida no intervalo $(0, 1)$, enquanto uma equação linear pode assumir qualquer valor real, então a solução é encontrar uma função que ligue o parâmetro às covariáveis, mas que garanta que a restrição seja satisfeita. Na literatura existem várias funções que satisfazem essa restrição, como a log-log, probit e logit, mas em nossos estudos nós usaremos apenas a função logit, a qual os resultados são mais facilmente interpretados.

2.1 Modelo

Seja Y uma variável aleatória, com distribuição Bernoulli(π), em que $\pi = P(Y = 1|\boldsymbol{x})$ é a probabilidade de sucesso dado um conjunto de variáveis independentes, $\boldsymbol{x} = (x_1, x_2..x_p)^\top$ e $\boldsymbol{\beta} = (\beta_1, \beta_2, ..., \beta_p)^\top$ o vetor de parâmetros desconhecidos a serem estimados. A função logística é dada por:

$$\text{logit}(\pi_i) = \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_p x_{pi},$$

em que

$$\text{logit}(\pi_i) = \log\left(\frac{\pi_i}{1-\pi_i}\right). \tag{2.1}$$

Nós podemos reescrever (2.1), em função de π_i, em que $x_{1i} = 1$ para $i = (1, \ldots, n)$ e β_1 é

o intercepto, da seguinte forma:

$$\pi_i = \frac{\exp(\beta_1 + \beta_2 x_{2i} + ... + \beta_p x_{pi})}{1 + \exp(\beta_1 + \beta_2 x_{2i} + ... + \beta_p x_{pi})}.$$

2.2 Estimação

Para estimarmos o vetor de parâmetros $\boldsymbol{\beta}$, usaremos o método da máxima verossimilhança (EMV). Além das suposições já feitas, é necessário assumir que as n amostras, $Y_1 \ldots, Y_n$, são independentes.

A função de verossimilhança é dada por:

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} L_i(\boldsymbol{\beta}) = \prod_{i=1}^{n} \pi_i^{y_i}(1-\pi_i)^{1-y_i}. \tag{2.2}$$

Aplicando o logaritmo natural na equação (2.2), temos que

$$\begin{aligned} \ell(\boldsymbol{\beta}) = \ell(\boldsymbol{\beta}|\boldsymbol{y}, \boldsymbol{X}) &= \sum_{i=1}^{n} \{y_i \log \pi_i + (1-y_i)\log(1-\pi_i)\} && (2.3) \\ &= \sum_{i=1}^{n} y_i \log \frac{\pi_i}{(1-\pi_i)} + \sum_{i=1}^{n} \log(1-\pi_i) \\ &= \sum_{i=1}^{n} y_i(\beta_1 x_{1i} + \beta_2 x_{2i} + ... + \beta_p x_{pi}) && (2.4) \\ &\quad - \sum_{i=1}^{n} \log(1 + \exp(\beta_1 x_{1i} + \beta_2 x_{2i} + ... + \beta_p x_{pi})). \end{aligned}$$

Derivando em relação aos parâmetros, temos que:

$$\frac{\delta \ell}{\delta \beta_j} = -\sum_{i=1}^{n} x_{ij} y_i - \sum_{i=1}^{n} x_{ij} \pi(x_i) = 0, \forall j \in \{1, 2, ..., p\} \tag{2.5}$$

Para encontrar os estimadores $\hat{\boldsymbol{\beta}}$, é necessario resolver a equação:

$$\sum_{i=1}^{n} x_{ij}(y_i - \pi(x_i)) = 0. \tag{2.6}$$

para $j \in \{1, 2.., p\}$. O vetor score $\boldsymbol{U}(\beta)$ pode ser escrito como:

$$\boldsymbol{U}(\boldsymbol{\beta}) = \boldsymbol{X}^\top(\boldsymbol{y} - \boldsymbol{\pi}), \tag{2.7}$$

em que, $\boldsymbol{X}^\top = (\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n)$, $\boldsymbol{y} = (1, 2, \cdots, n)^\top$, $\boldsymbol{\pi} = (\pi_1, \cdots, \pi_n)^\top$.

Não é possível resolver a equação analiticamente, já que é uma forma não linear nos parâmetros, mas podemos encontrar um resultado aproximado usando métodos numéricos, aqui optamos pelo método de Newton-Raphson.

A matriz de informação de Fisher obtida derivando a função de verossimilhança duas vezes é dada por:

$$\boldsymbol{K} = \boldsymbol{K}(\boldsymbol{\beta}) = \boldsymbol{X}^\top \boldsymbol{\Pi}^{(1)} \boldsymbol{X}, \tag{2.8}$$

em que $\boldsymbol{\Pi}^{(1)} = \text{diag}\{\pi_1(1-\pi_1), \ldots, \pi_n(1-\pi_n)\}$. Sob condições de regularidades usuais para os estimadores de máxima verossimilhança (ver, por exemplo, p. 245, Sen, Singer e Lima (2010)), temos, quando o tamanho amostral é grande, que

$$\widehat{\boldsymbol{\beta}} \sim \text{N}(\boldsymbol{\beta}, \boldsymbol{K}^{-1}),$$

em que $\boldsymbol{K}^{-1}$ é inversa da matriz de informação de Fisher, dada em (2.8).

2.3 Estatísticas de teste

Quando uma regressão é feita, o pesquisador deseja conseguir representar os dados de forma mais fiel a realidade possível, de modo que, testar os coeficientes é uma etapa importante a ser seguida, só assim será possível determinar se as variáveis explicativas

pré-selecionadas interferem, de fato, no modelo.

Considere que o vetor de paramêtros $\boldsymbol{\beta}$ pode ser particionado como $\boldsymbol{\beta} = (\boldsymbol{\beta_1}^\top, \boldsymbol{\beta_2}^\top)^\top$, em que $\boldsymbol{\beta}_1 = (\beta_1, \ldots, \beta_q)^\top$ é um vetor q-dimensional de parâmetros de interesse e $\boldsymbol{\beta}_2 = (\beta_{q+1}, \ldots, \beta_p)^\top$, um vetor de parâmetros de pertubação, $(p-q)$-dimensional. Nós queremos testar as hipóteses

$$\mathcal{H} : \boldsymbol{\beta}_1 = \boldsymbol{\beta}_1^{(0)} \text{ vs } \mathcal{A} : \boldsymbol{\beta}_1 \neq \boldsymbol{\beta}_1^{(0)},$$

sendo $\boldsymbol{\beta}_1^{(0)}$ um vetor especificado. Sejam $\widehat{\boldsymbol{\beta}} = \left(\widehat{\boldsymbol{\beta}}_1^\top, \widehat{\boldsymbol{\beta}}_2^\top\right)^\top$ e $\widetilde{\boldsymbol{\beta}} = \left(\boldsymbol{\beta}_1^{(0)\top}, \widetilde{\boldsymbol{\beta}}_2^\top\right)^\top$, as estimativas de máxima verossimilhança irrestrita e restrita de $\boldsymbol{\beta}$ a $\mathcal{H}$, respectivamente. A partição $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^\top, \boldsymbol{\beta}_2^\top)^\top$ conduz às correspondentes partições no vetor escore $\boldsymbol{U}(\boldsymbol{\beta}) = \left(\boldsymbol{U}(\boldsymbol{\beta}_1)^\top, \boldsymbol{U}(\boldsymbol{\beta}_2)^\top\right)^\top$, com $\boldsymbol{U}(\boldsymbol{\beta}_1) = \boldsymbol{X}_1^\top(\boldsymbol{y} - \boldsymbol{\pi})$, $\boldsymbol{U}(\boldsymbol{\beta}_2) = \boldsymbol{X}_2^\top(\boldsymbol{y} - \boldsymbol{\pi})$, na matriz de informação de Fisher e na sua inversa, dadas por

$$\boldsymbol{K} = \begin{pmatrix} \boldsymbol{K}_{11} & \boldsymbol{K}_{12} \\ \boldsymbol{K}_{21} & \boldsymbol{K}_{22} \end{pmatrix} = \begin{pmatrix} \boldsymbol{X}_1^\top \boldsymbol{\Pi}^{(1)} \boldsymbol{X}_1 & \boldsymbol{X}_1^\top \boldsymbol{\Pi}^{(1)} \boldsymbol{X}_2 \\ \boldsymbol{X}_2^\top \boldsymbol{\Pi}^{(1)} \boldsymbol{X}_1 & \boldsymbol{X}_2^\top \boldsymbol{\Pi}^{(1)} \boldsymbol{X}_2 \end{pmatrix} \text{ e } \boldsymbol{K}^{-1} = \begin{pmatrix} \boldsymbol{K}^{11} & \boldsymbol{K}^{12} \\ \boldsymbol{K}^{21} & \boldsymbol{K}^{22} \end{pmatrix}, \tag{2.9}$$

