**Avaliação de propostas de coeficientes
de determinação do tipo R^2 em modelos
de regressão logística com resposta nominal**

Ângelo de Souza Chiode

DISSERTAÇÃO APRESENTADA
AO
INSTITUTO DE MATEMÁTICA E ESTATÍSTICA
DA
UNIVERSIDADE DE SÃO PAULO
PARA
OBTENÇÃO DO TÍTULO
DE
MESTRE EM CIÊNCIAS

Programa: Estatística

Orientadora: Prof^a. Dr^a. Silvia Nagib Elian

São Paulo, Novembro de 2020

Avaliação de propostas de coeficientes de determinação do tipo R^2 em modelos de regressão logística com resposta nominal

Esta dissertação contém as correções e alterações sugeridas pela Comissão Julgadora durante a defesa da versão original do trabalho, realizada em 04/03/2021. Uma cópia da versão original está disponível no Instituto de Matemática e Estatística da Universidade de São Paulo.

Comissão Julgadora:

- Prof^a. Dr^a. Silvia Nagib Elian (orientadora) - IME-USP
- Prof. Dr. Bruno Ramos dos Santos - UFES
- Prof. Dr. Rinaldo Artes - INSPER

Agradecimentos

Agradeço à minha orientadora, Prof^a Dra^a Sílvia Nagib Elian, por me conduzir e tranquilizar ao longo desta jornada. Sua imensa disponibilidade e compreensão foram essenciais para o sucesso do meu trabalho, e todas as suas contribuições técnicas foram de extrema relevância.

Agradeço, também, ao Instituto de Matemática e Estatística da Universidade de São Paulo, pela excelente estrutura curricular oferecida ao longo da minha graduação e mestrado. Em especial, agradeço a todos os professores que contribuíram com a minha formação intelectual durante estas etapas.

Agradeço às minhas gestoras na empresa Marketdata, Karin e Karina, que me incentivaram a cursar o mestrado e forneceram todos os meios possíveis para que eu pudesse honrar, com excelência, todas as minhas responsabilidades. Seu apoio foi fundamental para o meu desenvolvimento pessoal, profissional e acadêmico.

Por fim, agradeço à minha família, Ana Lúcia e Rafael, por todo o seu amor, paciência e apoio incondicionais. Vocês são a principal razão das minhas lutas e das minhas conquistas.

Resumo

CHIODE A. S. **Avaliação de propostas de coeficientes de determinação do tipo R^2 em modelos de regressão logística com resposta nominal**. 2020. 124 f. Dissertação (Mestrado) - Instituto de Matemática e Estatística, Universidade de São Paulo, São Paulo, 2020.

Ao contrário de modelos de regressão linear múltipla, em que o coeficiente de determinação R^2 representa uma forma intuitiva de avaliar a proporção de variabilidade da resposta que é devidamente explicada por meio das variáveis independentes, não existe uma forma bem estabelecida de se mensurar a variabilidade explicada em modelos de regressão logística. Diversas proposições já foram realizadas em trabalhos anteriores, especialmente para casos em que existe uma escala contínua latente associada à variável resposta. Entretanto, não há consenso sobre a melhor alternativa para respostas de natureza nominal. Este trabalho teve como objetivo revisar 12 propostas de coeficientes de determinação do tipo R^2 que já foram discutidas em trabalhos anteriores, a fim de avaliar a sua adequabilidade para modelos de regressão logística com resposta nominal. Foi conduzido um estudo de simulação, que apontou que os dois coeficientes apresentados por [McFadden \(1974\)](#) - um para a regressão simples, e outro corrigido, para a regressão múltipla - são os mais apropriados entre todas as alternativas analisadas, levando em conta a sua interpretação intuitiva; variabilidade no intervalo razoável de valores de 0 a 1; e maior conformidade com o índice de sobreposição ([Bura e Gastwirth, 2001](#)) e com medidas de acurácia preditiva, tais como o KS ([Hand, 2012](#)) e a área abaixo da curva ROC - ou, equivalentemente, o índice de Gini.

Palavras-chave: regressão logística; coeficiente de determinação; R^2 ; variabilidade explicada; acurácia preditiva; resposta nominal.

Abstract

CHIODE, A. S. **Evaluation of proposals of coefficients of determination (R^2 -analogs) in logistic regression models with a nominal response.** 2020. 124 s. Dissertation (Master degree) - Instituto de Matemática e Estatística, Universidade de São Paulo, São Paulo, 2020.

Unlike multiple linear regression models, in which the coefficient of determination R^2 represents an intuitive way of evaluating the proportion of variation in the response that is adequately explained by using the independent variables, there is no well-established way of measuring the explained variation in logistic regression models. Several proposals have already been made in previous works, especially for cases in which there is a continuous latent scale associated with the response variable. However, there is no consensus on the best alternative for nominal responses. This study aimed to review 12 proposals of coefficients of determination (R^2 -analogs) that have already been discussed in previous studies, in order to assess their suitability for logistic regression models with a nominal response. A simulation study was conducted, which pointed out that the two coefficients presented by [McFadden \(1974\)](#) - one for simple regression, and the other corrected for multiple regression - are the most appropriate among all the alternatives analyzed, taking into account their intuitive interpretation; variability in a reasonable range of values from 0 to 1; and greater consistency with the overlap index ([Bura e Gastwirth, 2001](#)) and with predictive accuracy measures, such as KS ([Hand, 2012](#)) and the area under the ROC curve - or, equivalently, the Gini index.

Keywords: logistic regression; coefficient of determination; R^2 ; explained variation; predictive accuracy; nominal response.

Sumário

Lista de Abreviaturas	vii
Lista de Símbolos	viii
1 Introdução	1
2 Regressão Logística e Coeficiente de Determinação	4
2.1 Regressão Logística Múltipla	4
2.2 Regressão Logística Múltipla Multinomial	6
2.3 Coeficiente de Determinação Tradicional (R^2)	8
2.4 Adaptação do R^2 para a Regressão Logística (R^2_{MQ})	9
2.5 Considerações Finais do Capítulo	11
3 Acurácia Preditiva	12
3.1 O Problema de Classificação	13
3.2 Índices de Acurácia Preditiva	18
3.2.1 Acurácia Geral e por Categoria	18
3.2.2 Área Abaixo da Curva ROC (AUC) e Índice de Gini (G)	19
3.2.3 Índices de Youden (J) e Kolmogorov-Smirnov (KS)	20
3.2.4 Outros índices	22
3.3 Considerações Finais do Capítulo	24
4 Propostas de Coeficientes de Determinação do Tipo R^2	26

4.1	Medidas Baseadas na Verossimilhança	26
4.1.1	Função de Verossimilhança	27
4.1.2	Razão de Logaritmos das Verossimilhanças (R_{MF}^2)	30
4.1.3	Coefficiente de Contingência Simples (R_{AN}^2) e Corrigido (R_{HM}^2)	33
4.1.4	Incremento Quadrático Médio por Observação (R_{MA}^2) e corrigido (R_N^2)	34
4.2	Medidas Não Baseadas na Verossimilhança	36
4.2.1	Correlações de Pearson (R_P^2), Spearman (R_S^2) e Kendall (R_K^2)	36
4.2.2	Coefficientes de Haberman: Entropia (R_H^2) e Concentração (R_G^2)	38
4.2.3	Coefficiente de Ganho Total Padronizado (R_{BG}^2)	40
4.2.4	Coefficiente de Discriminação de Tjur (R_T^2)	42
4.3	Estratégia para Avaliação dos Coeficientes	45
4.4	Considerações Finais do Capítulo	47
5	Simulação	49
5.1	Esquema de Simulação	49
5.1.1	Fatores de variação	50
5.1.2	Etapas da simulação	51
5.2	Resultados	56
5.2.1	Análises iniciais	56
5.2.2	Panorama geral dos coeficientes R^2	58
5.2.3	Análises por cenário	62
5.2.4	Relação com a proporção de sobreposição final (p'_S)	63
5.2.5	Relações com os índices de acurácia preditiva	65
5.2.6	Acréscimo de variável explicativa	71
5.3	Considerações Finais do Capítulo	73
6	Conclusões	74

A Tabelas e Figuras	77
B Códigos da Simulação	113
Referências Bibliográficas	120

Lista de Abreviaturas

AUC	Área baixo da curva ROC (<i>Area under ROC curve</i>)
KS	Índice de Kolmogorov-Smirnoff
SQE	Soma de quadrados dos resíduos
SQR	Soma de quadrados da regressão
SQT	Soma de quadrados total
ROC	Característica de operação do receptor (<i>Receiver operating characteristic</i>)

Lista de Símbolos

χ^2 Estatística qui-quadrado para tabelas de contingência

ϕ_c Coeficiente ϕ de Pearson (ou Correlação de Matthews)

Capítulo 1

Introdução

A regressão logística é uma das principais técnicas de modelagem da probabilidade de ocorrência associada a um evento de natureza categórica, por meio de um conjunto de variáveis explicativas. Conforme estudo histórico realizado por [Cramer \(2002\)](#), a técnica se consolidou no final da década de 1960, especialmente após as contribuições de David R. Cox, e continuou a ser desenvolvida e popularizada, gradativamente, por meio de inúmeros outros trabalhos de pesquisa publicados nas três décadas seguintes.

As aplicações da regressão logística estendem-se a diversas áreas do conhecimento, conforme salientado por [Hosmer *et al.* \(2013\)](#), com ênfase para as ciências biológicas (ex.: medicina, epidemiologia, botânica), ciências humanas e sociais (ex.: psicologia, criminologia, linguística) e questões de negócios (ex.: administração, marketing, finanças). Nas últimas duas décadas, com o advento da inteligência artificial e dos modelos de aprendizado de máquina, a regressão logística também passou a ser empregada em problemas relacionados ao reconhecimento de padrões em textos e imagens; identificação de mensagens eletrônicas indesejadas (*spams*); detecção automática de fraudes em transações financeiras; entre tantos outros exemplos. A simplicidade algébrica e interpretabilidade do modelo de regressão logística contribuíram para que tal metodologia mantivesse a sua relevância de forma perene, mesmo em meio ao desenvolvimento de algoritmos de *machine learning* mais sofisticados.

Uma das necessidades frequentes no âmbito dos modelos de regressão consiste em mensurar o grau de variabilidade explicada da variável resposta, por meio das variáveis explicativas. Para modelos de regressão linear múltipla, sabe-se que o coeficiente de determinação R^2 é uma medida comumente empregada para essa finalidade. Tal coeficiente surge de forma natural, mediante partição da soma de quadrados total, resultante da estimação por mínimos quadrados ([Neter *et al.*, 2004](#)). A sua utilização encontra-se bem estabelecida na teoria associada à regressão linear, não

cabendo demasiada discussão a respeito de índices alternativos (Mittlböck e Schemper, 1996). Já no cenário da regressão logística, assim como em demais modelos não lineares em que a estimação de parâmetros é realizada pelo método de máxima verossimilhança, a utilização do coeficiente R^2 usual não se faz necessariamente adequada (Menard, 2000). Em tais contextos, R^2 não representaria uma medida razoável da proporção de variabilidade explicada, pois só poderia ser empregado sob determinadas condições relacionadas às probabilidades de sucesso e não atingiria o valor máximo de 1, conforme apontado por Cox e Wermuth (1992) e Ryan (2008), o que motivaria a busca por outros coeficientes.

Diversas proposições de coeficientes substitutos ao R^2 da regressão linear foram apresentadas em trabalhos anteriores a este, mas não houve consenso sobre qual delas seria a mais apropriada para a regressão logística. Em particular, alguns resultados interessantes foram obtidos para o caso específico de modelos com variáveis resposta categóricas ordinais, quando se pode inferir a existência de uma escala contínua latente, que veio a ser discretizada por conta dos interesses da pesquisa ou de limitações na coleta dos dados. Nesse sentido, merecem destaque as simulações realizadas por Hagle e Mitchell (1992), Veall e Zimmermann (1994) e Mittlböck e Schemper (1996). Tais autores buscaram por coeficientes alternativos que mais se aproximassem algebricamente do R^2 tradicional, calculado para a resposta latente em um modelo de regressão linear. Adicionalmente, no caso de variáveis resposta em escala ordinal, outras técnicas de modelagem podem se fazer mais apropriadas do que a regressão logística tradicional (Okura, 2008). Já para situações em que a variável resposta da regressão logística é de natureza estritamente nominal, ou seja, sem nenhum tipo de ordenação implícita entre as suas categorias, os resultados apresentados em estudos anteriores são pouco conclusivos e não convergentes entre si.

Convém ressaltar que o julgamento a respeito do grau de variabilidade explicada consiste em uma tarefa distinta da análise de diagnóstico. Eventualmente, todas as suposições teóricas acerca do modelo podem estar devidamente satisfeitas, mas a força de explicabilidade da variável resposta por meio das variáveis independentes pode ser pequena. Um maior aprofundamento no que diz respeito ao diagnóstico de modelos de regressão logística pode ser encontrado nas dissertações de Mestrado de Farhat (2003) e Nakamura (2013), também provenientes do Instituto de Matemática e Estatística da Universidade de São Paulo.

Tendo em vista a conjuntura atual, os objetivos deste trabalho consistem em:

- Apresentar uma compilação abrangente dos principais coeficientes de determinação do tipo R^2 que foram propostos na literatura de regressão logística, expondo os aspectos favoráveis e desfavoráveis descritos a respeito de cada um deles.
- Avaliar a adequabilidade dos coeficientes apresentados, de forma quantitativa, por meio de

um estudo de simulação. Consequentemente, propor a melhor alternativa de coeficiente do tipo R^2 , para mensuração do grau de variabilidade explicada no modelo de regressão logística com resposta de natureza nominal.