em que $\boldsymbol{X} = [\boldsymbol{X}_1 \ \boldsymbol{X}_2]$, $\boldsymbol{X}_1$, $\boldsymbol{X}_2$ sendo $n \times q$ e $n \times (p-q)$, respectivamente. As estatísticas da razão de verossimilhanças (WILKS, 1938), Wald (WALD, 1943), escore (RAO, 1948) e gradiente (TERRELL, 2002) para testar $\mathcal{H}$ são definidas, respectivamente, como

$$\begin{aligned} S_1 &= 2\left[\ell(\widehat{\boldsymbol{\beta}}) - \ell(\widetilde{\boldsymbol{\beta}})\right], \\ S_2 &= \left(\widehat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1^{(0)}\right)^\top \left\{\widehat{\boldsymbol{K}}^{11}\right\}^{-1} \left(\widehat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1^{(0)}\right), \\ S_3 &= \boldsymbol{U}(\widetilde{\boldsymbol{\beta}}_1)^\top \widetilde{\boldsymbol{K}}^{11} \boldsymbol{U}(\widetilde{\boldsymbol{\beta}}_1), \\ S_4 &= \boldsymbol{U}(\widetilde{\boldsymbol{\beta}}_1)^\top \left(\widehat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1^{(0)}\right). \end{aligned} \tag{2.10}$$

Sob $\mathcal{H}$, as quatro estatísticas têm distribuição de qui-quadrado com q graus de liberdade, com um erro de ordem n^{-1}. A hipótese nula é rejeitada, para um dado nível nominal α e uma estatística especificada, se a estatística do teste for maior que o quantil

$100(1-\alpha)\%$ de uma χ^2_q. Buse (1982) apresenta de uma forma didática a interpretação geométrica dos testes da razão de verossimilhanças, escore e Wald para o caso de hipóteses simples, mais recentemente, Montoril e Souza (2013) fazem a mesma análise incluindo a estatística gradiente.

Especificamente, para o modelo de regressão logística, levando em conta o logarítmo da função de verosssimilhança, o vetor escore e a matriz de informação de Fisher dados, respectivamente, em (2.3) a (2.8), as estatísticas apresentadas em (2.10), podem ser escritas da seguinte forma:

$$
\begin{aligned}
S_1 &= 2\sum_{i=1}^{n}\left\{y_i \log\left(\frac{\hat{\pi}_i}{\tilde{\pi}_i}\right) + (1-y_i)\log\left(\frac{1-\hat{\pi}_i}{1-\tilde{\pi}_i}\right)\right\},\\
S_2 &= \left(\widehat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1^{(0)}\right)^{\top}\left(\widehat{\boldsymbol{R}}^{\top}\widehat{\boldsymbol{\Pi}}^{(1)}\widehat{\boldsymbol{R}}\right)\left(\widehat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1^{(0)}\right),\\
S_3 &= (\boldsymbol{y}-\widetilde{\boldsymbol{\pi}})^{\top}\boldsymbol{X}_1\left(\widetilde{\boldsymbol{R}}^{\top}\widetilde{\boldsymbol{\Pi}}^{(1)}\widetilde{\boldsymbol{R}}\right)\boldsymbol{X}_1^{\top}(\boldsymbol{y}-\widetilde{\boldsymbol{\pi}}),\\
S_4 &= (\boldsymbol{y}-\widetilde{\boldsymbol{\pi}})^{\top}\boldsymbol{X}_1\left(\widehat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1^{(0)}\right).
\end{aligned}
$$

em que $\boldsymbol{R} = \boldsymbol{X}_1 - \boldsymbol{X}_2\boldsymbol{C}$, com $\boldsymbol{C} = \left(\boldsymbol{X}_2^{\top}\boldsymbol{\Pi}^{(1)}\boldsymbol{X}_2\right)^{-1}\boldsymbol{X}_2^{\top}\boldsymbol{\Pi}^{(1)}\boldsymbol{X}_1$ representando uma matriz $q \times (p-q)$ cuja as colunas são os coeficientes de regressão estimados obtido em modelo de regressão normal ponderado entre as colunas de $\boldsymbol{X}_1$ e a matriz modelo $\boldsymbol{X}_2$ com $\boldsymbol{\Pi}^{(1)}$ como a matriz de ponderação.

2.4 Diagnóstico do Modelo

Por fim, a última etapa a ser seguida, é verificar se o modelo foi bem ajustado, por meio de uma análise de diagnóstico, cujo objetivo é usar técnicas que permitem saber se as suposições do modelo estão sendo satisfeitas, se há presença de *outliers* e se o modelo foi bem ajustado para o conjunto de variáveis ou se alguma teve uma influência maior na estimação do que as demais.

2.4.1 Resíduo de Pearson e Deviance

Um dos principais pontos a serem analisados são os resíduos (diferença entre um valor estimado e o valor observado). Geralmente os resíduos são calculamos para J

padrões de covariáveis no lugar de cada observação X_i. Para isso, nós denotaremos m_j como o número total de ensaios ou indivíduos com $\boldsymbol{x} = \boldsymbol{x}_j$, para $j = 1, 2, ..., J$

O resíduo de Pearson é dado por:

$$r(y_j; \hat{\pi_j}) = \frac{(y_j - m_j\hat{\pi_j})}{\sqrt{m_j\hat{\pi_j}(1 - \hat{\pi_j})}}.$$

O resíduo Deviance é dado por:

$$d(y_j; \hat{\pi_j}) = \pm\sqrt{2\left[y_j \log\left(\frac{y_j}{m_j\hat{\pi_j}}\right) + (m_j - y_j)\log\left(\frac{(m_j - y_j)}{m_j(1 - \hat{\pi_j})}\right)\right]}.$$

O sinal será definido pelo sinal de $(y_j - m_j\hat{\pi_j})$

A estatística será a soma de quadrados do resíduo, e são dados respectivamente por:

$$X^2 = \sum_{j=1}^{J} r(y_j; \hat{\pi_j})^2,$$

$$D = \sum_{j=1}^{J} d(y_j; \hat{\pi_j})^2.$$

Ambas as estatísticas tem distribuição qui-quadrado com $J - (p + 1)$ graus de liberdade, e sob hipótese nula de que o modelo com $J - (p+1)$ é o correto, quanto menor os valores das estatísticas, melhor será o ajuste do modelo

2.4.2 Teste de Hosmer e Lemeshow

O teste de Hosmer e Lemeshow avalia o modelo pela diferença entre as probabilidades ajustadas e as probabilidades observadas. Nós podemos representar a hipótese nula como o modelo está bem ajustado e por consequência a hipótese alternativa como o modelo não está bem ajustado.

O teste é feito ordenando-se as probabilidades, e então criando g grupos (um valor usual é $g = 10$, possuindo aproximadamente o mesmo tamanho), suponhamos que se tenha 100 observações: o primeiro grupo contará com as 10 menores probabilidades, o segundo grupo com as 10 menores probabilidades restantes, e assim por diante até que o útimo grupo contenha as 10 maiores probabilidades. A partir da divisão dos grupos

observaremos o número de sucessos ($y_j = 1$) correspondentes, e calcularemos os valores esperados em cada grupo. Para grandes amostras nós podemos compará-los usando a estatistica qui-quadrado de Pearson com $(g - 2)$ graus de liberdade.

2.4.3 Leverage values (ou pontos de alavanca)

A matriz de projeção para o modelo de regressão logística, é definida por:

$$\mathbf{H} = \mathbf{Q^{1/2}X(X^\top QX)^{-1}X^\top Q^{1/2}},$$

em que, $\boldsymbol{Q} = \text{diag}[\boldsymbol{\pi}(\boldsymbol{x})(\mathbf{1} - \boldsymbol{\pi}(\boldsymbol{x}))]$.

A diagonal principal da matriz $\boldsymbol{H}$ contém os pontos de alavanca (são as observações extremas de $\boldsymbol{X}$, ou seja, é uma análise das variáveis explicativas), e pode ser expressa por:

$$\hat{h_{ii}} = \hat{\pi}(\boldsymbol{x_i})(1 - \hat{\pi}(\boldsymbol{x_i}))\boldsymbol{x_i}^\top[\boldsymbol{K(\hat{\beta})}]^{-1}\boldsymbol{x_i}; i = 1, 2, ..., n.$$

2.5 R^2 Ajustado

Como a variável resposta não é contínua como na Regressão Linear, não é possível calcular R^2. Mas existem estatísticas com o objetivo parecido conhecidas como R^2 ajustado ou pseudoR^2. Nós usaremos a modificação proposta por Cox e Sneel (1989), representada por:

$$R^2 = 1 - \left(\frac{L(\beta)_0}{L(\beta)_M}\right)^{\frac{2}{n}},$$

em que $L(\beta)_0$ é a função de verossimilhança sem variáveis explicativas, $L(\boldsymbol{\beta})_M$ é a função de verossimilhança do modelo estimado e, n é o tamanho da amostra.

Diferente do R^2, o valor máximo não chega a 1. Como não existe um valor considerado ótimo na literatura, muitas vezes é usado apenas para comparar o desempenho de modelos concorrentes e, prefere-se o modelo que possuir um valor mais elevado.

2.6 Interpretação do Modelo

Após a consolidação do modelo, o nosso interesse agora é interpretá-lo. O modelo com a função de ligação logit possui vantagem sobre os demais nesse quesito: a possibilidade de calcular a razão de chances, também conhecida como *Odds Ratio*, relacionando a probabilidade de sucesso ($\pi(x) = 1$) à probabilidade de fracasso ($\pi(x) = 0$).

Seja x_{j+1} e x_j variáveis explicativas, suas chances de sucesso podem ser descritas respectivamente por:

$$odds_{x_{j+1}} = \frac{\pi_{x_{j+1}}}{1 - \pi_{x_{j+1}}} = \exp(\beta_1 + \beta_2 x_{j+1})$$

e

$$odds_{x_j} = \frac{\pi_{x_j}}{1 - \pi_{x_j}} = \exp(\beta_1 + \beta_2 x_j).$$

O resultado a seguir indica a chance do indivíduo x_{j+1} ter sucesso em relação ao indivíduo x_j:

$$OR_{x_{j+1}/x_j} = \frac{odds_{x_{j+1}}}{odds_{x_j}} = \frac{\exp(\beta_1 + \beta_2 x_{j+1})}{\exp(\beta_1 + \beta_2 x_j)} = \exp(\beta_2(x_{j+1} - x_j)).$$

Se a diferença de x_{j+1} e x_j for de 1 unidade, o resultado se reduz à $\exp(\beta_2)$.

2.7 Predição

Para predizer novas observações é necessário comparar a probabilidade estimada π_i com um ponto de corte, pertencente ao intervalo (0,1), e dessa forma, classificar:

$$\hat{Y} = \begin{cases} 1, & \text{caso a probabilidade estimada seja maior que o ponto de corte,} \\ 0, & \text{caso contrário.} \end{cases}$$

Existem duas formas comumente utilizadas para definição dos pontos de corte, a primeira é um método mais simples, já que é só adotar como ponto de corte o valor $0,5$.

Apesar de intuitivo, esse método não garante a minimização de predições erradas.

O segundo método, é através da curva ROC ou *Receiver Operating Characteristic Curve*, obtida calculando a sensibilidade ($P(\hat{Y} = 1|Y = 1)$)) e a especificidade ($P(\hat{Y} = 0|Y = 0)$) para cada ponto de corte, no gráfico representamos (1-especificidade, sensibilidade). O objetivo é encontrar um ponto que maximiza tanto a especificidade quanto a sensibilidade, isto é, o ponto que maximiza a classificação certa.

Uma aplicação comum quando o objetivo é a predição em Regressão Logística, são os modelos de Dose-Resposta, cujo interesse principal é a predição da probabilidade de sucesso $\pi(x)$ para uma dosagem x. Além disso, também visam encontrar a dose letal (dosagem necessária para uma probabilidade de sucesso π). É importante observar que os modelos de dose-resposta não se limitam a Toxicologia, em outros casos, a dose pode representar uma idade ou uma resistência de material por exemplo.

Apesar de útil a predição em modelos de regressão, seu uso deve ser restrito. Frost (2020) ilusta situações em que a predição não é aconselhável, como a predição fora do intervalo dos dados que estimam o modelo, ou a previsão para uma população diferente da pertencente a amostra.

3 Aprendizado supervisionado

No Capítulo 2, o interesse era principalmente inferencial, neste tipo de situação todos os dados são utilizados para encontrar uma curva sigmóide. O método da máxima verossimilhança, pela propriedade de consistência para grandes amostras, tende a produzir estimadores não viesados, ou seja, o modelo tende a super ajustar a amostra (em *machine learning*, nós conceituamos essa situação como *overfitting*). Depois de estimado, nós temos interesse em interpretá-lo. Embora o modelo seja o de melhor ajuste, não há garantias que ele fará uma boa predição para novas observações, pois apesar do viés, que é a distância entre o valor observado e a sigmóide estimada, ser o menor possível, a variância que mede a distância entre o valor das novas observações e o valor predito pela sigmóide, tende a ser grande quando o modelo está super ajustado.

Quando nós estamos interessados na previsão, uma alternativa aos métodos estatísticos tradicionais é a divisão do conjunto de dados entre treinamento, em que o propósito será encontrar um modelo e, uma outra parte de teste, que nós usaremos para analisar o desempenho do modelo, testando-o. A ideia do aprendizado de máquinas é deixar que os métodos (ou algoritmos) aprendam com os dados disponíveis no problema e façam modelos.

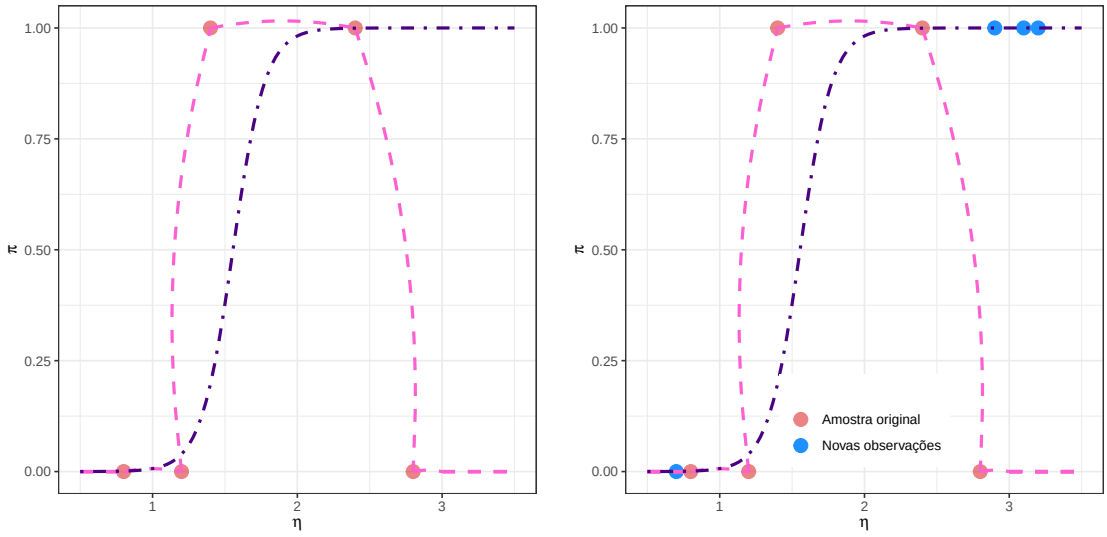
Essa área possui certas subdivisões, que se diferenciam pelo estado em que o dado é entregue ao problema, dentre elas há o aprendizado supervisionado, que os dados apresentam variáveis independentes (ou entradas) e dependentes (ou saídas), ou seja foram especificadas (ou rotuladas) pelo fornecedor. Já o aprendizado não supervisionado, não possui variável de saída. Este trabalho será voltado para o primeiro caso.

Diferentemente da terminologia estatística que considera os problemas como sendo de regressão, seja a variável quantitativa ou qualitativa, em aprendizado supervisionado há distinção da denominação de acordo com a variável resposta (desfecho). Se a variável for quantitativa é referido como um problema de regressão, e se for qualitativa, como um problema de classificação. James et al. (2013) comentam sobre essa distinção de terminologias não ser tão nítida. Mas é possível perceber isso, uma vez que ambos modelos fazem

parte da classe de Modelos Lineares Generalizados, propostos por Nelder e Wedderburn (1972).

3.1 Viés e Variância

Para ilustrar os conceitos de viés e variância, nós faremos uso de um exemplo simulado de uma amostra contendo 5 observações. Na Figura 3.1 (a), nós temos os dados ajustados por duas funções, uma delas é uma sigmóide ajustada por Regressão Logística e, a outra uma densidade de uma distribuição uniforme. Nós podemos ver que a densidade da uniforme se ajusta muito bem aos dados, já a sigmóide não se ajustou tão bem e, consequentemente apresentou um alto viés.



(a) (b)

Figura 3.1: Amostra ajustada na curva uniforme (em rosa) e na curva sigmóide (em roxo).

Agora suponhamos que novas observações surgiram, e a Figura 3.1 (b) mostra a relação das novas observações com as curvas obtidas anteriormente. A curva uniforme que antes estava bem ajustada não representa bem as novas observações, apresentando uma alta variância, diferentemente da curva sigmóide que apesar de não ter ajustado perfeitamente as primeiras observações ajustou bem às demais. Com relação a curva uniforme, podemos perceber que essas observações também estão fora do intervalo de estimação e, como visto na Seção 2.7, isso não é recomendado.