Os conceitos gerais a respeito do modelo de regressão logística múltipla serão introduzidos no Capítulo 2, bem como discutir-se-á a aplicabilidade do R^2 tradicional, proveniente da estimação por mínimos quadrados. Já no Capítulo 3, serão elencados os principais índices de acurácia preditiva, já bem estabelecidos no que diz respeito à avaliação de modelos de regressão logística, em contextos em que existe a necessidade de classificar/predizer a resposta associada a novas observações. Em seguida, 12 alternativas de coeficientes de determinação do tipo R^2 serão apresentadas, no Capítulo 4, juntamente com os principais comentários acerca de cada uma delas em estudos anteriores. Tais proposições serão avaliadas e comparadas com outros índices de referência, por meio de um estudo de simulação, cujos resultados serão discutidos no Capítulo 5. Por fim, apresentar-se-á as conclusões do trabalho no Capítulo 6. O Apêndice A contém tabelas e figuras complementares, e o Apêndice B apresenta os códigos em linguagem R utilizados para condução da simulação.

Capítulo 2

Regressão Logística e Coeficiente de Determinação

Inicialmente, será realizada uma breve revisão a respeito do modelo de regressão logística múltipla e do conceito de coeficiente de determinação do tipo R^2 , introduzindo a notação que será empregada no decorrer deste trabalho. Ainda, neste capítulo, discute-se a respeito da aplicabilidade do R^2 usual, enquanto medida de explicação da variabilidade da resposta no modelo de regressão logística.

2.1 Regressão Logística Múltipla

Conforme apontado por [Hosmer *et al.* \(2013\)](#), os modelos de regressão logística possuem como objetivo explicar a probabilidade de ocorrência de determinado evento ou característica de interesse (*variável resposta*) de natureza categórica, a partir de um conjunto de outras características, que demonstram alguma força de explicação (*variáveis explicativas*). Busca-se realizar tal tarefa com qualidade, ou seja, estabelecendo uma expressão algébrica e suposições estatísticas que tenham a maior aderência possível aos dados observados. Ao mesmo tempo, deve-se zelar pela parcimônia, utilizando a menor quantidade possível de variáveis explicativas.

Em sua forma mais tradicional, a regressão logística é aplicada a variáveis resposta de natureza dicotômica: sim/não; sucesso/fracasso etc. Neste trabalho, não estão sendo consideradas as situações em que a variável resposta de interesse original possui natureza contínua; ou seja, sua natureza é, de fato, nominal. Alguns exemplos de situações práticas em que se pode ter o interesse em modelar uma variável dicotômica nominal, em contextos bastante distintos, são apresentados a seguir.

- Identificação dos hábitos de vida associados à ocorrência/não ocorrência de determinada variação de câncer de estômago.
- Identificação de características de consumidores que cancelam/não cancelam seu contrato de serviços de telefonia, prestados por uma empresa do ramo.
- Identificação de fatores que levam clientes a serem adimplentes/inadimplentes a um empréstimo de dinheiro realizado por determinado banco.

Em muitas ocasiões, os modelos de regressão logística são empregados com caráter preditivo, ou seja, como ferramenta hábil para prever se observações futuras serão designadas como “sucessos” ou “fracassos” quanto ao evento de interesse. Em outros casos, os modelos são utilizados apenas de forma retrospectiva, a fim de identificar e quantificar a influência de diferentes variáveis explicativas sobre os sucessos e fracassos já observados.

A formulação do modelo é dada por:

$$Y_i \sim \text{Bernoulli}(\pi_1(\mathbf{x}_i)),$$

$$\log \left(\frac{\pi_1(\mathbf{x}_i)}{1 - \pi_1(\mathbf{x}_i)} \right) = \beta_0 + \beta_1 x_{i,1} + \dots + \beta_{p-1} x_{i,p-1} = \mathbf{x}_i^T \boldsymbol{\beta}, \quad (2.1)$$

em que:

- n é a quantidade de observações;
- $i = 1, \dots, n$ é o índice associado à observação;
- Y_i é uma variável resposta dicotômica, que assume valores 1 (“sucesso”) ou 0 (“fracasso”), para a i -ésima observação;
- $\pi_1(\mathbf{x}_i)$ é a probabilidade condicional de ocorrência de sucesso, dados os valores $\mathbf{x}_i$ das $p - 1$ variáveis explicativas associadas à i -ésima observação;
- $\boldsymbol{\beta} = [\beta_0, \dots, \beta_{p-1}]$ é um vetor de p parâmetros desconhecidos;
- $\mathbf{x}_i^T = [1, x_{i,1}, \dots, x_{i,p-1}]$ é um vetor de valores observados e fixados, correspondentes a $p - 1$ variáveis explicativas, para a i -ésima observação.

Como consequência direta de (2.1), tem-se a seguinte expressão para a probabilidade condicional de sucesso, como função das variáveis explicativas:

$$\pi_1(\mathbf{x}_i) = \frac{\exp(\mathbf{x}_i^T \boldsymbol{\beta})}{1 + \exp(\mathbf{x}_i^T \boldsymbol{\beta})} = \frac{1}{1 + \exp(-\mathbf{x}_i^T \boldsymbol{\beta})}. \quad (2.2)$$

Consequentemente, a probabilidade condicional de fracasso é dada por:

$$\pi_0(\mathbf{x}_i) = 1 - \pi_1(\mathbf{x}_i) = \frac{1}{1 + \exp(\mathbf{x}_i^T \boldsymbol{\beta})}.$$

Ao longo deste trabalho, denotar-se-á as proporções observadas de fracasso e sucesso, em uma determinada amostra, por $\bar{\pi}_0$ e $\bar{\pi}_1$, respectivamente.

2.2 Regressão Logística Múltipla Multinomial

Uma extensão natural do modelo de regressão logística com resposta dicotômica pode ser realizada para respostas com três ou mais categorias, conforme apresentada nesta seção. Novamente, nota-se que serão levadas em conta apenas variáveis resposta com natureza puramente nominal, e não ordinal. Alguns exemplos de situações práticas em que se pode ter interesse em empregar esta técnica são:

- Identificação das características demográficas do eleitorado que estão associadas à intenção de voto em cinco candidatos ao governo de um estado.

Respostas possíveis: *candidato 1; candidato 2; ...; candidato 5.*

- Identificação de fatores sanguíneos associados à presença concomitante de duas variações de determinada bactéria no organismo.

Respostas possíveis: *ambas as variações presentes; apenas variação A presente; apenas variação B presente; ambas as variações ausentes.*

- Identificação de aspectos transacionais que podem interferir na resposta dos clientes de uma loja à seguinte pergunta, proveniente de pesquisa de satisfação: “Qual o principal aspecto que pode ser melhorado em sua experiência de compra?”.

Respostas possíveis: *atendimento; preço; variedade de produtos.*

Denotando por k a quantidade de categorias da variável resposta, e assumindo, sem perda de generalidade, a k -ésima categoria como sendo de referência, a formulação do modelo é dada por um conjunto de $k - 1$ equações:

$$Y_i \sim Multinomial(\pi_1(\mathbf{x}_i), \dots, \pi_k(\mathbf{x}_i)),$$

$$\log \left(\frac{\pi_j(\mathbf{x}_i)}{\pi_k(\mathbf{x}_i)} \right) = \beta_{j,0} + \beta_{j,1} x_{i,1} + \dots + \beta_{j,p-1} x_{i,p-1} = \mathbf{x}_i^T \boldsymbol{\beta}_j, \quad 1 \leq j \leq k-1, \quad (2.3)$$

em que:

- n é a quantidade de observações.
- $i = 1, \dots, n$ é o índice associado à observação.
- Y_i é uma variável resposta categórica, assumindo valores inteiros de 1 a k , para a i -ésima observação;
- $j = 1, \dots, k-1$ é o índice associado à categoria da variável resposta.
- $\pi_j(\mathbf{x}_i)$ é a probabilidade condicional de ocorrência da categoria j , dados os valores $\mathbf{x}_i$ das $p-1$ variáveis explicativas associadas à i -ésima observação;
- $\boldsymbol{\beta}_j = [\beta_{j,0}, \dots, \beta_{j,p-1}]$ é o vetor de p parâmetros desconhecidos, associados ao logito envolvendo a j -ésima categoria;
- $\mathbf{x}_i^T = [1, x_{i,1}, \dots, x_{i,p-1}]$ é um vetor de valores observados e fixados, correspondentes a $p-1$ variáveis explicativas, para a i -ésima observação.

Os logitos associados a quaisquer pares a e b de categorias da variável resposta ($1 \leq a, b \leq k$) são calculados a partir dos $k-1$ valores obtidos pelo sistema de equações (2.3):

$$\log \left(\frac{\pi_a(\mathbf{x}_i)}{\pi_b(\mathbf{x}_i)} \right) = \log \left(\frac{\pi_a(\mathbf{x}_i)/\pi_k(\mathbf{x}_i)}{\pi_b(\mathbf{x}_i)/\pi_k(\mathbf{x}_i)} \right) = \log \left(\frac{\pi_a(\mathbf{x}_i)}{\pi_k(\mathbf{x}_i)} \right) - \log \left(\frac{\pi_b(\mathbf{x}_i)}{\pi_k(\mathbf{x}_i)} \right).$$

Alternativamente, pode-se escrever o modelo a partir das probabilidades condicionais associadas a cada categoria da variável resposta, como função das variáveis explicativas:

$$\pi_j(\mathbf{x}_i) = \frac{\exp(\mathbf{x}_i^T \boldsymbol{\beta}_j)}{1 + \sum_{m=1}^{k-1} \exp(\mathbf{x}_i^T \boldsymbol{\beta}_m)}, \quad 1 \leq j \leq k-1, \quad (2.4)$$

e, para a categoria de referência,

$$\pi_k(\mathbf{x}_i) = 1 - \sum_{j=1}^{k-1} \pi_j(\mathbf{x}_i) = \frac{1}{1 + \sum_{m=1}^{k-1} \exp(\mathbf{x}_i^T \boldsymbol{\beta}_m)}. \quad (2.5)$$

Ao longo deste trabalho, as proporções observadas em cada categoria de resposta, para uma determinada amostra, serão denotadas como $\bar{\pi}_j$, $1 \leq j \leq k$.

2.3 Coeficiente de Determinação Tradicional (R^2)

Cabe lembrar o conceito do coeficiente de determinação R^2 , em seu contexto original do modelo de regressão linear múltipla, cuja expressão é dada por:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \dots + \beta_{p-1} X_{i,p-1} + \varepsilon_i, \quad (2.6)$$

em que:

- Y_i é o valor da variável resposta, de natureza contínua, para a i -ésima observação;
- $\beta_0, \dots, \beta_{p-1}$ são parâmetros desconhecidos;
- $X_{i1}, \dots, X_{i,p-1}$ são valores observados e fixados, correspondentes a $p-1$ variáveis explicativas, para a i -ésima observação;
- ε_i é um erro aleatório, com média zero e variância constante, para a i -ésima observação, sendo que ε_i e ε_j são independentes $\forall i \neq j$;
- $i = 1, \dots, n$ é o índice associado à observação.

Denotando por $\hat{y}_i$ o valor ajustado para a i -ésima observação da resposta (y_i), segundo o modelo (2.6) e utilizando o método de mínimos quadrados, pode-se realizar a decomposição:

$$\sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2 + \sum_{i=1}^n (\hat{y}_i - \bar{y})^2, \quad (2.7)$$

(SQT)
(SQE)
(SQR)

sendo que:

- SQT é a *soma de quadrados total*, que denota a variação total entre as observações y_i ;

- SQE é a *soma de quadrados dos resíduos*, que denota a variação das observações y_i em relação aos valores $\hat{y}_i$ ajustados pelo modelo;
- SQR é a *soma de quadrados da regressão*, que denota a diferença entre os valores $\hat{y}_i$ ajustados pelo modelo e a média amostral da variável resposta.

Quanto maior o valor de SQR , também chamado de *variação explicada*, maior a capacidade que o modelo tem de explicar a resposta de interesse. Em particular, caso haja um ajuste perfeito (ou seja, $\hat{y}_i = y_i \forall i$), é fácil ver que:

$$SQE = 0 \quad \text{e} \quad SQR = SQT,$$

na Equação (2.7).

Dessa forma, o coeficiente de determinação R^2 é definido de maneira muito plausível, segundo a Equação (2.8) a seguir, cujo objetivo é quantificar a proporção da variabilidade total de Y que é satisfatoriamente explicada pelas $p - 1$ variáveis explicativas do modelo:

$$R^2 = \frac{SQR}{SQT} = 1 - \frac{SQE}{SQT}, \quad 0 \leq R^2 \leq 1. \quad (2.8)$$

A denominação R^2 deve-se ao fato de essa quantidade coincidir com o quadrado do coeficiente de correlação linear de Pearson, no caso da regressão linear simples, entre os valores y_i da variável resposta e os valores x_i da variável explicativa, para $1 \leq i \leq n$.

2.4 Adaptação do R^2 para a Regressão Logística (R_{MQ}^2)

Autores como Lave (1970) e Efron (1978) sugeriram, para os modelos de regressão logística, a utilização de um coeficiente análogo ao R^2 da regressão linear, apresentado na Equação (2.8). Tal coeficiente será denotado por R_{MQ}^2 e é apresentado na Equação (2.9), a seguir.

$$R_{MQ}^2 = 1 - \frac{\sum_{i=1}^n [y_i - \hat{\pi}_1(\mathbf{x}_i)]^2}{\sum_{i=1}^n (y_i - \bar{y})^2}, \quad 0 \leq R_{MQ}^2 \leq 1. \quad (2.9)$$

Aqui, as observações y_i assumem apenas valores 0 ou 1, enquanto os valores $\hat{\pi}_1(\mathbf{x}_i)$ tratam-se das probabilidades preditas de sucesso associadas a cada observação.