Para evitar situações como essa, em que os novas observações não se adequam

ao modelo ajustado, precisamos adotar novos artifícios, como a técnica de particionar o conjunto de dados entre uma parte de treino para o ajuste e outra de teste para validação. Dessa maneira é possível testar o poder de generalização do modelo. A forma usada para partição é a Validação Cruzada, que será vista na próxima seção.

3.2 Validação Cruzada

Na Seção anterior vimos que particionar o conjunto de dados era importante para melhorar a qualidade da previsão, mas ainda é necessário saber a melhor proporção entre dados de treino e teste. Para essa definição, nós usaremos a Validação Cruzada k- fold (k- *fold cross validation*), que é um método que particiona o conjunto em k grupos.

Com uma amostra de tamanho $n = 100$, por exemplo e, k sub-amostras de tamanho 20 é possível estimar 5 modelos logísticos diferentes com a mesma abordagem do Capítulo 2. Como é possível visualizar na Figura 3.2, após a divisão, o primeiro subconjunto será usado para teste e os demais para estimação, logo em seguida o processo será repetido iterativamente. A cada iteração é possível mensurar o modelo para que ao fim do processo o melhor modelo seja escolhido. A forma mais comum para fazer isso, é usando a matriz de confusão, que será vista posteriormente, mas é possível também usar a Deviance ou o Desvio de Pearson, descritos na Seção 2.4.1.

Não existe critério definido para a escolha do k, é possível repetir o processo descrito acima para diferentes tamanhos, nesse caso, será feita a média das métricas obtidas para cada k. O k escolhido será aquele que apresentar melhor média. Entretanto realizar essa etapa, pode acarretar em um alto custo computacional.

3.3 Estimação

O método de estimação da Regressão Logística, definida na Seção 2.1, continua válido para predição, desde que a amostra seja dividida em duas partes: treino e teste. E apenas a parte do treino seja usada na estimação. Essa divisão é essencial para melhora dos resultados de previsão, apesar disso podemos ter o interesse de reduzir ainda mais a variância e, por conta disso, nós podemos optar por modelos viesados, mas que possuem

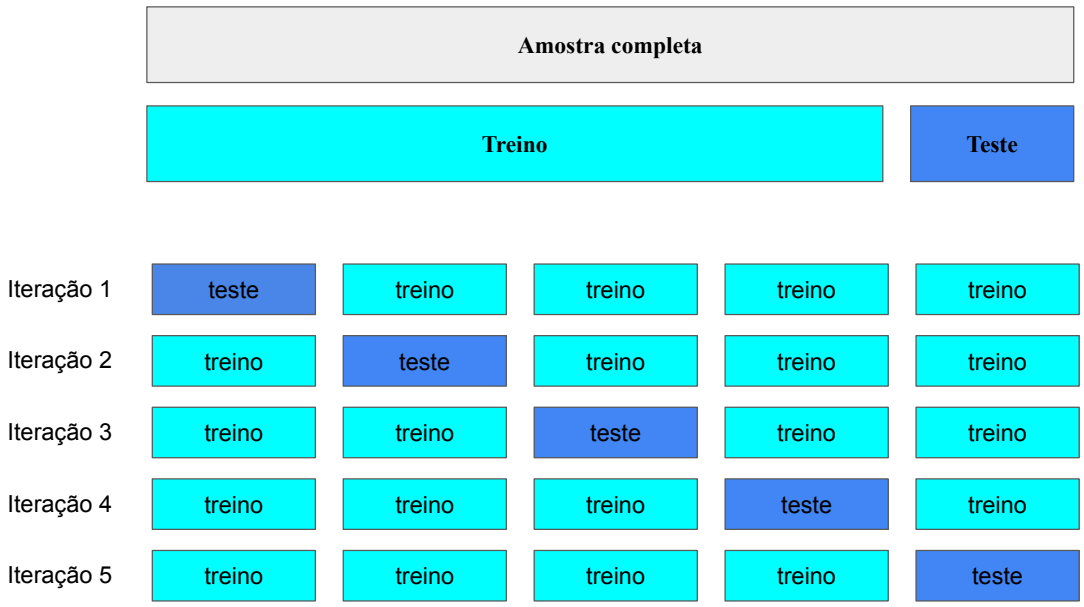


Figura 3.2: Esquema para Validação Cruzada

variância menor. É interessante observar a diferença em relação a ótica inferencial, que tinha como propósito encontrar estimadores com o menor viés possível.

3.4 Penalização

Anteriormente, vimos no Método de *Cross Validation* uma maneira de reduzir a variância, mas mesmo separando o conjunto de dados, a estimação por máxima verossimilhança tende a produzir estimadores que buscam o super ajuste. Existem outros métodos que possuem a finalidade de diminuir o ajuste excessivo dos dados, como o método Regularização que veremos a seguir e, tem como objetivo penalizar o modelo por meio da inserção de viés.

3.4.1 Regressão Ridge

A Regressão Ridge (HOERL; KENNARD, 1970) foi originalmente proposta para contornar a multicolinearidade acrescentando uma penalização por meio da norma L_2, também conhecida como penalização quadrática. A função de máxima verossimilhança penalizada pode ser descrita como:

$$l_R(\boldsymbol{\beta}) = l(\boldsymbol{\beta}) - \lambda \sum_{j=2}^{p} \beta_j^2, \tag{3.1}$$

em que $l(\boldsymbol{\beta})$ é o logaritmo da função de verossimilhança da Regressão Logística, definida em (2.3), λ é um hiperparâmetro, em outras palavras, ele é o peso do viés adicionado na verossimilhança original. Os parâmetros β_j são obtidos através da maximização de (3.1) e λ é comumente estimado pelo método de Validação Cruzada.

Para evitar erros nos resultados que possam vir a ocorrer pela diferença de escala das variáveis é conveniente padronizar pela média (ver, p.63, Hastie, Tibshirani e Friedman (2009)). Além disso, β_1 não entra na penalização e pode ser obtido pela média de $\boldsymbol{y}$.

3.4.2 Regressão Lasso

A Regressão Lasso, *least absolute shrinkage and selection operador*; (TIBSHIRANI, 1996), surgiu com intuito de selecionar variáveis, ou seja, algumas estimativas dos parâmetros podem assumir o valor 0 quando há variáveis com alta correlação.

A penalização é obtida pela norma L_1 e, assim como na Regressão Ridge o hiperparâmetro λ pode ser estimado por Validação Cruzada e, também é recomendado que as variáveis independentes sejam padronizadas. A seguir, a função de máxima verossimilhança penalizada:

$$l_L(\boldsymbol{\beta}) = l(\boldsymbol{\beta}) - \lambda \sum_{j=2}^{p} |\beta_j|. \tag{3.2}$$

A seleção de variáveis torna o método atrativo quando comparado à Regressão Ridge, pois um modelo com menos variáveis pode ser interpretado mais facilmente.

3.4.3 Penalização Firth

A proposta do método de Firth (FIRTH, 1993) surgiu para correção de viés, porém diferentemete do seu objetivo inicial, nós podemos usá-lo também para acrescen-

tar viés. Para distribuições pertencentes a família exponencial, a penalização otimiza a seguinte função:

$$l_F(\boldsymbol{\beta}) = l(\boldsymbol{\beta}) + \frac{1}{2}\log|\boldsymbol{K}|,$$

em que, como visto anteriomente $l(\boldsymbol{\beta})$ é o logaritmo da função de verossimilhança da Regressão Logística, $\boldsymbol{K}$ é a matriz de informação de Fisher, definida em (2.9) e, $|.|$ é a notação para determinante. Diferente das penalizações anteriores, λ é um valor fixo (1/2).

Note que os métodos das Seções 3.4.1 e 3.4.2 também foram criados com objetivos diferentes à inserção de viés ao modelo, mas são comumente usadas para esta finalidade.

3.4.4 Penalização Uniforme

Diferente das outras penalizações descritas acima, em que os novos parâmetros já são calculados de imediato, para a penalização Uniforme (HOUWELINGEN; CESSIE, 1990) é necessário calcular o fator S_{LU}, baseado na estatística de razão de verossimilhança do modelo ajustado e g, que é o número de graus de liberdade obtido pelo número de preditores do modelo. O fator pode ser representado da seguinte forma:

$$S_{LU} = \frac{\chi^2_{model} - g}{\chi^2_{model}}.$$

Após o cálculo do fator e estimação do modelo, β_j penalizado será:

$$\hat{\beta}_{jLU} = S_{LU}\hat{\beta}_j.$$

3.5 Qualidade do Ajuste

Note que, nós apresentamos diversos modelos com o mesmo objetivo, diminuir a variância e consequentemente melhorar a predição. É desejável saber o modelo mais adequado para seus dados, a seguir nós apresentaremos medidas que irão pssibilitar com-

pará-los.

3.5.1 Matriz de Classificação

A matriz de classificação ou confusão pode ser obtida contabilizando as observações que foram preditas corretamente (n_{00} e n_{11}) e erroneamente (n_{01} e n_{10}). As linhas representam os valores observados e, as colunas os valores preditos de acordo com o ponto de corte, geralmente nós usamos a média de y_i para essa classificação.

A partir da construção da matriz de confusão é possível calcular algumas métricas. As mais comuns são:

Acurácia: com ela é possível mensurar a porcentagem de predições corretas do modelo. É necessário se atentar aos casos em que a matriz apresenta muitas observações em apenas uma lacuna. Podemos representá-la da seguinte forma:

$$\frac{n_{00} + n_{11}}{n}.$$

Precisão: Mensura a porcentagem de predições positivas classificadas corretamente, dentre todas predições positivas. Devemos nos atentar ao fato que os falsos positivos não são contabilizados. Nós podemos descrevê-la como:

$$\frac{n_{11}}{n_{11} + n_{01}}. \tag{3.3}$$

Recall ou sensibilidade: Mensura a porcentagem positiva observada que foi classificada corretamente, isto é:

$$\frac{n_{11}}{n_{11} + n_{10}}. \tag{3.4}$$

Nós podemos observar que o que difere as métricas (3.3) e (3.4) é o denominador, enquanto na Precisão, contabilizamos apenas os valores preditos, no Recall, contabilizamos

os valores observados. É importante ressaltar que apenas uma métrica não é capaz de classificar o modelo como bom, sendo necessário fazer uma análise conjunta.

Tabela 3.1: Matriz de Confusão

		Predito		Total
		$\hat{y}_i = 0$	$\hat{y}_i = 1$	
Observado	$y_i = 0$	n_{00}	n_{01}	$n_{0.}$
	$y_i = 1$	n_{10}	n_{11}	$n_{1.}$
Total		$n_{.0}$	$n_{.1}$	n

3.5.2 Area Under the ROC Curve - AUC

Além de determinar o melhor ponto de corte para classificação de y_i, é possível usar a curva ROC para comparar modelos concorrentes. A aréa abaixo à curva representa a capacidade do classificador em separar as duas classes. Nós podemos entender esse método como uma extensão da Matriz de Confusão, vista na Seção 3.5.1 e, por gerar apenas um valor para análise, é mais usual.

Nós desejamos que a curva apresente alta sensibilidade (valores proximos a 1 no eixo Y), e baixa $1-$ especificidade ou a probabilidade de classificar erroneamente o evento como sucesso (valores próximos a 0 representados no eixo X). A área está contida no intervalo $(0, 1)$ e, quanto mais próximo de 1, maior será o poder preditivo do modelo.

4 Simulação

Com auxílio do Software R Core Team (2020) nós simulamos três cenários de regressão logística, como definido em 2.1, pelo método de Monte Carlo, com 10000 replicas e 5 tamanhos de amostras distintos, com $n = (50, 70, 90, 120, 200)$. As covariáveis seguem distribuição normal padrão e, estão fixas em cada uma das 10000 replicas. No primeiro cenário, nós teremos $\boldsymbol{\beta} = (0{,}1; 0{,}1; 0{,}1)^\top$ e proporção média de sucesso igual a $0,525$, no segundo cenario $\boldsymbol{\beta} = (3; 3; 3)^\top$ e proporção média de sucesso de $0,953$ e, no terceiro cenário $\boldsymbol{\beta} = (-2; -2; -2)^\top$ e proporção média de sucesso de $0,119$. Nós calcularemos as estimativas, o erro padrão (E.P), o viés e também seu valor absoluto e, o erro quadrático médio (EQM).

Tabela 4.1: Resultado da simulação para $\boldsymbol{\beta} = (0, 1, 0, 1, 0, 1)^\top$

n	Parâmetro	Estimativa	EP	Viés	|Viés|	EQM
50	β_1	0,103	0,317	0,035	0,250	0,100
	β_2	0,115	0,341	0,015	0,268	0,116
	β_3	0,111	0,355	0,011	0,278	0,126
70	β_1	0,104	0,258	0,004	0,203	0,066
	β_2	0,108	0,274	0,008	0,216	0,075
	β_3	0,106	0,280	0,006	0,219	0,078
90	β_1	0,103	0,223	0,003	0,177	0,050
	β_2	0,104	0,245	0,004	0,193	0,060
	β_3	0,107	0,232	0,008	0,182	0,054
120	β_1	0,105	0,192	0,005	0,153	0,037
	β_2	0,106	0,205	0,006	0,162	0,042
	β_3	0,107	0,201	0,007	0,159	0,041
200	β_1	0,098	0,145	-0,002	0,116	0,021
	β_2	0,103	0,156	0,003	0,124	0,024
	β_3	0,103	0,151	0,003	0,120	0,023

Para a Tabela 4.1 nós podemos perceber que mesmo com uma amostra pequena, a estimativa dos parâmetros é bem próxima do valor real e, a medida que o tamanho da amostra aumenta, o erro padrão, o erro quadrático médio, o viés e seu módulo diminuem.

Para a Tabela 4.2 percebemos que para uma amostra de tamanho 50 e 70, as estimativas dos parâmetros não são boas, mas à medida que o tamanho da amostra

Tabela 4.2: Resultado da simulação para $\boldsymbol{\beta} = (3, 3, 3)^\top$

n	Parâmetro	Estimativa	EP	Viés	|Viés|	EQM
50	β_1	36,937	322,762	33,937	34,319	105316,488
	β_2	36,050	289,700	33,051	33,319	84992,503
	β_3	38,134	356,904	35,134	35,627	128601,952
70	β_1	9,078	81,580	6,078	6,449	6675,184
	β_2	9,264	90,894	6,264	6,653	8300,104
	β_3	9,051	79,513	6,051	6,480	6358,266
90	β_1	3,416	1,115	0,416	0,782	1,416
	β_2	3,436	1,175	0,436	0,838	1,570
	β_3	3,445	1,235	0,445	0,869	1,723
120	β_1	3,307	0,878	0,307	0,636	0,866
	β_2	3,301	0,932	0,301	0,671	0,960
	β_3	3,317	0,904	0,317	0,663	0,918
200	β_1	3,161	0,560	0,161	0,429	0,339
	β_2	3,162	0,575	0,162	0,447	0,357
	β_3	3,181	0,625	0,182	0,785	0,424

aumenta as estimativas convergem para o valor real, e o erro padrão, o viés e o EQM diminuem.

Tabela 4.3: Resultado da simulação para $\boldsymbol{\beta} = (-2, -2, -2)^\top$

n	Parâmetro	Estimativa	EP	Viés	|Viés|	EQM
50	β_1	-6,638	147,707	-4,638	4,939	21836,792
	β_2	-6,237	112,045	-4,237	4,589	12570,741
	β_3	-7,228	164,531	-5,228	5,613	27095,038
70	β_1	-2,641	14,058	-0,641	0,928	198,016
	β_2	-2,632	14,313	-0,632	0,931	205,237
	β_3	-2,681	15,736	-0,681	1,021	248,071
90	β_1	-2,192	0,614	-0,192	0,457	0,414
	β_2	-2,208	0,663	-0,208	0,497	0,483
	β_3	-2,215	0,704	-0,215	0,527	0,542
120	β_1	-2,144	0,499	-0,144	0,382	0,270
	β_2	-2,133	0,530	-0,133	0,411	0,298
	β_3	-2,115	0,527	-0,152	0,409	0,300
200	β_1	-2,069	0,334	-0,069	0,264	0,116
	β_2	-2,077	0,365	-0,077	0,289	0,139
	β_3	-2,087	0,383	-0,087	0,303	0,155

Nós podemos perceber que para $\boldsymbol{\beta} = (-2, -2, -2)$, assim como na Tabela 4.2 as estimativas para amostras pequenas diferem muito do valor real, e a medida que o tamanho amostral aumenta as estimativas convergem para o valor real.

O objetivo maior da realização destas simulações era avaliar os estimadores de

máxima verossimilhança e, vimos que quando as proporções de sucesso estão próximas as bordas do espaço parametrico (na prática, significa que estamos lidando com eventos raros), as estimativas para n pequeno são ruins e, a medida que n cresce as estimativas convergem para o valor real e, as estimativas do erro padrão, viés e EQM diminuem.

5 Aplicação

Para ilustrar as metodologias apresentadas nesta monografia, nós a aplicaremos em um conjunto de dados reais.