Segundo crítica de Ryan (2008), não se faria adequada a aplicação do coeficiente de deter-

minação R_{MQ}^2 em contextos nos quais se deseja modelar uma resposta dicotômica. O autor afirmou que a analogia com a estimação por mínimos quadrados, nesse contexto, poderia induzir a violações importantes das principais suposições a respeito dos erros, tal como a de que eles possuem variância constante, conforme apontado também por Aldrich e Cnudde (1975). Ainda, Ryan (2008) alegou que R_{MQ}^2 seria um coeficiente muito sensível, que pode decrescer de forma expressiva por conta de pequenos desvios de ajuste do modelo. Por exemplo, modelos com alto grau de explicabilidade da variável resposta poderiam apresentar, com frequência, $R_{MQ}^2 \approx 0,10$.

Menard (2000) observou, de forma adicional, que a expressão de R_{MQ}^2 é matematicamente análoga à da regressão linear, mas não é conceitualmente análoga, uma vez que os processos de mínimos quadrados e de máxima verossimilhança lidam com diferentes quantidades a serem minimizadas e maximizadas, respectivamente. Cox e Wermuth (1992) também exploraram tal analogia. Ao realizar o ajuste de um modelo de regressão linear simples por mínimos quadrados, considerando como resposta as probabilidades condicionais de sucesso e uma única variável explicativa, tiveram uma aproximação muito boa para a curva logística nos casos em que $0,2 \leq \pi_1(x_i) \leq 0,8$, conforme observa-se na Figura 2.1.

Figura 2.1: Exemplo de curva logística (em preto) versus reta da regressão linear simples (em vermelho), ajustada em relação às probabilidades condicionais de sucesso, com destaque para os pontos em que $\pi_1(x) = 0.2$ e $\pi_1(x) = 0.8$, entre os quais a aproximação linear é razoável.

Consequentemente, o valor obtido de R_{MQ}^2 na regressão logística só seria coerente com o da regressão linear caso a distribuição da variável explicativa fosse controlada, de tal forma que $0,2 \leq \pi_1(x_i) \leq 0,8 \forall x_i$, o que seria contraintuitivo. Ao impor tal restrição na distribuição de X e nas probabilidades condicionais de sucesso, os autores demonstraram que o valor máximo assumido

por R_{MQ}^2 é de 0,36, mesmo em situações em que a variável explicativa contribui de forma expressiva para explicar a probabilidade do evento de interesse.

Já [Mittlböck e Schemper \(1996\)](#) defenderam a utilização de R_{MQ}^2 , levando em conta a sua consistência numérica com o R^2 da regressão linear e alegando facilidade de interpretação. Os autores responderam a diversas críticas de seus predecessores, afirmando que foram feitas sob a ótica de considerar o coeficiente de determinação como uma medida de qualidade de ajuste do modelo, e não de variabilidade explicada. Por outro lado, reconheceram a limitação no intervalo de valores de R_{MQ}^2 , apontada por [Cox e Wermuth \(1992\)](#), decidindo restringir suas simulações aos casos em que a distribuição da variável explicativa fosse tal que $0,2 \leq \pi_1(x_i) \leq 0,8 \ \forall x_i$, possibilitando a utilização de um modelo linear para basear suas conclusões a respeito de coeficientes R^2 alternativos.

Na busca por outros coeficientes, é natural que a própria função de verossimilhança sirva como base para estabelecer critérios alternativos de mensuração da variabilidade explicada, conforme mencionou [Menard \(2002\)](#). Diversas outras proposições, não baseadas na verossimilhança, também foram apresentadas para o modelo de regressão logística, como será visto ao longo do Capítulo 4.

2.5 Considerações Finais do Capítulo

Neste capítulo, foi realizada uma revisão das principais definições acerca do modelo de regressão logística múltipla, tanto com resposta dicotômica quanto multinomial, bem como do conceito de coeficiente de determinação na regressão linear múltipla. Argumentou-se a respeito do problema existente na utilização de um coeficiente análogo ao R^2 na regressão logística, o que leva à necessidade de se buscar por outros coeficientes que sejam adequados para este contexto.

No Capítulo 3, a seguir, serão apresentados os principais índices de acurácia preditiva empregados para avaliação de modelos de regressão logística, mediante fomento do problema da classificação (ou predição) de observações.

Capítulo 3

Acurácia Preditiva

Uma primeira forma de avaliar o grau de variabilidade explicada de um conjunto de variáveis, para o modelo de regressão logística, estabelece-se por meio dos índices de acurácia preditiva. Tais índices são baseados na comparação entre os valores observados e os valores preditos para a variável resposta, considerando a sua natureza categórica. A predição é realizada por meio de algum tipo de regra ou aplicação de ponto de corte às probabilidades condicionais de sucesso, obtidas a partir do modelo e dadas pelas Equações (2.2) e (2.4), para os casos de resposta dicotômica e multinomial, respectivamente.

Diferentemente do cálculo dos coeficientes de determinação do tipo R^2 , que leva em conta diretamente as probabilidades de sucesso preditas pelo modelo, a mensuração da acurácia preditiva envolve somente a comparação direta entre o “real” e o que foi, de alguma maneira, “previsto” pelo modelo. Os resultados são sumarizados em forma de uma tabela de classificação (Agresti, 2002), também conhecida como matriz de confusão (Hastie *et al.*, 2008). Os índices de acurácia são baseados, geralmente, nas frequências observadas em tal tabela (Hand, 2012) ou na teoria de curva ROC (Hosmer *et al.*, 2013).

Os índices de acurácia são de especial importância quando se elabora um modelo de regressão logística com finalidade preditiva. Um modelo preditivo tem como objetivo não apenas explicar os fatores que interferem na variável resposta e atribuir seus respectivos pesos, mas, também, estabelecer uma regra eficaz, por meio da qual novas observações possam ser classificadas em alguma categoria de resposta, de forma antecipada à observação da realidade. Mesmo quando não existe um interesse primário em realizar predições, os índices de acurácia podem servir como referência para se estabelecer relações com coeficientes do tipo R^2 , tendo em vista que estas são as duas abordagens essenciais que dizem respeito à avaliação da capacidade explanatória do modelo de regressão logística (Menard, 2000).

Neste capítulo, serão apresentados os principais índices de acurácia preditiva presentes na literatura, com especial ênfase para modelos de resposta dicotômica, para os quais o assunto foi explorado de forma mais extensiva.

3.1 O Problema de Classificação

Em diversas situações práticas, existe a necessidade de se classificar (ou “rotular”) novas observações com relação a um evento de natureza categórica, de maneira assertiva, a fim de se tomar uma ação específica. Em tais ocasiões, o interesse primário não se encontra nos valores em si das probabilidades associadas, mas sim na predição correta das observações segundo as classes da variável resposta.

Alguns exemplos de cenários em que a classificação pode ser necessária, quando se dispõe de respostas de natureza dicotômica, são:

- Futuros pacientes atendidos no sistema público de saúde, em determinada região do país, podem desenvolver ou não complicações clínicas associadas a determinada doença pré-existente. Caso sejam classificados como propensos a manifestar tais complicações, serão previamente tratados com a medicação A, mais potente; caso contrário, serão tratados com a medicação B, menos agressiva.
- Clientes de uma marca de vestuário podem aceitar ou não uma oferta, que estimula a compra de determinado produto. Caso sejam preditos como altamente propensos ao aceite, serão interpelados por meio de uma comunicação de marketing apropriada (e-mail, mala direta etc.); caso contrário, os clientes sequer serão abordados, de forma a reduzir os custos associados ao envio de comunicações.
- Futuras transações realizadas em cartões de crédito, intermediadas por uma empresa do segmento bancário, podem ser consideradas como potencialmente fraudulentas ou não. Caso sejam classificadas como não fraudulentas, a instituição bancária autorizará sua efetivação; caso contrário, serão bloqueadas no ato da compra, mitigando o impacto financeiro causado pelas fraudes.

Outros exemplos, quando se tem respostas de natureza multinomial, são:

- Quatro tipos de dieta podem ser atribuídos para filhotes de determinada espécie animal, com base em características genéticas e antropométricas (comprimento, peso etc.). Mediante pré-classificação da dieta mais adequada para cada indivíduo, pode-se administrá-la sem que haja

necessidade de testar sempre as quatro opções, poupando o animal de sucessivos processos de readaptação alimentar.

- Potenciais clientes de uma grande empresa de telefonia podem ser classificados com relação ao pacote de serviços mais adequado às suas necessidades, entre cinco opções disponíveis. Pode-se tomar como base as suas características sociodemográficas, bem como informações de renda e movimentações bancárias, fornecidas por instituições financeiras. O pacote mais adequado para cada indivíduo pode ser oferecido, mediante ligação telefônica de prospecção realizada pela empresa, maximizando a taxa de aceite e incrementando a carteira de clientes de forma efetiva.
- Viajantes de determinada companhia aérea, que já emitiram alguma passagem de avião no passado, podem ter a sua próxima viagem prevista como sendo de um dos três seguintes tipos: (i) viagem de negócios; (ii) viagem a lazer, sozinho; (iii) viagem a lazer, com acompanhante. Fundamentando-se no tipo de viagem mais provável, diferentes estratégias de comunicação e estímulo podem ser adotadas pela companhia, de forma personalizada por cliente, para incentivar a compra de novas passagens aéreas.

Vale ressaltar que, em todos os exemplos mencionados, as variáveis resposta são de natureza verdadeiramente nominal; ou seja, não se pode estabelecer um paralelo com variáveis resposta em escala contínua, que tenham sido discretizadas para fins de modelagem.

Amostras de construção e teste

Quando se deseja utilizar um modelo de classificação com efeito preditivo, isto é, para previsão da variável resposta associada a novas observações, costuma-se dividir a amostra total disponível em duas novas amostras (Bishop, 2006), a saber:

- *Amostra de construção (ou de treino)*

Subconjunto aleatório de observações destinadas ao ajuste do modelo, ou seja, para a seleção de variáveis explicativas e estimação dos parâmetros. Em geral, representa entre 70% e 80% da amostra inicial.

- *Amostra de teste*

Subconjunto complementar à amostra de construção, com observações destinadas à aplicação do modelo ajustado. Representa entre 20% e 30% da amostra inicial.

Uma das principais motivações para a separação de amostras de construção e teste consiste em evitar problemas de sobreajuste (*“overfitting”*), ou seja, situações em que o ajuste do modelo

é demasiadamente específico ao conjunto de dados considerado, podendo haver falhas graves na predição da resposta para novas observações. Essa preocupação é particularmente relevante no contexto de modelagem exaustiva, por meio de algoritmos sequenciais de aprendizado de máquina (*machine learning*). Dessa forma, a avaliação da acurácia de classificação tem como foco a amostra de teste, que é independente dos dados utilizados na construção (James *et al.*, 2013). Para efeito de notação, denotar-se-á como n_C o total de observações na amostra de construção e n_T o total de observações na amostra de teste, de forma que $n_C + n_T = n$.

A divisão aleatória dos conjuntos de construção e teste pode ser realizada diversas vezes durante o processo de modelagem, por meio da utilização de métodos de reamostragem, tais como o *bootstrap* e diferentes variantes de validação cruzada. Tais técnicas não são abordadas com detalhe neste trabalho, mas pode-se consultar mais detalhes em Hastie *et al.* (2008), James *et al.* (2013) e Dinov (2018).

Regra de decisão

Após definição das amostras para construção e teste do modelo, faz-se necessário o estabelecimento de uma regra de decisão (ou regra de classificação). Por meio de tal regra, pode-se, finalmente, atribuir a cada observação uma das categorias da resposta dicotômica ou multinomial, com base nas probabilidades preditas pelo modelo.

No caso de modelos com resposta dicotômica, em que cada observação pode ser predita como um “sucesso” ou um “fracasso”, o problema se trata de aplicar um ponto de corte c , $0 \leq c \leq 1$, às probabilidades preditas de sucesso $\hat{\pi}_1(\mathbf{x}_i)$, de tal forma que:

$$\hat{y}_i = \begin{cases} 1, & \text{se } \hat{\pi}_1(\mathbf{x}_i) \geq c, \\ 0, & \text{se } \hat{\pi}_1(\mathbf{x}_i) < c, \end{cases} \quad (3.1)$$

para $1 \leq i \leq n_T$.

Em muitos casos, admite-se $c = 0,5$ (Hosmer *et al.*, 2013). Entretanto, outros pontos de corte devem ser considerados, a depender da distribuição das probabilidades preditas na amostra selecionada, bem como a depender da importância atribuída a cada um dos dois possíveis erros de decisão: (i) prever uma nova observação como “sucesso”, quando a resposta real seria “fracasso”; ou (ii) prever uma nova observação como “fracasso”, quando a resposta real seria “sucesso”.

Uma vez definida a regra para atribuição das categorias preditas, pode-se construir uma tabela com o resumo dos resultados da classificação, para a amostra de observações utilizada para teste

do modelo (Tabela 3.1).

Real	Predito		Total
	Fracasso	Sucesso	
Fracasso	n_{00}	n_{01}	$n_{0\cdot}$
Sucesso	n_{10}	n_{11}	$n_{1\cdot}$
Total	$n_{\cdot 0}$	$n_{\cdot 1}$	n_T

Tabela 3.1: Tabela de classificação das observações da amostra de teste, para modelos com resposta dicotômica.

Já para modelos com resposta multinomial, a escolha da regra de classificação é mais sofisticada. Em trabalhos práticos de diferentes áreas do mercado, é comum observar a adoção de uma regra de decisão análoga à do caso dicotômico, em que se atribui como resposta predita a categoria com maior probabilidade associada, para cada observação. Ou seja:

$$\hat{y}_i = \begin{cases} 1, & \text{se } \max \{ \hat{\pi}_1(\mathbf{x}_i), \hat{\pi}_2(\mathbf{x}_i), \dots, \hat{\pi}_k(\mathbf{x}_i) \} = \hat{\pi}_1(\mathbf{x}_i), \\ 2, & \text{se } \max \{ \hat{\pi}_1(\mathbf{x}_i), \hat{\pi}_2(\mathbf{x}_i), \dots, \hat{\pi}_k(\mathbf{x}_i) \} = \hat{\pi}_2(\mathbf{x}_i), \\ \dots & \\ k, & \text{se } \max \{ \hat{\pi}_1(\mathbf{x}_i), \hat{\pi}_2(\mathbf{x}_i), \dots, \hat{\pi}_k(\mathbf{x}_i) \} = \hat{\pi}_k(\mathbf{x}_i), \end{cases}$$

para $1 \leq i \leq n_T$.

Quando existe desbalanceamento nas probabilidades gerais associadas a cada categoria na amostra selecionada, convém notar que a regra de classificação anterior pode penalizar radicalmente as categorias com baixa representatividade. Se tal desbalanceamento for significativo, grande parte das observações será predita como pertencente à categoria modal; ao passo que poucas observações (ou nenhuma, eventualmente) serão preditas como pertencentes às categorias que, por natureza, são mais esporádicas.

Uma estratégia alternativa consiste em comparar, para cada observação, as probabilidades preditas em cada categoria de resposta ($\hat{\pi}_1(\mathbf{x}_i), \dots, \hat{\pi}_k(\mathbf{x}_i)$) com as suas respectivas probabilidades gerais associadas ($\bar{\pi}_1, \dots, \bar{\pi}_k$). Dessa forma, pode-se estabelecer como regra de decisão a atribuição da classe com maior aumento relativo de probabilidade, também denominado “alavancamento” ou *lift* (Kuhn e Johnson, 2013; Tufféry, 2011). Ou seja:

$$\hat{y}_i = \begin{cases} 1, & \text{se } \max \{ \hat{r}_1(\mathbf{x}_i), \hat{r}_2(\mathbf{x}_i), \dots, \hat{r}_k(\mathbf{x}_i) \} = \hat{r}_1(\mathbf{x}_i), \\ 2, & \text{se } \max \{ \hat{r}_1(\mathbf{x}_i), \hat{r}_2(\mathbf{x}_i), \dots, \hat{r}_k(\mathbf{x}_i) \} = \hat{r}_2(\mathbf{x}_i), \\ \dots & \\ k, & \text{se } \max \{ \hat{r}_1(\mathbf{x}_i), \hat{r}_2(\mathbf{x}_i), \dots, \hat{r}_k(\mathbf{x}_i) \} = \hat{r}_k(\mathbf{x}_i), \end{cases}$$

em que $\hat{r}_j = \hat{\pi}_j(\mathbf{x}_i)/\bar{\pi}_j$, com $1 \leq j \leq k$ e $1 \leq i \leq n_T$.

Apesar de ser intuitivamente razoável, a regra baseada no maior alavancamento penaliza a classificação em categorias de alta representatividade, de forma inversa à regra que considerava a maior probabilidade, especialmente quando há desbalanceamento nas proporções gerais das classes. Com efeito, é fácil perceber que alavancamentos altos são mais prováveis de ocorrer para categorias cuja proporção geral na amostra é pequena (ex.: 1%) do que para categorias com proporções gerais mais altas (ex.: 50%).

Soluções para o problema de classificação em modelos com resposta multinomial vêm sendo desenvolvidas na literatura mais recente, especialmente no contexto dos algoritmos de *machine learning*. Algumas técnicas consistem em construir um modelo que já minimize os erros por categoria, penalizando-o com base em uma matriz $k \times k$ de “custos” de classificação errada (Cichosz, 2015). Outras envolvem rebalanceamento da amostra por meio da exclusão de observações das categorias de resposta mais representativas (Chawla *et al.*, 2002).

Pode-se resumir as quantidades de observações associadas a cada categoria de resposta real e de resposta predita, para a amostra de teste de um modelo multinomial, por meio da Tabela 3.2. A soma dos valores de sua diagonal principal resulta no total de observações corretamente classificadas.

Real	Predito			Total
	Categoria 1	...	Categoria k	
Categoria 1	n_{11}	...	n_{1k}	$n_{1\cdot}$
...	...	...	...	...
Categoria k	n_{k1}	...	n_{kk}	$n_{k\cdot}$
Total	$n_{\cdot 1}$	...	$n_{\cdot k}$	n_T

Tabela 3.2: Tabela de classificação das observações da amostra de teste, para modelos com resposta multinomial.

3.2 Índices de Acurácia Preditiva

A seguir, apresenta-se os principais índices de acurácia preditiva utilizados em modelos de regressão logística, especialmente para respostas dicotômicas. [Menard \(2000\)](#) considera que os índices de acurácia possuem estreita relação com os coeficientes de determinação do tipo R^2 , que serão examinados no Capítulo 4.

3.2.1 Acurácia Geral e por Categoria

A acurácia geral é a primeira medida que surge, naturalmente, para avaliação da qualidade preditiva de um modelo de classificação. Esta medida corresponde ao percentual de observações na amostra de teste para as quais se previu corretamente a categoria de resposta por meio do modelo, associado à regra de decisão envolvendo o ponto de corte c .

No modelo com resposta dicotômica, define-se a acurácia geral como:

$$A_c = \frac{n_{00} + n_{11}}{n_T}. \quad (3.2)$$

Analogamente, pode-se definir índices de acurácia associados a cada categoria de resposta. Quando a resposta de interesse é dicotômica, tais índices são mais conhecidos como especificidade e sensibilidade ([Hosmer *et al.*, 2013](#)), e são dados, respectivamente, por:

$$E_c = \frac{n_{00}}{n_{0.}}, \quad (3.3)$$

$$S_c = \frac{n_{11}}{n_{1.}}. \quad (3.4)$$

Evidentemente, o valor adotado para o ponto de corte c interfere diretamente nos resultados dos índices de acurácia geral e por categoria. A partir da Equação (3.1), é fácil perceber que, quanto menor o valor de c , maior o percentual de sensibilidade e menor o de especificidade, e vice-versa. A relação entre essas duas medidas será explorada nas Seções 3.2.2 e 3.2.3, a seguir.

Já para um modelo com resposta multinomial, a medida de acurácia geral é dada por:

$$A = \frac{\sum_{j=1}^k n_{jj}}{n_T}.$$

De forma análoga, pode-se calcular k índices de acurácia por categoria:

$$A_j = \frac{n_{jj}}{n_{j.}}, \quad 1 \leq j \leq k.$$

3.2.2 Área Abaixo da Curva ROC (AUC) e Índice de Gini (G)

Para modelos de regressão logística com resposta dicotômica, a curva ROC (do inglês *receiver operating characteristic*) é uma representação visual da relação entre as medidas de sensibilidade (S_c) e especificidade (E_c). Ela é definida a partir de um conjunto de pontos da forma $(1 - E_c, S_c)$, calculados para diversos valores do ponto de corte c , no intervalo de 0 a 1.

As curvas ROC apresentam formatos semelhantes aos exemplificados na Figura 3.1.

Exemplos de curvas ROC

Figura 3.1: Gráficos ilustrativos de curvas ROC, representando uma gradual diminuição de explicação da resposta por meio das variáveis explicativas (do exemplo 1 ao 4).

O valor da área abaixo da curva ROC, comumente abreviado como AUC , constitui uma medida tradicional para avaliação da acurácia de um modelo preditivo. Tal medida, que varia no intervalo

de 0,5 a 1, consiste em uma generalização da sensibilidade e especificidade, pois não se baseia em um ponto de corte único e previamente definido, mas sim em vários pares ordenados $(1 - E_c, S_c)$ obtidos para diferentes pontos de corte c . Visualmente, corresponde à união das áreas hachuradas em azul claro e azul escuro na Figura 3.1.

Hosmer *et al.* (2013) propuseram as seguintes faixas de valores de referência para o AUC , permitindo um julgamento acerca da capacidade discriminatória do modelo:

- $AUC \geq 0,9$: discriminância excepcional
- $0,8 \leq AUC < 0,9$: discriminância excelente
- $0,7 \leq AUC < 0,8$: discriminância aceitável
- $0,5 \leq AUC < 0,7$: discriminância baixa

Os autores apontaram, entretanto, que modelos com AUC superior a 0,9 são muito difíceis de se obter na prática, a não ser em situações de separação quase completa (Ryan, 2008), em que quase todos os sucessos observados estão associados aos valores mais altos (ou mais baixos) de alguma variável explicativa. Já quando o AUC é igual a 0,5, tem-se uma situação em que o modelo provê discriminância semelhante a de uma atribuição aleatória dos sucessos e fracassos, com mesma probabilidade; ou seja, as variáveis explicativas não ajudariam a explicar a resposta de interesse.

Algumas extensões da teoria de curva ROC foram propostas para modelos com resposta multinomial, especialmente para o caso de três categorias, com destaque para o trabalho de Scurfield (1996). O assunto foi mais amplamente discutido, entretanto, para o caso específico de respostas ordinais, tal como por Nakas e Yiannoutsos (2004) e van Belle *et al.* (2006).

Já o índice de Gini (Hand, 2012) consiste em uma padronização do AUC para o intervalo de 0 a 1. Seu valor é proporcional à área compreendida entre a curva ROC e a reta $S_c = 1 - E_c$, ou seja, a região hachurada em azul escuro, na Figura 3.1. Sua expressão é dada por:

$$G = 2 \cdot AUC - 1. \quad (3.5)$$

O índice de Gini pode ser interpretado, também, como o acréscimo relativo no valor do AUC quando se utiliza as variáveis explicativas, em relação a uma classificação aleatória. A correspondência entre esses dois índices se dá tanto do ponto de vista geométrico quanto algébrico.

3.2.3 Índices de Youden (J) e Kolmogorov-Smirnov (KS)

Outro índice tradicional, proposto por Youden (1950) e aplicável para modelos com resposta dicotômica, baseia-se em um balanceamento idêntico das medidas de sensibilidade e especificidade,

dado um ponto de corte c específico:

$$J_c = S_c + E_c - 1.$$

Do ponto de vista algébrico, o índice J_c de Youden pode variar entre -1 e 1. Entretanto, quando $J_c = 0$, tem-se uma situação em que o modelo não provê discriminância de sucessos e fracassos superior à de uma classificação aleatória, pois, neste caso, espera-se que $S_c + E_c = 1$. Dessa forma, os valores plausíveis de J_c para modelos com alguma discriminância encontram-se no intervalo de 0 a 1.

Segundo Hand (2012), um índice derivado do J_c de Youden que merece especial destaque é o de Kolmogorov-Smirnoff, abreviado como KS , bastante utilizado em modelos de *score* de crédito com resposta dicotômica (Fang e Chen, 2019). É definido como sendo o valor máximo do índice de Youden, ou seja:

$$KS = \max_c (S_c + E_c - 1). \quad (3.6)$$

A partir de sua expressão, é fácil notar que KS corresponde à distância vertical máxima entre a curva ROC e a reta $S_c = 1 - E_c$ (Figura 3.2), podendo variar entre 0 e 1.

Figura 3.2: Gráfico ilustrativo do índice KS , corresponde à distância máxima entre a curva ROC e a reta $S_c = 1 - E_c$.

Além de sua facilidade de interpretação, o fato de atribuir pesos iguais para sensibilidade (S_c) e especificidade (E_c) representa grande vantagem do índice KS em relação à medida de acurácia

geral. Tal premissa é essencial quando se tem desbalanceamento nas probabilidades gerais de sucesso e fracasso, isto é, proporções razoavelmente distantes do cenário em que $\bar{\pi}_0 = \bar{\pi}_1 = 0,5$.

Alternativamente, pode-se propor uma generalização do índice *KS*. Trata-se de definir uma função de perda, baseada na ponderação das proporções amostrais de sucessos e fracassos que foram classificados de forma incorreta, dado um ponto de corte c e dois “custos” distintos, h_0 e h_1 (Hand, 2012):

$$L_c = h_0 \cdot \bar{\pi}_0(1 - E_c) + h_1 \cdot \bar{\pi}_1(1 - S_c). \quad (3.7)$$

A definição dos valores h_0 e h_1 é uma questão arbitrária, que depende da importância que se deseja atribuir à classificação errada de sucessos e fracassos, em cada diferente contexto prático. Em seguida, pode-se proceder à escolha numérica de um ponto de corte ótimo c' , de forma a minimizar a função de perda da Equação (3.7), ou seja:

$$c' = \arg \min_c L_c.$$

3.2.4 Outros índices

Alguns outros índices já foram propostos para avaliação da acurácia preditiva, balanceando sensibilidade e especificidade, bem como outras taxas relacionadas à classificação real *versus* predita. Eles são apresentados a seguir.

- ***F*-score (ou F_1 -score)**

Proposto por Chinchor (1992), corresponde à média harmônica entre a sensibilidade (S_c) e a quantidade $PPV = n_{11}/n_{.1}$, correspondente à porcentagem de sucessos reais entre as observações que foram preditas como sucesso. Ou seja,

$$F_c = 2 \cdot \frac{S_c \cdot PPV}{S_c + PPV}.$$

Apesar de ser frequentemente utilizado na avaliação de modelos de *machine learning* em aplicações práticas, o *F*-score apresenta ao menos duas limitações importantes. Sua expressão atribui mesmo peso aos índices S_c e PPV , o que é algo bastante questionável, segundo Hand e Christen (2018). Adicionalmente, não leva em conta a classificação correta de fracassos, o que pode ser de grande impacto em situações de grande desbalanceamento (Powers, 2011).