5.1 Titanic

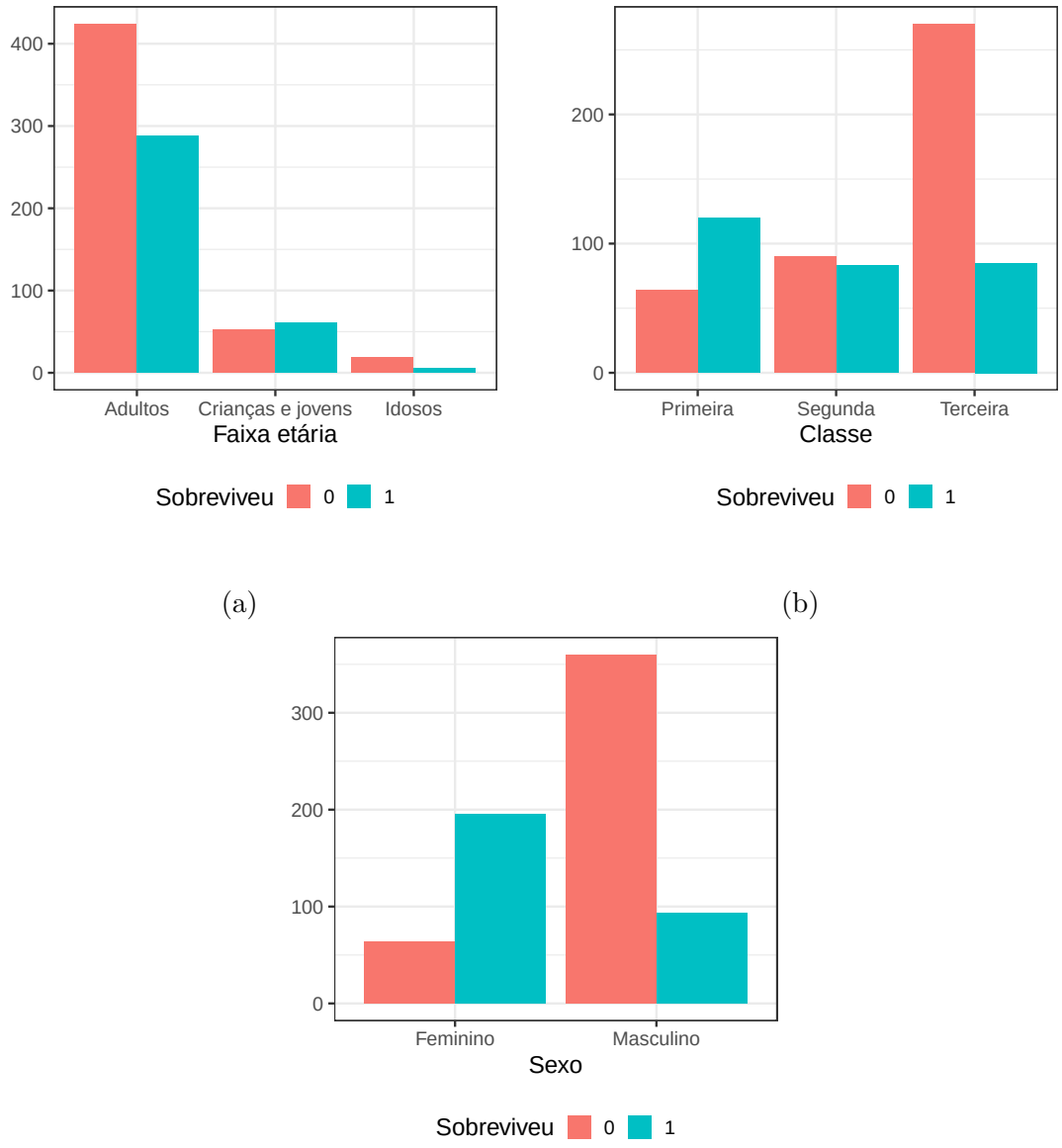
O Titanic foi um navio britânico do Século XX, criado para realizar viagens transatlânticas, mas fez apenas uma, em 1912, quando naufragou. O navio foi construído com a ideia de ser luxuoso e extremamente seguro, criando uma lenda de que seria inafundável.

Os dados do Titanic são uma referência para ilustrar problemas de classificação. Nós o usamos para ilustrar uma situação hipotética e, com isso é possível verificar a chance de sobreviver ou não, de um possível passageiro a bordo. Além disso também fizemos um modelo dos dados, como estudado no Capítulo 2. Os dados pertencem ao pacote de mesmo nome, Titanic (HENDRICKS, 2015), disponível no Software (R Core Team, 2020). A amostra original possui 1.309 observações, mas só temos acesso à 891 observações.

O banco de dados contém 11 variáveis, sendo a variável resposta sobreviveu (representado por 1), ou não sobreviveu (representado por 0) e, as demais: nome do passageiro, classe que estava alojado, sexo, idade, número de cônjuges e/ou irmãos à bordo, número de pais ou filhos a bordo, número da passagem, preço da passagem, número da cabine que o passageiro estava alojado e, a cidade em que embarcou. As variáveis nome, número da passagem e número da cabine são códigos e portanto não serão usadas nesta análise. Além disso, o banco possui 177 observações faltantes e, nós escolhemos eliminá-las.

Inicialmente, nós fizemos uma análise exploratória dos dados. A seguir os principais gráficos de nosso estudo.

No primeiro gráfico, inicialmente a variável idade era numérica, mas nós agrupamos a idade em três grupos, os menores ou iguais a 18 anos pertecem ao grupo “Criança e



(a) (b)

(c)

Figura 5.1: Idade, Classe e Sexo em relação à sobrevivência.

jovens" e, pessoas entre 19 e 59 estão no grupo "Adultos"e, os demais no grupo "Idosos". Observando a Figura 5.1 (a), percebemos que adultos representam a maior parcela dos dados, já o grupo de crianças e jovens, apesar de somarem uma parcela pequena, é o único grupo que a frequência relativa de sobreviventes é maior e, no grupo idosos, os não sobreviventes representam aproximadamente o dobro de sobreviventes, porém, é o grupo que possui o menor número de observações. Já na Figura 5.1 (b), o gráfico relaciona a sobrevivência do passageiro à classe que ele estava alojado e, nós podemos observar que a classe com mais sobreviventes foi a primeira e a classe mais populosa é a terceira, em que

Tabela 5.1: Ajuste final dos dados do Titanic

Coeficientes	**Estimativa**	**Erro Padrão**	**Valor Z**	**Valor P**
Intercepto	4,334	0,451	9,617	<0,001
Idade	-0,045	0,008	-5,442	<0,001
Sexo (masculino)	-2,628	0,215	-12,235	<0,001
Classe (segunda)	-1,414	0,285	-4,967	<0,001
Classe (terceira)	-2,653	0,286	-9,280	<0,001
Número de cônjuges/ ou irmãos à bordo	-0,380	0,122	-3,129	0,002

o grupo que não sobreviveu é aproximadamente a soma dos não sobreviventes da primeira e da segunda classe. E por fim, nós temos o terceiro gráfico, 5.1 (c), que relaciona o sexo à sobrevivência e, é possível observar que apesar de maior frequência do sexo masculino, a frequência relativa de sobreviventes é maior para o sexo feminino.

Após uma análise exploratória preliminar, nós identificamos as possíveis tendências entre a sobrevivência dos passageiros e as demais variáveis do banco de dados. Agora nós passaremos para a análise inferencial e veremos quais variáveis foram realmente significativas para o modelo.

O primeiro modelo foi construído com todas as variáveis, porém apenas a idade, o sexo, a classe do passageiro e, o número de cônjuges e/ou irmãos à bordo foram significativos a um nível de 5%. Após análises prévias, nós chegamos no seguinte modelo:

$$\begin{aligned}\text{logit}[\pi(x_i)] = 4,334 - 0,045 \times \text{Idade}_i - 2,628 \times \text{Sexo (masc)}_i - 1,414 \times \text{Classe (segunda)}_i \\ - 2,653 \times \text{Classe (terceira)}_i - 0,380 \times \text{Cônjuges/irmãos}_i\end{aligned}$$

Para esse modelo o pseudo-R^2 de Cox e Snell resultou em $0,37$. Já usando o teste de Hosmer e Lemeshow, visto na seção 2.4.2, nossa estatística Qui-Quadrado resultante foi 712, com 8 graus de liberdade e, ao nível de significância de 5% podemos dizer que as proporções esperadas não são as mesmas das proporções observadas, ou seja, há indícios que o modelo não se ajustou tão bem aos dados.