- **Kappa de Cohen (K)**

Apresentado por [Cohen \(1960\)](#) e especialmente utilizado em estudos na área de Psicologia, tem como objetivo balizar a acurácia geral do modelo (A_c), dada pela Equação (3.2), a partir de uma medida de acurácia esperada (A_E), que corresponde à proporção de observações que seriam classificadas corretamente de forma “natural”, mesmo quando não são consideradas variáveis explicativas. Sua expressão é dada por meio da razão:

$$K_c = \frac{A_c - A_E}{1 - A_E},$$

em que A_E é a acurácia esperada naturalmente. No caso de resposta dicotômica, tem-se que:

$$A_E = \bar{\pi}_0^2 + \bar{\pi}_1^2.$$

A generalização para respostas multinomiais é imediata:

$$A_E = \sum_{i=1}^k \bar{\pi}_k^2.$$

O coeficiente K_c varia de 0 a 1, sendo que seu valor pode ser interpretado como o percentual de incremento na acurácia proporcionado pelo modelo, em relação a uma classificação aleatória das observações.

Conforme apontado por [Powers \(2011\)](#), o coeficiente de Cohen propaga a mesma limitação do índice de acurácia geral, ou seja, não é devidamente penalizado para casos de grande desbalanceamento entre sucessos e fracassos. Para ilustrar tal problema, pode-se tomar como exemplo um cenário com resposta dicotômica e $\bar{\pi}_1 = 0,03$. Caso fosse adotada uma regra por meio da qual 100% das observações fossem classificadas como fracassos (isto é, $c = 1$), obter-se-ia uma acurácia de 0,97. Por outro lado, caso a predição de sucessos e fracassos fosse realizada de forma aleatória, respeitando a proporção geral de sucessos na amostra, a acurácia natural seria de $0,97^2 + 0,03^2 = 0,9418$. Com base nesses números, o coeficiente Kappa de Cohen associado seria igual a:

$$K_c = \frac{0,97 - 0,9418}{1 - 0,9418} \approx 0,485.$$

Ou seja, esta regra de classificação “degenerada” proporcionaria cerca de 48,5% de todo o incremento possível na acurácia, em relação ao cenário de aleatoriedade.

- **Coefficiente ϕ de Pearson (ou Correlação de Matthews)**

Consiste no coeficiente de correlação de Pearson usual, aplicado a uma tabela de contingência com duas variáveis dicotômicas. Pode ser escrito em função da estatística qui-quadrado, como:

$$\phi_c = \sqrt{\frac{\chi^2}{n_T}},$$

ou em função das frequências absolutas da tabela de contingência (Tabela 3.1), como:

$$\phi_c = \frac{n_{00} \cdot n_{11} - n_{01} \cdot n_{10}}{\sqrt{n_{\cdot 0} \cdot n_{\cdot 1} \cdot n_{0 \cdot} \cdot n_{1 \cdot}}}.$$

Originalmente, o coeficiente ϕ_c destina-se à investigação do grau de associação entre duas variáveis qualitativas dicotômicas. Quando aplicado ao problema de classificação em modelos de *machine learning*, comparando resposta real *versus* resposta predita, é comumente denominado como coeficiente de correlação de Matthews ou, abreviadamente, MCC (Matthews, 1975).

Autores como Powers (2011) e Chicco e Jurman (2020) defenderam este índice como uma opção mais adequada para avaliação da acurácia preditiva, se comparado ao percentual de acurácia geral, ao *F*-score ou ao coeficiente Kappa de Cohen. Entre todas essas propostas, o coeficiente ϕ é o único que leva em conta, em sua composição algébrica, todas as principais taxas de interesse da classificação binária: verdadeiros positivos e negativos; falsos positivos e negativos. Dessa forma, ϕ assumirá valores altos à medida que o modelo, associado a uma regra de classificação adequada, apresentar a capacidade de prever corretamente tanto os sucessos quanto os fracassos.

3.3 Considerações Finais do Capítulo

Após introdução acerca do problema de classificação de observações, assunto de especial interesse em modelos de natureza preditiva, realizou-se uma revisão a respeito dos principais índices de acurácia propostos na literatura, com ênfase para respostas dicotômicas.

Levando em conta as críticas apresentadas, alguns índices mostraram-se mais adequados, por lidarem melhor com situações de grande desbalanceamento nas proporções associadas a cada categoria; ou seja, não menosprezam o acerto associado a classes de ocorrência menos frequente. Tais índices são o de Kolmogorov-Smirnov (*KS*), o de Gini (ou, equivalentemente, o *AUC*) e o coeficiente ϕ_c de Pearson. O cálculo deste último, particularmente, depende da definição prévia de um ponto de corte c , ao contrário dos anteriores.

Estes índices poderão servir como balizadores para avaliação de diversas alternativas de coeficientes de determinação do tipo R^2 , que serão apresentadas no próximo capítulo.

Capítulo 4

Propostas de Coeficientes de Determinação do Tipo R^2

Diversas alternativas de coeficientes de determinação, denominados comumente como sendo de tipo R^2 (R^2 -analogs, pseudo- R^2 s etc.), foram derivadas para utilização em modelos de regressão logística. Tal como no caso da regressão linear, eles também podem ser chamados de índices de variabilidade explicada (Schemper, 2003), tendo em vista que compartilham do objetivo de quantificar o grau de variabilidade da variável resposta, agora de natureza dicotômica ou multinomial, que é satisfatoriamente explicada por meio das variáveis explicativas utilizadas no modelo.

Neste capítulo, serão analisadas as principais propostas elencadas em trabalhos anteriores, examinando seus aspectos teóricos, bem como as posições favoráveis e desfavoráveis dos autores a respeito de cada coeficiente.

4.1 Medidas Baseadas na Verossimilhança

Uma das fontes naturais para buscar alternativas de coeficiente de determinação é a função de verossimilhança ou, alternativamente, o seu logaritmo, tendo em vista que a estimação de modelos de regressão logística é realizada, usualmente, por meio do procedimento de máxima verossimilhança.

Conforme defendido por Menard (2000), coeficientes de determinação baseados na função de verossimilhança apresentariam uma vantagem de interpretação, pois, para alguns deles, pode-se estabelecer uma analogia algébrica direta com as somas de quadrados obtidas da estimação por mínimos quadrados. Por esse motivo, o autor restringiu sua investigação apenas a propostas de coeficientes baseados na verossimilhança.

Já Cameron e Windmeijer (1997) argumentaram que tais alternativas de R^2 são de especial interesse quando se busca por um coeficiente de determinação que apresente valores necessariamente crescentes à medida que novas variáveis sejam acrescentadas ao modelo, tal como é o comportamento da função de verossimilhança. Caso contrário, outras medidas - ou correções apropriadas - deveriam ser empregadas.

Tais medidas de R^2 envolvem, essencialmente, comparações entre o modelo completo – ou seja, com todas as variáveis explicativas relevantes para o evento de interesse – e o modelo basal, que possui apenas o intercepto. As comparações podem ser feitas por meio dos valores da função de verossimilhança, ou, alternativamente, do seu logaritmo, após estimação dos parâmetros. Por meio desse procedimento, tem-se como objetivo quantificar a proporção de variabilidade da variável resposta (no caso, a probabilidade condicional de sucesso) que é explicada pelo modelo completo, em relação ao modelo basal.

4.1.1 Função de Verossimilhança

Modelo de resposta dicotômica

Para o modelo de regressão logística múltipla introduzido na Seção 2.1, cuja resposta é dicotômica, a expressão da função de verossimilhança associada à i -ésima observação é dada por

$$L_{M_i}(\boldsymbol{\beta} | y_i, \mathbf{x}_i) = [\pi_1(\mathbf{x}_i)]^{y_i} \cdot [1 - \pi_1(\mathbf{x}_i)]^{1-y_i}, \quad (4.1)$$

em que $\pi_1(\mathbf{x}_i) = 1/[1 + \exp(-\mathbf{x}_i^T \boldsymbol{\beta})]$, conforme Equação (2.2).

Consequentemente, pode-se derivar a verossimilhança conjunta do modelo, para uma amostra aleatória de n observações, em função de $\boldsymbol{\beta}$:

$$\begin{aligned} L_M(\boldsymbol{\beta} | \mathbf{y}, \mathbf{x}_1, \dots, \mathbf{x}_n) &= \prod_{i=1}^n \left\{ [\pi_1(\mathbf{x}_i)]^{y_i} \cdot [1 - \pi_1(\mathbf{x}_i)]^{1-y_i} \right\} \\ &= \prod_{i=1}^n \left\{ \left(\frac{\exp(\mathbf{x}_i^T \boldsymbol{\beta})}{1 + \exp(\mathbf{x}_i^T \boldsymbol{\beta})} \right)^{y_i} \cdot \left(\frac{1}{1 + \exp(\mathbf{x}_i^T \boldsymbol{\beta})} \right)^{1-y_i} \right\} \\ &= \prod_{i=1}^n \left\{ [\exp(\mathbf{x}_i^T \boldsymbol{\beta})]^{y_i} \cdot [1 + \exp(\mathbf{x}_i^T \boldsymbol{\beta})]^{-1} \right\}. \end{aligned} \quad (4.2)$$

A partir da Equação (4.2), pode-se obter o logaritmo da função de verossimilhança conjunta do modelo:

$$\ell_M(\boldsymbol{\beta} | \mathbf{y}, \mathbf{x}_1, \dots, \mathbf{x}_n) = \sum_{i=1}^n \left\{ y_i \mathbf{x}_i^T \boldsymbol{\beta} - \log [1 + \exp(\mathbf{x}_i^T \boldsymbol{\beta})] \right\}. \quad (4.3)$$

O cálculo da verossimilhança e do logaritmo da verossimilhança de um particular modelo ajustado é feito, respectivamente, a partir das Equações (4.2) e (4.3), substituindo $\boldsymbol{\beta}$ por $\hat{\boldsymbol{\beta}}$, que é o vetor que contém os valores estimados dos parâmetros do modelo, obtidos pelo processo de máxima verossimilhança:

$$L_M(\hat{\boldsymbol{\beta}} | \mathbf{y}, \mathbf{x}_1, \dots, \mathbf{x}_n) = \prod_{i=1}^n \left\{ \left[\exp(\mathbf{x}_i^T \hat{\boldsymbol{\beta}}) \right]^{y_i} \cdot \left[1 + \exp(\mathbf{x}_i^T \hat{\boldsymbol{\beta}}) \right]^{-1} \right\}, \quad (4.4)$$

$$\ell_M(\hat{\boldsymbol{\beta}} | \mathbf{y}, \mathbf{x}_1, \dots, \mathbf{x}_n) = \sum_{i=1}^n \left\{ y_i \mathbf{x}_i^T \hat{\boldsymbol{\beta}} - \log [1 + \exp(\mathbf{x}_i^T \hat{\boldsymbol{\beta}})] \right\}. \quad (4.5)$$

Convém, ainda, escrever as expressões da verossimilhança e de seu logaritmo para o modelo basal, ajustado apenas com um intercepto $\beta_0 = \hat{\beta}_0$ e sem variáveis explicativas:

$$L_0(\hat{\boldsymbol{\beta}} | \mathbf{y}) = L_0(\hat{\beta}_0 | \mathbf{y}) = \prod_{i=1}^n \left\{ \left[\exp(\hat{\beta}_0) \right]^{y_i} \cdot \left[1 + \exp(\hat{\beta}_0) \right]^{-1} \right\}, \quad (4.6)$$

$$\ell_0(\hat{\boldsymbol{\beta}} | \mathbf{y}) = \ell_0(\hat{\beta}_0 | \mathbf{y}) = \sum_{i=1}^n \left\{ \hat{\beta}_0 y_i - \log [1 + \exp(\hat{\beta}_0)] \right\}. \quad (4.7)$$

Modelo de resposta multinomial

Já para o modelo de regressão logística múltipla com resposta multinomial, introduzido na Seção 2.2, são definidas, inicialmente, variáveis aleatórias dicotômicas Y_j , $1 \leq j \leq k$, que assumem os seguintes valores para a i -ésima observação, com $1 \leq i \leq n$:

$$Y_{ji} = \begin{cases} 1, & \text{se } Y_i = j, \\ 0, & \text{caso contrário.} \end{cases}$$

A expressão da função de verossimilhança associada à i -ésima observação é dada por

$$L_{M_i}(\boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_{k-1} | y_{1,i}, \dots, y_{k,i}, \mathbf{x}_i) = [\pi_1(\mathbf{x}_i)]^{y_{1,i}} \cdot \dots \cdot [\pi_k(\mathbf{x}_i)]^{y_{k,i}}, \quad (4.8)$$

em que $\pi_j(\mathbf{x}_i)$ são dados pela Equações (2.4) e (2.5).