Para a análise dos coeficientes, estaremos observando a chance do sucesso, que é ter sobrevivido, do indivíduo X_{j+1}, em relação ao indivíduo X_j. Para variável idade,

temos que, a cada ano a mais, a chance de sobrevivência do passageiro diminui $0,96\%$. Já em relação ao sexo, temos que a chance de sobrevivência das mulheres é 93% maior que a dos homens. Em relação à classe, temos que a primeira tinha 76% e 93% maior chance de sobrevivência em relação a segunda e terceira classe, respectivamente. E a cada unidade a mais na contagem do número de cônjuges e/ou irmão à bordo, a probabilidade de sobrevivência cai 32%.

Após a análise feita de forma mais tradicional, faremos uso dos artifícios que auxiliam na melhora da predição. Os modelos feitos à partir daqui, possuem o banco de dados dividido em 2/3 para o treinamente e, 1/3 para validação. Além disso usaremos validação cruzada para estimação do λ nos modelos.

O primeiro ajuste nós fizemos apenas separando o banco de dados entre dados de treinamento e teste e estimando da forma tradicional. Após ajustes prévios, o modelo final é dado por:

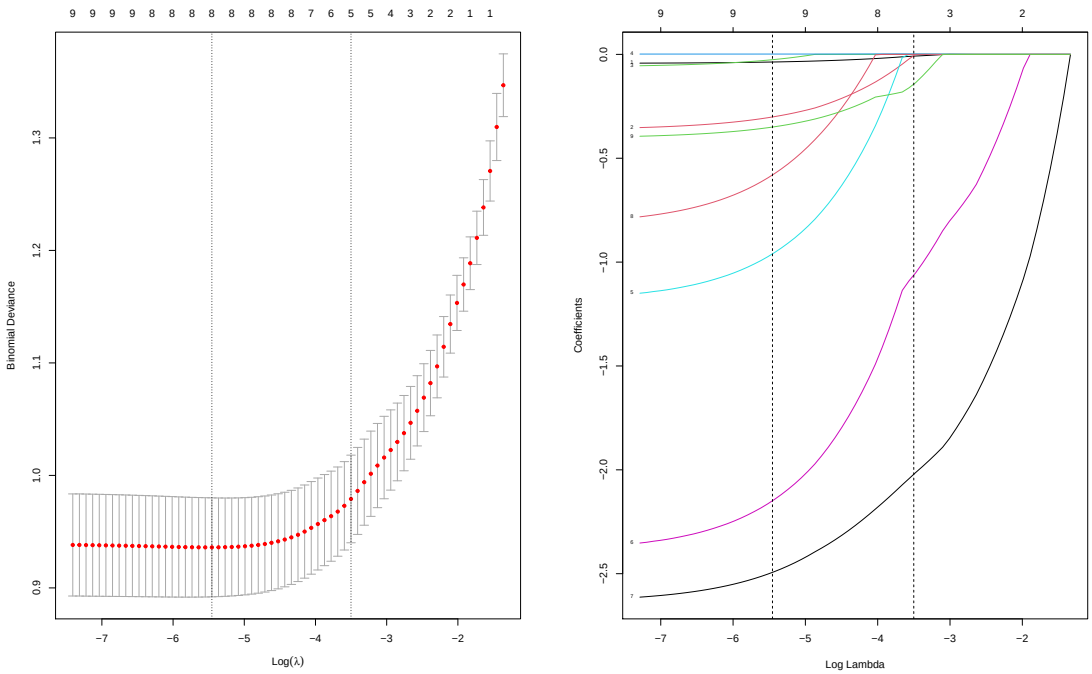
$$\begin{aligned}\text{logit}[\pi(x)] = 4,369 &- 0.053 \times \text{Idade}_i - 0.508 \times \text{Cônjuges/irmãos}_i \\ &- 1,280 \times \text{Classe (segunda)}_i - 2.551 \times \text{Classe (terceira)}_i - 2,592 \times \text{Sexo (masc)}_i\end{aligned}$$

Tabela 5.2: Matriz de Confusão do Modelo tradicional com banco de dados dividido

		Predito		Total
		$\hat{y}_i = 0$	$\hat{y}_i = 1$	
Observado	$y_i = 0$	119	26	145
	$y_i = 1$	22	76	98
Total		141	102	243

Pela matriz de confusão obtida a partir do modelo tradicional apenas com a divisão dos dados, podemos obter a acurácia de $80,25\%$, a precisão de $74,51\%$ e a sensibilidade de $77,55\%$.

O próximo ajuste nós faremos usando a penalização Lasso. Observando a Figura 5.2 (a), que mostra a curva de Validação Cruzada com $k = 10$, vemos que a primeira reta com $\lambda = 0.004$ produz o menor desvio. Já na Figura 5.2 (b), em que cada linha representa um coeficiente de acordo com a variação de λ, podemos observar que $\lambda = 0.004$ manteve todos coeficientes, já $\lambda = 0.045$ possui apenas 5 coeficientes.



(a) (b)

Figura 5.2: Curva de Validação Cruzada em (a) e Comportamento dos coeficientes em relação ao $\log(\lambda)$ em (b)

Nosso modelo final será dado por:

$$\text{logit}[\pi(x)] = 1,564 - 0.015 \times \text{Idade}_i - 0.129 \times \text{Cônjuges/irmãos}_i$$
$$0,003 \times \text{Preço}_i - 0,945 \times \text{Classe (terceira)}_i - 1,983 \times \text{Sexo (masc)}_i$$

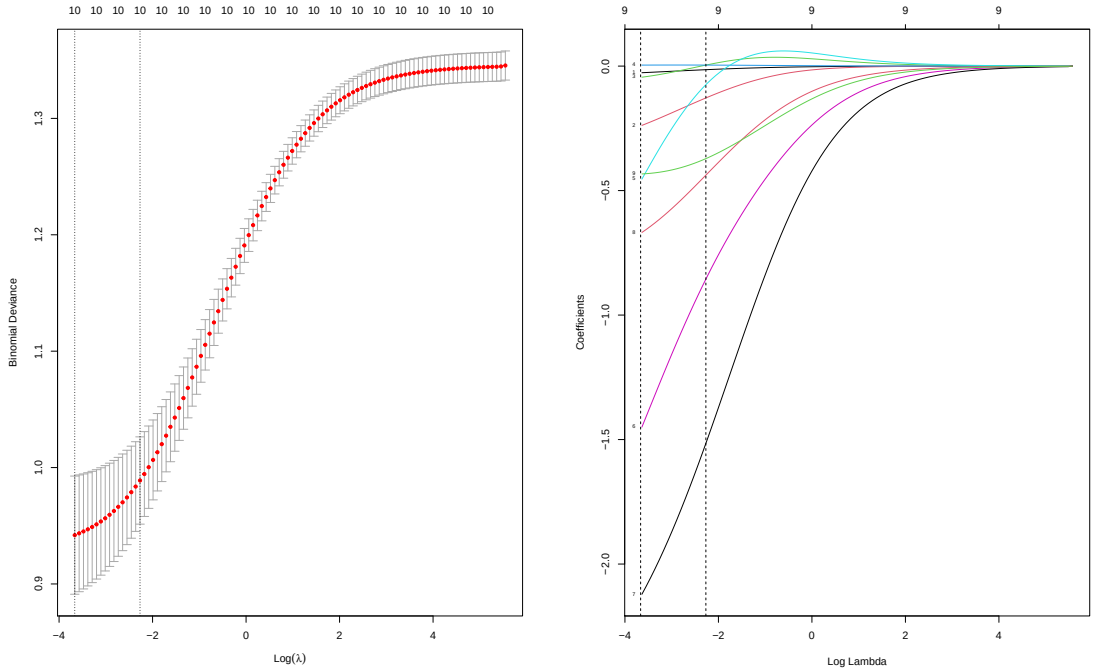
Tabela 5.3: Matriz de Confusão do Modelo Lasso

		Predito		Total
		$\hat{y}_i = 0$	$\hat{y}_i = 1$	
Observado	$y_i = 0$	122	33	155
	$y_i = 1$	19	69	88
Total		141	102	243

Pela matriz de confusão obtida a partir do modelo Lasso, a acurácia será de $78,6\%$, a precisão de $67,65\%$ e a sensibilidade de $78,41\%$.

Agora faremos um modelo por meio da penalização Ridge. Observando a Figura 5.3 (a), é possível ver que os valores de $\log(\lambda)$ no intervalo de -4 a -2 produzem os menores desvios. Já na Figura 5.3 (b), em que cada linha representa um coeficiente de

acordo com a variação de λ, nós podemos observar que para qualquer valor de λ nós teremos todos os coeficientes, diferente do Lasso.



(a) (b)

Figura 5.3: Curva de Validação Cruzada em (a) e Comportamento dos coeficientes em relação ao $\log(\lambda)$ em (b)

$$\begin{aligned}\text{logit}[\pi(x_i)] = 1,515 &- 0,018 \times \text{Idade}_i - 0,187 \times \text{Cônjuges/irmãos}_i + 0,033 \times \text{Pais/filhos}_i \\ &+ 0,004 \times \text{Preço}_i - 0,017 \times \text{Classe (segunda)}_i - 0,775 \times \text{Classe (terceira)}_i \\ &- 1,487 \times \text{Sexo (masc)}_i - 0,329 \times \text{Embarque (Q)}_i - 0,286 \times \text{Embarque (S)}_i\end{aligned}$$

Tabela 5.4: Matriz de Confusão do Modelo Ridge

		Predito		Total
		$\hat{y}_i = 0$	$\hat{y}_i = 1$	
Observado	$y_i = 0$	117	30	147
	$y_i = 1$	24	72	96
Total		141	102	243

Pela matriz de confusão obtida a partir do modelo Ridge, as métricas de acurácia precisão e sensibilidade serão, respectivamente $77,78\%$, $75,00\%$ e, $79,59\%$.

O modelo que obteve melhor resultado nas métricas acurácia e precisão foi o modelo tradicional com o banco de dados particionado, como podemos visualizar na Tabela abaixo, representado por Logístico*. Já para a métrica sensibilidade, o modelo Lasso apresentou melhor resultado.

Tabela 5.5: Métricas dos modelos

Métricas	Logístico*	Lasso	Ridge
Acurácia	0,802	0,786	0,779
Precisão	0,745	0,676	0,706
Sensibilidade	0,775	0,784	0,750

6 Considerações Finais

Neste trabalho nós revisamos os conceitos de Regressão Logística: o modelo, a inferência, o diagnóstico, a partir da ótica inferencial. Porém, nos últimos anos esse modelo vem sendo usado no aprendizado supervisionado cujo objetivo é a previsão de novas observações. Para essa ótica mais preditivista, além de validação cruzada, podemos aplicar penalizações, como Lasso e Ridge, com uma intenção diferente da proposta: a de acrescentar viés. Depois da revisão de conceitos, nós fizemos um estudo de simulação para visualizar o comportamento assintótico dos estimadores de máxima verossimilhança e vimos que, quando a proporção de sucesso está nas extremidades, para amostras pequenas o estimador não produz bons resultados. Nós também aplicamos algumas técnicas de melhora de previsão em um banco de dados muito comum nos estudos de aprendizado supervisionado e, o modelo com o banco de dados particionado apresentou maior acurácia e precisão e, o modelo Lasso apresentou maior sensibilidade, apesar disso o desempenho dos três modelos resultantes foi similar e fica a cargo do pesquisador escolher o modelo com a métrica que deseja priorizar.

Como Diniz e Magalhães (2020) mencionam métodos de correção (refinamentos assintóticos) para tamanhos amostrais pequenos, como sugestão de trabalhos futuros, está a expansão das simulações do Capítulo 4, levando em consideração os métodos de refinamento assintótico, tanto para os estimadores, como para as estatísticas do teste. Uma segunda sugestão é expandir a simulação do Capítulo 4, levando em consideração o foco preditivista, como Calster et al. (2020), e também as metodologias de refinamento.

Bibliografia

BUSE, A. The likelihood ratio, wald and lagrange multiplier tests: an expository note. *The American Statistician*, v. 36, n. 3, p. 153–157, 1982.

CALSTER, B. V. et al. Regression shrinkage methods for clinical prediction models do not guarantee improved performance: Simulation study. *Statistical Methods in Medical Research*, v. 29, n. 11, p. 3166–3178, nov. 2020. ISSN 0962-2802, 1477-0334. Disponível em: ⟨http://journals.sagepub.com/doi/10.1177/0962280220921415⟩.

COX, D. R.; SNEEL, E. J. *The analysis of binary data*. [S.l.]: Chapman and Hall, 1989.

DINIZ, M. A.; MAGALHãES, T. M. Logistic Regression and Related Methods. In: PIANTADOSI, S.; MEINERT, C. L. (Ed.). *Principles and Practice of Clinical Trials*. Cham: Springer International Publishing, 2020. p. 1–23. ISBN 9783319526775. Disponível em: ⟨https://doi.org/10.1007/978-3-319-52677-5_122-1⟩.

FIRTH, D. Bias Reduction of Maximum Likelihood Estimates. *Biometrika*, v. 80, n. 1, p. 27–38, 1993. ISSN 0006-3444. Disponível em: ⟨https://www.jstor.org/stable/2336755⟩.

FROST, J. *Regression Analysis: An Intuitive Guide for Using and Interpreting Linear Models*. [S.l.]: Statistics By Jim Publishing, 2020.

GALTON, F. Natural inheritance. *London: Macmillan*, 1889.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. *The Elements of Statistical Learning*. New York, NY: Springer New York, 2009. (Springer Series in Statistics). ISBN 9780387848570 9780387848587. Disponível em: ⟨http://link.springer.com/10.1007/978-0-387-84858-7⟩.

HENDRICKS, P. *titanic: Titanic Passenger Survival Data Set*. [S.l.], 2015. R package version 0.1.0. Disponível em: ⟨https://CRAN.R-project.org/package=titanic⟩.

HOERL, A. E.; KENNARD, R. W. Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics*, v. 12, n. 1, p. 55–67, fev. 1970. ISSN 0040-1706, 1537-2723. Disponível em: ⟨http://www.tandfonline.com/doi/abs/10.1080/00401706.1970.10488634⟩.

HOUWELINGEN, J. C. V.; CESSIE, S. L. Predictive value of statistical models. *Statistics in Medicine*, v. 9, n. 11, p. 1303–1325, nov. 1990. ISSN 02776715, 10970258. Disponível em: ⟨http://doi.wiley.com/10.1002/sim.4780091109⟩.

JAMES, G. et al. *An Introduction to Statistical Learning: with Applications in R*. Springer, 2013. Disponível em: ⟨https://faculty.marshall.usc.edu/gareth-james/ISL/⟩.

MONTORIL, M. H.; SOUZA, E. A. Estatística gradiente: propriedades e aplicações. *Revista Brasileira de Biometria*, v. 31, n. 1, p. 43–60, 2013.

NELDER, J. A.; WEDDERBURN, R. W. M. Generalized Linear Models. *Journal of the Royal Statistical Society. Series A (General)*, v. 135, n. 3, p. 370–384, 1972. ISSN 0035-9238. Disponível em: ⟨https://www.jstor.org/stable/2344614⟩.

R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria, 2020. Disponível em: ⟨https://www.R-project.org/⟩.

RAO, C. R. Large sample tests of statistical hypotheses concerning several parameters with applications to problems of estimation. *Proceedings of the Cambridge Philosophical Society*, v. 44, n. 1, p. 50–57, 1948.

SEN, P. K.; SINGER, J. M.; LIMA, A. C. P. *From Finite Sample to Asymptotic Methods in Statistics*. New York: Cambridge University Press, 2010.

TERRELL, G. R. The gradient statistic. *Computing Science and Statistics*, v. 34, p. 206–215, 2002.

TIBSHIRANI, R. Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, v. 58, n. 1, p. 267–288, 1996. ISSN 0035-9246. Disponível em: ⟨https://www.jstor.org/stable/2346178⟩.

WALD, A. Test of statistical hypotheses concerning several parameter when the number of observations is large. *Transactions of the American Mathematical Society*, v. 54, n. 3, p. 426–482, 1943.

WILKS, S. S. The large-sample distribution of likelihood ratio for testing composite hypotheses. *The Annals of Mathematical Statistics*, v. 9, n. 1, p. 60–62, 1938.

A Apêndice

$$
\begin{aligned}
\ell(\boldsymbol{\beta}) = \ell(\boldsymbol{\beta}|\boldsymbol{y}, \boldsymbol{X}) &= \sum_{i=1}^{n} \{y_i \log \pi_i + (1 - y_i) \log(1 - \pi_i)\} \\
&= \sum_{i=1}^{n} y_i \log(\pi_i) + \sum_{i=1}^{n} (1 - y_i) \log(1 - \pi_i) \\
&= \sum_{i=1}^{n} y_i \log(\pi_i) + \sum_{i=1}^{n} \log(1 - \pi_i) - \sum_{i=1}^{n} y_i \log(1 - \pi_i) \\
&= \sum_{i=1}^{n} y_i [\log(\pi_i) - \log(1 - \pi_i)] + \sum_{i=1}^{n} \log(1 - \pi_i) \\
&= \sum_{i=1}^{n} y_i \log \frac{\pi_i}{(1 - \pi_i)} + \sum_{i=1}^{n} \log(1 - \pi_i) \\
&= \sum_{i=1}^{n} y_i (\beta_1 + \beta_2 x_2 + ... + \beta_p x_p) - \sum_{i=1}^{n} \log(1 + exp(\beta_1 + \beta_2 x_2 + ... + \beta_p x_p))
\end{aligned}
$$