Denotando por $\mathbf{y}_j$ os vetores de n valores observados para as variáveis Y_j , $1 \leq j \leq k$, a verossimilhança conjunta do modelo, para uma amostra aleatória de n observações e em função de $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_{k-1})^T$, pode ser calculada como:

$$\begin{aligned}
 L_M(\boldsymbol{\beta} | \mathbf{y}_1, \dots, \mathbf{y}_k, \mathbf{x}_1, \dots, \mathbf{x}_n) &= \prod_{i=1}^n \left\{ [\pi_1(\mathbf{x}_i)]^{y_{1,i}} \cdot \dots \cdot [\pi_k(\mathbf{x}_i)]^{y_{k,i}} \right\} \\
 &= \prod_{i=1}^n \left\{ \left[\frac{\pi_1(\mathbf{x}_i)}{\pi_k(\mathbf{x}_i)} \right]^{y_{1,i}} \cdot \dots \cdot \left[\frac{\pi_{k-1}(\mathbf{x}_i)}{\pi_k(\mathbf{x}_i)} \right]^{y_{k-1,i}} \cdot [\pi_k(\mathbf{x}_i)]^{y_{1,i}} \cdot \dots \cdot [\pi_k(\mathbf{x}_i)]^{y_{k,i}} \right\} \\
 &= \prod_{i=1}^n \left[\frac{\prod_{m=1}^{k-1} \exp(\mathbf{x}_i^T \boldsymbol{\beta}_m)^{y_{m,i}}}{1 + \sum_{m=1}^{k-1} \exp(\mathbf{x}_i^T \boldsymbol{\beta}_m)} \right]. \tag{4.9}
 \end{aligned}$$

A partir da Equação (4.9), pode-se obter o logaritmo da função de verossimilhança conjunta do modelo multinomial:

$$\ell_M(\boldsymbol{\beta} | \mathbf{y}_1, \dots, \mathbf{y}_k, \mathbf{x}_1, \dots, \mathbf{x}_n) = \sum_{i=1}^n \left\{ \left(\sum_{m=1}^{k-1} y_{m,i} \mathbf{x}_i^T \boldsymbol{\beta}_m \right) - \log \left(1 + \sum_{m=1}^{k-1} \exp(\mathbf{x}_i^T \boldsymbol{\beta}_m) \right) \right\}. \tag{4.10}$$

O cálculo da verossimilhança e do logaritmo da verossimilhança de um particular modelo ajustado é feito, respectivamente, a partir das Equações (4.9) e (4.10), substituindo $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_{k-1})^T$ pelos valores estimados $\hat{\boldsymbol{\beta}} = (\hat{\boldsymbol{\beta}}_1, \dots, \hat{\boldsymbol{\beta}}_{k-1})^T$:

$$L_M(\hat{\boldsymbol{\beta}} | \mathbf{y}_1, \dots, \mathbf{y}_k, \mathbf{x}_1, \dots, \mathbf{x}_n) = \prod_{i=1}^n \left(\frac{\prod_{m=1}^{k-1} \exp(\mathbf{x}_i^T \hat{\boldsymbol{\beta}}_m)^{y_{m,i}}}{1 + \sum_{m=1}^{k-1} \exp(\mathbf{x}_i^T \hat{\boldsymbol{\beta}}_m)} \right), \tag{4.11}$$

$$\ell_M(\hat{\boldsymbol{\beta}} | \mathbf{y}_1, \dots, \mathbf{y}_k, \mathbf{x}_1, \dots, \mathbf{x}_n) = \sum_{i=1}^n \left\{ \left(\sum_{m=1}^{k-1} y_{m,i} \mathbf{x}_i^T \hat{\boldsymbol{\beta}}_m \right) - \log \left(1 + \sum_{m=1}^{k-1} \exp(\mathbf{x}_i^T \hat{\boldsymbol{\beta}}_m) \right) \right\}. \tag{4.12}$$

Já as expressões da verossimilhança e de seu logaritmo para o modelo basal, ajustado apenas com interceptos $\beta_{m,0} = \hat{\beta}_{m,0}$, $1 \leq m \leq k-1$, e sem variáveis explicativas, são dadas, respectivamente, por:

$$L_0(\hat{\boldsymbol{\beta}} | \mathbf{y}_1, \dots, \mathbf{y}_k) = L_0(\hat{\beta}_{1,0}, \dots, \hat{\beta}_{k-1,0} | \mathbf{y}_1, \dots, \mathbf{y}_k)$$

$$= \prod_{i=1}^n \left(\frac{\prod_{m=1}^{k-1} \exp(\hat{\beta}_{m,0})^{y_{m,i}}}{1 + \sum_{m=1}^{k-1} \exp(\hat{\beta}_{m,0})} \right), \quad (4.13)$$

$$\begin{aligned} \ell_0(\hat{\beta} | \mathbf{y}_1, \dots, \mathbf{y}_k) &= \ell_0(\hat{\beta}_{1,0}, \dots, \hat{\beta}_{k-1,0} | \mathbf{y}_1, \dots, \mathbf{y}_k) \\ &= \sum_{i=1}^n \left\{ \left(\sum_{m=1}^{k-1} y_{m,i} \cdot \hat{\beta}_{m,0} \right) - \log \left(1 + \sum_{m=1}^{k-1} \exp(\hat{\beta}_{m,0}) \right) \right\}. \end{aligned} \quad (4.14)$$

Por facilidade de notação, os valores das verossimilhanças dos modelos ajustados serão denotados apenas por $L_M(\hat{\beta})$ e $L_0(\hat{\beta})$ nas próximas seções (substituindo L por ℓ , no caso dos logaritmos), tanto para os casos de resposta dicotômica quanto multinomial.

4.1.2 Razão de Logaritmos das Verossimilhanças (R_{MF}^2)

A razão de logaritmos das verossimilhanças, proposta inicialmente por [McFadden \(1974\)](#) e rediscutida por diversos outros autores ([DeMaris, 1992](#); [Dhrymes, 1986](#); [Hagle e Mitchell, 1992](#); [Knoke e Burke, 1980](#); [Menard, 2000](#)), é uma das alternativas mais conhecidas de coeficiente de determinação para modelos de regressão logística.

A expressão do coeficiente R_{MF}^2 é baseada na função desvio, muito empregada para avaliação da qualidade de ajuste de modelos lineares generalizados. Ela estabelece uma comparação razoável entre o modelo em questão e o modelo saturado, por meio da razão de verossimilhanças:

$$D = -2 \log \frac{L_M(\hat{\beta})}{L_S(\hat{\beta})},$$

em que L_M é dada pelas Equações (4.4) e (4.11) nos casos de resposta dicotômica e multinomial, respectivamente, e L_S é a verossimilhança do modelo saturado, que possui a mesma quantidade de parâmetros e de observações.

No modelo saturado, os valores preditos das probabilidades condicionais de sucesso coincidem com os próprios valores observados. Ou seja, no caso de resposta dicotômica,

$$\hat{\pi}_1(\mathbf{x}_i) = y_i,$$

e no caso de resposta multinomial,

$$\hat{\pi}_j(\mathbf{x}_i) = y_{ji}.$$

Utilizando tais resultados nas Equações (4.1) e (4.8), respectivamente, tem-se que o logaritmo da verossimilhança da i -ésima observação, no modelo saturado, se reduz a:

$$\ell_{Si}(y_i) = y_i \log y_i I(y_i = 1) + (1 - y_i) \log(1 - y_i) I(y_i = 0),$$

no modelo com resposta dicotômica, e a:

$$\ell_{Si}(y_{1,i}, \dots, y_{k,i}) = y_{1,i} \log(y_{1,i}) I(y_{1,i} = 1) + \dots + y_{k,i} \log(y_{k,i}) I(y_{k,i} = 1),$$

no modelo com resposta multinomial, sendo $I(\cdot)$ a função indicadora, empregada devido à indefinição da quantidade $\log(0)$.

É fácil ver que as duas expressões anteriores assumem valor 0 para todas as n observações, tendo em vista que $y_i = 0$ ou $y_i = 1 \forall i$, no caso dicotômico, bem como $y_{j,i} = 0$ ou $y_{j,i} = 1 \forall i, j$, no caso multinomial. Dessa forma, as funções desvio, para o modelo sob investigação e para o modelo que contém apenas o intercepto, respectivamente, são simplificadas na regressão logística como:

$$D = -2 \ell_M(\hat{\beta}) \quad \text{e} \quad D_0 = -2 \ell_0(\hat{\beta}),$$

em que, para resposta dicotômica, ℓ_M e ℓ_0 são dados pelas Equações (4.5) e (4.7); e para resposta multinomial, são dados pelas Equações (4.12) e (4.14), respectivamente.

A utilização da função logarítmica e a multiplicação pelo fator -2 apresentam uma finalidade. Assintoticamente, a quantidade $D_0 - D$ segue uma distribuição qui-quadrado com k graus de liberdade, sendo k igual à quantidade de parâmetros utilizados no modelo sob investigação, excluindo o intercepto (Haggle e Mitchell, 1992; Menard, 2000). Tal resultado possibilita a realização de testes de hipóteses, aspecto que não será avaliado no escopo deste trabalho.

Conforme salientado por DeMaris (1992), os valores D_0 e D possuem uma analogia com as somas de quadrados SQT e SQE , respectivamente, tendo em vista que SQE é proporcional ao logaritmo da verossimilhança do modelo de regressão linear, quando respeitada a suposição de normalidade da distribuição condicional dos erros. Além disso, Menard (2000) apontou outra semelhança. O autor recordou que, do mesmo modo que se minimiza a soma de quadrados dos resíduos (SQE) na estimação por mínimos quadrados, de forma que o R^2 tradicional serve para

quantificar a “força” dessa minimização, busca-se, na regressão logística, maximizar a verossimilhança do modelo final, ou, alternativamente, minimizar a quantidade de interesse $-2 \ell_M$. Segundo essa interpretação, a proposta de coeficiente de determinação R_{MF}^2 surge naturalmente, de forma análoga à Equação (2.8), como:

$$R_{MF}^2 = \frac{D_0 - D}{D_0} = 1 - \frac{\ell_M(\hat{\beta})}{\ell_0(\hat{\beta})}. \quad (4.15)$$

Ben-Akiva e Lerman (1985) propuseram um ajuste ao coeficiente de McFadden (1974) a partir dos graus de liberdade, cuja utilização é desejável quando se tem duas ou mais variáveis explicativas (Equação (4.16)). Segundo os autores, caso sejam incluídas variáveis que não ajudem a explicar de forma satisfatória a resposta, o coeficiente ajustado, aqui denotado como R_{MFadj}^2 , pode decrescer, ao contrário do que ocorreria com R_{MF}^2 . Sua expressão é dada por:

$$R_{MFadj}^2 = 1 - \frac{\ell_M(\hat{\beta}) - p}{\ell_0(\hat{\beta})}, \quad (4.16)$$

em que p é a quantidade de parâmetros estimados no modelo.

Dhrymes (1986) mencionou as boas propriedades da medida R_{MF}^2 , ressaltando que ela varia no intervalo de 0 a 1 e que, quanto mais próxima de 1, maior a proporção de variabilidade explicada pelas variáveis explicativas, de forma claramente reconhecível. O autor defendeu esta como sendo a única alternativa plausível de coeficiente de determinação para a regressão logística.

Por outro lado, Cox e Snell (1989) fizeram uma crítica pertinente a essa proposta, no contexto de seu trabalho. Devido ao fato de R_{MF}^2 não tratar de uma comparação direta entre valores observados e valores preditos pelo modelo, os autores alegaram que não seria possível denominá-lo como uma verdadeira medida de qualidade de ajuste. Entretanto, tal objeção não apresenta grande relevância no contexto deste trabalho, que tem como objetivo tratar tais medidas sob a ótica de coeficientes de determinação, e não como medidas de diagnóstico da qualidade de ajuste do modelo como um todo.

DeMaris (1992) mencionou outra limitação deste coeficiente. Segundo ele, a interpretação de R_{MF}^2 é prejudicada devido à presença do fator -2 na fórmula da função desvio, o que impediria de estabelecer um paralelo direto com a proporção da variabilidade explicada. Alguns anos depois, Menard (2000) contestou tal colocação, defendendo que R_{MF}^2 pode ser claramente identificado como um coeficiente que mensura a explicação da variabilidade da variável resposta por meio do modelo, não havendo problemas de interpretação. Inclusive, o autor considerou R_{MF}^2 como sendo a melhor entre um conjunto de diversas medidas, devido à razoabilidade da sua expressão, tal como

argumentado também por [Hauser \(1977\)](#).

Outro alerta apresentado por [Hagle e Mitchell \(1992\)](#), com base em resultados de simulações, foi de que os valores assumidos por R_{MF}^2 tendem a ser subestimados quando a variável resposta original, de natureza latente, segue uma distribuição contínua, tendo sido discretizada durante a coleta ou análise dos dados. Com relação a esta objeção, [DeMaris \(1992\)](#) respondeu que, em grande parte dos casos em que a regressão logística é utilizada, sequer existe uma distribuição contínua latente por trás da resposta de interesse, que pode ser verdadeiramente categórica nominal. Além disso, [Menard \(2000\)](#) defendeu que R_{MF}^2 apresenta resultados mais coerentes do que outras medidas, quando tal distribuição latente, se existir, for bimodal ou tiver assimetria marcante, o que pode ser bastante comum.

4.1.3 Coeficiente de Contingência Simples (R_{AN}^2) e Corrigido (R_{HM}^2)

Outro conhecido coeficiente de determinação, proposto por [Aldrich e Nelson \(1984\)](#), é dado pela expressão a seguir:

$$R_{AN}^2 = \frac{D_0 - D}{n + D_0 - D}. \quad (4.17)$$

Tendo em vista que a quantidade $D_0 - D$ segue, assintoticamente, uma distribuição qui-quadrado, [Menard \(2000\)](#) ressaltou a semelhança entre R_{AN}^2 e o quadrado do tradicional coeficiente de contingência para variáveis categóricas, dado por:

$$C = \sqrt{\frac{\chi^2}{n + \chi^2}}, \quad \text{em que } \chi^2 = \sum_{i,j} \frac{(O_{ij} - E_{ij})^2}{E_{ij}},$$

sendo que χ^2 é a estatística do teste de independência entre as duas variáveis categóricas examinadas na tabela de contingência. Considera-se que as variáveis sob investigação são fortemente associadas quando C assume valor próximo de 1, e fracamente associadas quando C é próximo de 0.

Nos anos seguintes, autores como [Hagle e Mitchell \(1992\)](#) e [Veall e Zimmermann \(1994\)](#) apontaram como desvantagem o fato de que R_{AN}^2 nunca assume valor igual a 1, mesmo quando as variáveis independentes explicam de forma perfeita a probabilidade de interesse. Tal limitação, que parece não ter sido levada em conta por [Aldrich e Nelson \(1984\)](#), se deve à presença da quantidade de observações (n) no denominador da expressão. [Menard \(2000\)](#) também criticou tal coeficiente, por considerá-lo como tendo uma interpretação menos intuitiva, apesar da analogia com o coeficiente de contingência.

Tanto para o caso dicotômico quanto multinomial, é fácil ver, pelas Equações (4.2) e (4.9), que um modelo que provê total explicabilidade da resposta tem valor de verossimilhança igual a 1, pois, neste caso, as probabilidades previstas assumem apenas valores 0 ou 1. Consequentemente, $D = -2\ell_M(\hat{\beta}) = 0$ e o valor máximo que o coeficiente R_{AN}^2 poderia assumir é dado por:

$$\text{LS}(R_{AN}^2) = \frac{D_0}{n + D_0} = \frac{-2\ell_0(\hat{\beta})}{n - 2\ell_0(\hat{\beta})}.$$

Com base em tal resultado, uma correção ao coeficiente R_{AN}^2 foi apresentada por Hagle e Mitchell (1992) e também por Veall e Zimmermann (1994). Sua expressão é dada por:

$$R_{HM}^2 = \frac{R_{AN}^2}{\text{LS}(R_{AN}^2)} = \frac{\left(\frac{D_0 - D}{n + D_0 - D}\right)}{\left(\frac{D_0}{n + D_0}\right)}. \quad (4.18)$$

Mediante estudos de simulação, ambos os trabalhos defenderam R_{HM}^2 como uma das melhores opções de coeficiente de determinação, no sentido de apresentar grande similaridade de resultados em relação a R_{MQ}^2 , quando se assume a existência de uma variável resposta contínua, não-observada, por trás da resposta categorizada. Segundo os autores, a proximidade com R_{MQ}^2 seria o principal critério para escolha de um coeficiente de determinação apropriado para a regressão logística.

Por outro lado, Menard (2000) apresentou críticas a este coeficiente. O autor argumentou que a medida de Aldrich e Nelson (1984) já sofre por não apresentar uma interpretação intuitiva, enquanto medida de variabilidade explicada; de forma que a sua versão corrigida, R_{HM}^2 , propaga tal limitação de forma ainda mais marcante. Criticou, ainda, o fato de as simulações realizadas nos trabalhos anteriores serem baseadas na premissa de existência de uma variável resposta latente, com escala contínua, o que não necessariamente seria válido. Com esse argumento, respondeu às críticas de seus antecessores à razão de logaritmos das verossimilhanças (R_{MF}^2), apresentada na Seção 4.1.2, que seria superior em termos de interpretabilidade, além de atender aos critérios básicos de adequabilidade de um coeficiente de determinação, segundo proposição de Kvalseth (1985) descrita mais adiante, na Seção 4.3.

4.1.4 Incremento Quadrático Médio por Observação (R_{MA}^2) e corrigido (R_N^2)

A proposta a seguir foi apresentada originalmente por Maddala (1983) e recomendada por outros autores nos anos seguintes, com destaque para Cox e Snell (1989) e Magee (1990). Observa-se, inicialmente, que a razão de verossimilhanças pode ser escrita como o produto das razões de

verossimilhanças de cada observação, ou seja:

$$\frac{L_M(\hat{\beta})}{L_0(\hat{\beta})} = \frac{\prod_{i=1}^n L_{M_i}(\hat{\beta})}{\prod_{i=1}^n L_{0_i}(\hat{\beta})} = \prod_{i=1}^n \frac{L_{M_i}(\hat{\beta})}{L_{0_i}(\hat{\beta})}.$$

Cada termo do último produtório representa o incremento de informação relevante trazida pela i -ésima observação no modelo completo, em relação ao modelo basal, para a variável resposta. Elevando a expressão anterior a $1/n$, chega-se à média geométrica de incremento por observação, tal como ressaltado por [Cox e Wermuth \(1992\)](#):

$$G = \left(\prod_{i=1}^n \frac{L_{M_i}(\hat{\beta})}{L_{0_i}(\hat{\beta})} \right)^{1/n}.$$

Com base em tal resultado, o coeficiente proposto por [Maddala \(1983\)](#) foi:

$$R_{MA}^2 = 1 - G^{-2} = 1 - \left(\frac{L_0(\hat{\beta})}{L_M(\hat{\beta})} \right)^{2/n}. \quad (4.19)$$

Autores como [Magee \(1990\)](#) e [Ryan \(2008\)](#) também apresentaram tal coeficiente, ressaltando como principal vantagem o fato de que ele é algebricamente idêntico ao tradicional R_{MQ}^2 , quando calculado para um modelo de regressão linear. Sua utilização para o modelo de regressão logística representaria, portanto, uma generalização natural. A mesma defesa foi realizada por [Tjur \(2009\)](#), segundo o qual este seria o único coeficiente de determinação razoável, entre todos os que são baseados na verossimilhança.

Já [Menard \(2000\)](#) apresentou uma opinião divergente. O autor ressaltou que o erro quadrático médio geométrico não é a medida que se busca minimizar no processo de estimação por máxima verossimilhança, o que desfavoreceria uma interpretação intuitiva de R_{MA}^2 . [Tjur \(2009\)](#) concordou parcialmente com tal afirmação, fato que lhe motivou a propor uma opção mais intuitiva de coeficiente, conforme será visto mais adiante, na Seção 4.2.4.

Alguns anos mais tarde, [Nagelkerke \(1991\)](#) propôs um coeficiente alternativo, baseando-se na advertência de que o coeficiente de [Maddala \(1983\)](#) nunca assumiria valor igual a 1. Considerando que o valor máximo da verossimilhança de um modelo de regressão logística é igual a 1, o valor limitante superior do coeficiente R_{MA}^2 seria igual a:

$$LS(R_{MA}^2) = 1 - L_0(\hat{\beta})^{2/n}.$$

Consequentemente, o coeficiente corrigido de Nagelkerke (1991) é dado por:

$$R_N^2 = \frac{R_{MA}^2}{\text{LS}(R_{MA}^2)} = \frac{1 - \left(\frac{L_0(\hat{\beta})}{L_M(\hat{\beta})} \right)^{2/n}}{1 - L_0(\hat{\beta})^{2/n}}. \quad (4.20)$$

Nagelkerke (1991) considerou o coeficiente de Maddala (1983) como a melhor alternativa de coeficiente de determinação para modelos não-lineares, exceto pelo fato de que necessitava de uma correção em seu intervalo de variação de valores. Entretanto, a correção que o autor apresentou não foi considerada suficiente por Mittlböck e Schemper (1996), que julgou tanto R_{MA}^2 quanto R_N^2 como coeficientes sem interpretação intuitiva. A mesma ressalva foi apresentada por Menard (2000), ainda com maior veemência.

4.2 Medidas Não Baseadas na Verossimilhança

Diversos outros coeficientes de determinação foram introduzidos pela literatura, ao longo das últimas décadas, cujas expressões não envolvem a função de verossimilhança ou seu logaritmo. As principais propostas são descritas a seguir.

4.2.1 Correlações de Pearson (R_P^2), Spearman (R_S^2) e Kendall (R_K^2)

Uma classe de coeficientes de natural interesse é fundamentada na correlação entre os valores observados da resposta dicotômica ($y_i, 1 \leq i \leq n$) e as probabilidades de sucesso preditas ($\hat{\pi}_1(\mathbf{x}_i), 1 \leq i \leq n$). Por sua natureza, tais coeficientes não se aplicam à regressão logística com resposta multinomial.

Um dos trabalhos de destaque na análise de tais coeficientes é o de Mittlböck e Schemper (1996). Estabelecendo uma correspondência com R_{MQ}^2 , que coincide com o quadrado do coeficiente de correlação de Pearson na regressão linear múltipla, a seguinte medida foi apresentada:

$$R_P^2 = \frac{\left[\sum_{i=1}^n (y_i - \bar{y}) [\hat{\pi}_1(\mathbf{x}_i) - \bar{\pi}_1] \right]^2}{\left[\sum_{i=1}^n (y_i - \bar{y})^2 \right] \cdot \left[\sum_{i=1}^n [\hat{\pi}_1(\mathbf{x}_i) - \bar{\pi}_1]^2 \right]}, \quad (4.21)$$

em que $\bar{\pi}_1 = \bar{y} = \sum_{i=1}^n y_i / n$.

Mittlböck e Schemper (1996) consideraram R_P^2 uma das opções mais razoáveis de coeficiente

de determinação para a regressão logística, devido ao fato de atender a todos os critérios de adequabilidade propostos pelos autores, ao contrário das demais alternativas analisadas em seu trabalho. Tais critérios serão apresentados mais adiante, na Seção 4.3. Além disso, Hosmer *et al.* (2013) também recomendaram a utilização do coeficiente R_P^2 .

Outras duas medidas, análogas aos coeficientes de correlação de Spearman e Kendall, podem ser deduzidas. A primeira, respectivamente, é dada por:

$$R_S^2 = \frac{\left[\sum_{i=1}^n [R(y_i) - \bar{R}] [R(\hat{\pi}_1(\mathbf{x}_i)) - \bar{R}] \right]^2}{\left[\sum_{i=1}^n [R(y_i) - \bar{R}]^2 \right] \cdot \left[\sum_{i=1}^n [R(\hat{\pi}_1(\mathbf{x}_i)) - \bar{R}]^2 \right]}, \quad \bar{R} = (n+1)/2, \quad (4.22)$$

ou seja, tem estrutura análoga ao coeficiente de correlação linear de Pearson, substituindo os valores originais de y_i e $\hat{\pi}_1(\mathbf{x}_i)$ pelos seus postos, denotados pela função $R(\cdot)$.

O coeficiente R_S^2 de Spearman é criticado por Mittlböck e Schemper (1996), que argumentaram que medidas de associação ordinal não são coerentes com a natureza da regressão logística. Como exemplo, os autores mencionaram a seguinte situação: se uma variável explicativa “fraca” é adicionada a um modelo que já possui uma variável explicativa “forte”, sendo ambas não correlacionadas, entende-se que o total de variabilidade explicada deveria sofrer uma alteração discreta; entretanto, o coeficiente R_S^2 pode decrescer de forma considerável nessa situação, devido a impactos significativos nos rankings.

Já o coeficiente de determinação análogo à correlação τ_b de Kendall, (mais apropriada que τ_a , devido à grande quantidade de empates em $\mathbf{y}$), é dada por:

$$R_K^2 = \frac{\left[\sum_{i < j} \text{sgn}(y_j - y_i) \cdot \text{sgn}[\hat{\pi}_1(\mathbf{x}_j) - \hat{\pi}_1(\mathbf{x}_i)] \right]^2}{\sum_{i < j} \text{sgn}^2(y_j - y_i) \sum_{i < j} \text{sgn}^2[\hat{\pi}_1(\mathbf{x}_j) - \hat{\pi}_1(\mathbf{x}_i)]}, \quad \text{sgn}(z) = \begin{cases} 1 & \text{se } z > 0, \\ 0 & \text{se } z = 0, \\ -1 & \text{se } z < 0. \end{cases} \quad (4.23)$$

Mittlböck e Schemper (1996) também desencorajaram a utilização de tal coeficiente, por não considerarem clara a sua interpretação, enquanto medida de variabilidade explicada. Além disso, destacaram que tal coeficiente não atingiria o valor de 1, mesmo em casos de perfeita explicação por meio das variáveis independentes. Por outro lado, tal coeficiente foi recomendado por Korn e Simon (1990), que alegaram que R_K^2 seria uma alternativa viável para mensurar a variabilidade explicada, em particular no contexto de análise de sobrevivência.

Existem diversas variações do coeficiente de correlação de Kendall, que não serão discutidas neste trabalho. Elas estão sujeitas às mesmas limitações de R_K^2 , pois diferem apenas nos tipos de penalizações utilizadas para casos de empates. Mais detalhes podem ser encontrados nos trabalhos de [Goodman e Kruskal \(1954\)](#) e [Somers \(1962\)](#).

4.2.2 Coeficientes de Haberman: Entropia (R_H^2) e Concentração (R_G^2)

Em suas análises sobre dispersão de variáveis categóricas em modelos log-lineares, [Haberman \(1982\)](#) propôs uma medida genérica de redução proporcional do erro:

$$\rho = \frac{\sum_{i=1}^n D(y_i) - \sum_{i=1}^n D(y_i|\mathbf{x}_i)}{\sum_{i=1}^n D(y_i)} = 1 - \frac{\sum_{i=1}^n D(y_i|\mathbf{x}_i)}{\sum_{i=1}^n D(y_i)},$$

em que $D(y_i|\mathbf{x}_i)$ e $D(y_i)$ representam medidas de dispersão dos valores y_i , condicional ou não aos valores $\mathbf{x}_i$ das variáveis explicativas consideradas, para $1 \leq i \leq n$.

Não se deve confundir a função $D(\cdot)$ com a função desvio, apresentada anteriormente em termos da verossimilhança do modelo, ainda que o racional envolvido seja análogo – vide derivação do coeficiente R_{MF}^2 , dado pela Equação (4.15).

Uma primeira possibilidade consiste em definir a função $D(\cdot)$ a partir da medida de entropia de [Shannon \(1948\)](#), utilizada justamente para avaliar a dispersão de uma variável aleatória de natureza categórica. Esta medida é definida como:

$$H_Y = - \sum_{j=1}^k p_j \log p_j,$$

em que p_j corresponde à probabilidade de a variável aleatória Y assumir o valor associado à j -ésima categoria, com $1 \leq j \leq k$. Para um valor fixado de k , quanto maior a variabilidade entre os valores p_j , mais próximo de zero tende a ser o valor de H_Y .

Logo, pode-se definir os seguintes valores, respectivamente, para as medidas de dispersão condicional e não condicional associadas a cada observação, na regressão logística multinomial:

$$\begin{aligned} D(y_i|\mathbf{x}_i) &= - \sum_{j=1}^k \hat{\pi}_j(\mathbf{x}_i) \log [\hat{\pi}_j(\mathbf{x}_i)], \\ D(y_i) &= - \sum_{j=1}^k \bar{\pi}_j \log (\bar{\pi}_j), \end{aligned}$$

em que $\bar{\pi}_j = \sum_{i=1}^n \hat{\pi}_j(\mathbf{x}_i)/n$.

Consequentemente, tem-se a seguinte proposta de coeficiente de determinação, apresentada por [Haberman \(1982\)](#) e baseada na medida de entropia:

$$R_H^2 = 1 - \frac{\sum_{i=1}^n \sum_{j=1}^k \hat{\pi}_j(\mathbf{x}_i) \log [\hat{\pi}_j(\mathbf{x}_i)]}{n \sum_{j=1}^k \bar{\pi}_j \log (\bar{\pi}_j)}. \quad (4.24)$$

É fácil notar a semelhança estrutural entre as fórmulas do coeficiente R_H^2 (Equação (4.24)) e o coeficiente R_{MF}^2 , proposto por [McFadden \(1974\)](#) (Equação (4.15)). Conforme apontado por [Mittlböck e Schemper \(1996\)](#), ambos os coeficientes são numericamente idênticos.

[Haberman \(1982\)](#) discutiu outra medida, estruturalmente análoga ao R_H^2 , mas pautada pelo índice de concentração de [Gini \(1912\)](#), também utilizado para mensurar dispersão de variáveis categóricas. Tal índice é dado por:

$$G_Y = 1 - \sum_{j=1}^k p_j^2,$$

em que p_j corresponde à probabilidade de a variável aleatória Y assumir o valor associado à j -ésima categoria, com $1 \leq j \leq k$. Tem-se os seguintes valores de dispersão condicional e não condicional associados a cada observação, respectivamente:

$$\begin{aligned} D(y_i|\mathbf{x}_i) &= 1 - \sum_{j=1}^k \hat{\pi}_j(\mathbf{x}_i)^2, \\ D(y_i) &= 1 - \sum_{j=1}^k \bar{\pi}_j^2. \end{aligned}$$

Com isso, chega-se ao seguinte coeficiente de determinação, apresentado por [Haberman \(1982\)](#) e discutido também por [Mittlböck e Schemper \(1996\)](#):

$$R_G^2 = 1 - \frac{n - \sum_{i=1}^n \sum_{j=1}^k \hat{\pi}_j(\mathbf{x}_i)^2}{n - n \sum_{j=1}^k \bar{\pi}_j^2}. \quad (4.25)$$

No caso de resposta dicotômica, a expressão de R_G^2 pode ser simplificada como:

$$R_G^2 = 1 - \frac{\sum_{i=1}^n \hat{\pi}_1(\mathbf{x}_i)[1 - \hat{\pi}_1(\mathbf{x}_i)]}{n\hat{\pi}_1(1 - \hat{\pi}_1)}.$$

Segundo [Mittlböck e Schemper \(1996\)](#), R_G^2 é consistente com todos os critérios de adequabilidade (Seção 4.3), podendo ser considerado uma alternativa plausível de coeficiente de determinação.

4.2.3 Coeficiente de Ganho Total Padronizado (R_{BG}^2)

Com ênfase em uma interpretação visual da capacidade explanatória do modelo de regressão logística, [Bura e Gastwirth \(2001\)](#) propuseram um novo coeficiente de determinação. Denominado como “ganho total padronizado”, ele se baseia na área da região compreendida entre a curva da regressão logística simples e a reta $y = \bar{\pi}_1$, correspondente à proporção observada de sucessos na amostra. Segundo os autores, quanto maior o valor dessa área, maior o grau de variabilidade explicada.

A expressão do ganho total, para o caso de uma única variável explicativa, é definida como uma integral em relação à distribuição acumulada da variável X , e é dada por:

$$TG = \begin{cases} 2 \int_{-\infty}^{x^*} [\bar{\pi}_1 - \pi_1(x)] dF(x) & \text{se } \beta_1 > 0, \\ 2 \int_{x^*}^{\infty} [\pi_1(x) - \bar{\pi}_1] dF(x) & \text{se } \beta_1 < 0, \end{cases}$$

em que $\pi_1(x) = P(Y = 1|X = x)$ e x^* representa a abscissa do ponto de intersecção da curva $y = \pi_1(x)$ com a reta $y = \bar{\pi}_1$.

Devido à simetria da curva da regressão logística, pode-se generalizar a expressão anterior de TG , independentemente do sinal de β_1 :

$$TG = 2 \left| \int_{-\infty}^{x^*} [\bar{\pi}_1 - \pi_1(x)] dF(x) \right|.$$

Já o ganho total padronizado, ainda no cenário de uma variável explicativa única, resulta da divisão do ganho total pelo seu limitante superior, que [Bura e Gastwirth \(2001\)](#) demonstraram ser igual a $2\bar{\pi}_1(1 - \bar{\pi}_1)$. Dessa forma, obtém-se um coeficiente que varia no intervalo de 0 a 1:

$$TG_{std} = \frac{TG}{2\bar{\pi}_1(1 - \bar{\pi}_1)}.$$

Exemplos: *binary regression quantile plots*

Figura 4.1: Gráficos ilustrativos da área entre a curva de probabilidades condicionais de sucesso e a reta da proporção geral de sucesso, na regressão logística simples, representando uma gradual diminuição de explicação da resposta por meio da variável explicativa (do exemplo 1 ao 4).

A Figura 4.1 exibe alguns exemplos de curvas de probabilidades condicionais de sucesso e retas da proporção geral de sucesso, representadas em relação a F_X , que é a distribuição acumulada da variável explicativa X . Este formato gráfico é denominado por Bura e Gastwirth (2001) como *binary regression quantile plot*. Ao ajustar um modelo em particular, a distribuição real F_X pode ser estimada por meio de sua correspondente distribuição empírica $\hat{F}_X$. Segundo os autores, a padronização do eixo horizontal em escala unitária propicia a facilidade de interpretação e comparação de diferentes modelos de regressão logística.

Definindo $t = F(x)$, e conseqüentemente $t^* = F(x^*)$, tem-se a seguinte expressão alternativa para o coeficiente de ganho total:

$$TG = 2 \int_0^{t^*} [\bar{\pi}_1 - \pi_1(F^{-1}(t))] dt.$$

Para o cálculo numérico da integral do coeficiente de ganho total, os autores derivaram um estimador e demonstraram a sua consistência. Sua expressão envolve a proporção geral de sucesso na amostra, as probabilidades previstas para cada observação e os valores estimados dos parâmetros β_0 e β_1 pelo modelo:

$$\widehat{TG} = 2 \left[\bar{\pi}_1 \cdot \frac{1}{n} \sum_{i=1}^n I(x_i \leq \hat{x}^*) - \frac{1}{n} \sum_{x_i \leq \hat{x}^*} \hat{\pi}_1(x_i) \right], \quad \text{em que } \hat{x}^* = \frac{\log \frac{\bar{\pi}_1}{1 - \bar{\pi}_1} - \hat{\alpha}}{\hat{\beta}},$$

em que $I(\cdot)$ é a função indicadora, assumindo valores 0 ou 1.

Por fim, o candidato proposto como coeficiente de determinação surge a partir da padronização da expressão anterior, para o intervalo de 0 a 1:

$$R_{BG}^2 = \frac{\widehat{TG}}{2\bar{\pi}_1(1 - \bar{\pi}_1)}. \quad (4.26)$$

Já no caso da regressão logística múltipla, [Bura e Gastwirth \(2001\)](#) mencionaram, de forma sucinta, uma generalização do caso mais simples. Pode-se considerar a combinação linear $\beta^T \mathbf{X}$ como um novo preditor único, de forma que cálculos análogos aos anteriores sejam realizados a partir de seus escores, em vez da variável X do caso da regressão logística simples. Gráficos análogos aos da Figura 4.1 também podem ser construídos.

Os autores não realizaram comparações numéricas do coeficiente R_{BG}^2 em relação às alternativas que já haviam sido apresentadas na literatura, bem como não argumentaram a respeito de sua possível superioridade. Destacaram, apenas, a sua facilidade de interpretação, por conta do apelo gráfico envolvido. Não houve trabalhos de destaque posteriores em que tenham sido realizadas comparações deste coeficiente com as propostas mais tradicionais.

4.2.4 Coeficiente de Discriminação de Tjur (R_T^2)

Na contramão de todos os trabalhos anteriores, [Tjur \(2009\)](#) recomendou um coeficiente novo e bastante simplificado para mensurar a variabilidade explicada, especificamente para o caso da regressão logística com resposta dicotômica.

O cálculo recomendado pelo autor consiste, meramente, na diferença entre as médias das probabilidades de sucesso previstas de dois grupos: as observações que são sucessos e as observações que são fracassos. Quanto maior essa diferença, maior seria a explicação da variabilidade da resposta. Ou seja:

Exemplo 1 **Exemplo 2** **Exemplo 3**

Figura 4.2: Histogramas ilustrativos da distribuição das probabilidades preditas de sucesso ($\hat{\pi}_1(x)$), a depender do valor da resposta real, representando um gradual aumento de explicação da resposta por meio da variável explicativa (do exemplo 1 ao 3), segundo o coeficiente R_T^2 .

$$R_T^2 = \overline{\hat{\pi}_1(\mathbf{x}|y=1)} - \overline{\hat{\pi}_1(\mathbf{x}|y=0)}, \quad (4.27)$$

em que $\overline{\hat{\pi}_1(\mathbf{x}|y=k)}$ representa a média das probabilidades de sucesso preditas pelo modelo, para observações com valor de resposta observada igual a k , com $k = 0$ ou 1 .

Intuitivamente, R_T^2 se baseia na ideia de que um modelo tem melhor capacidade explanatória à medida que as probabilidades estimadas de sucesso ($\pi_1(\mathbf{x}_i)$) sejam mais altas para observações que, de fato, correspondam a sucessos; bem como sejam mais baixas, para observações que correspondam a fracassos. Este racional pode ser verificado graficamente, por meio de histogramas comparativos das distribuições das probabilidades preditas (Figura 4.2).

Para justificar tal proposição, define-se três outros coeficientes de determinação hipotéticos, conceitualmente análogos ao R^2 da regressão linear, tomando como base as somas de quadrados apresentadas na Equação (2.7). São os seguintes:

$$R_{mod}^2 = \frac{SQR}{SQT} = \frac{\sum_{i=1}^n [\hat{\pi}_1(\mathbf{x}_i) - \bar{y}]^2}{\sum_{i=1}^n (y_i - \bar{y})^2},$$

$$R_{res}^2 = 1 - \frac{SQE}{SQT} = 1 - \frac{\sum_{i=1}^n [y_i - \hat{\pi}_1(\mathbf{x}_i)]^2}{\sum_{i=1}^n (y_i - \bar{y})^2},$$

$$R_{cor}^2 = \frac{\left[\sum_{i=1}^n (y_i - \bar{y}) [\hat{\pi}_1(\mathbf{x}_i) - \bar{\pi}_1] \right]^2}{\left[\sum_{i=1}^n (y_i - \bar{y})^2 \right] \cdot \left[\sum_{i=1}^n [\hat{\pi}_1(\mathbf{x}_i) - \bar{\pi}_1]^2 \right]},$$

sendo que R_{cor}^2 é idêntico ao coeficiente R_P^2 , apresentado na Seção 4.2.1.

Tjur (2009) ressaltou que tais coeficientes não são algebricamente idênticos entre si na regressão logística, tal como são na regressão linear, devido à não ortogonalidade entre os vetores de valores ajustados e de resíduos. Entretanto, demonstrou que eles são assintoticamente equivalentes, ou seja, assumem o mesmo valor quando $n \rightarrow \infty$.

O coeficiente R_T^2 apresenta estreita relação com os três coeficientes anteriores. As seguintes igualdades foram demonstradas por Tjur (2009):

$$R_T^2 = \frac{1}{2}(R_{mod}^2 + R_{res}^2),$$

$$R_T^2 = \sqrt{R_{mod}^2 R_{cor}^2}.$$

Dessa forma, segundo o autor, R_T^2 seria uma medida numericamente compatível com o conceito de coeficiente de determinação da regressão linear. O autor também defendeu, com particular ênfase, a facilidade de interpretação desse coeficiente, em relação às propostas dos trabalhos predecessores.

4.3 Estratégia para Avaliação dos Coeficientes

Para pautar a escolha de um coeficiente de determinação do tipo R^2 pertinente, que sirva como alternativa ao coeficiente tradicional, oito critérios de adequabilidade foram propostos por Kvalseth (1985). Tais critérios foram mencionados em diversos trabalhos posteriores, para fins de discussão e comparação de outros coeficientes (Magee, 1990; Menard, 2000; Tjur, 2009), e são apresentados a seguir.

Critérios de Kvalseth

1. O coeficiente deve ser útil para medir a qualidade de ajuste do modelo, bem como ter uma interpretação razoável.
2. O coeficiente deve ser independente das unidades de medida das variáveis explicativas do modelo.
3. O coeficiente deve assumir valores entre 0 e 1, sendo o limite inferior correspondente à total falta de ajuste e o limite superior correspondente a um ajuste perfeito.
4. O coeficiente deve ser aplicável a qualquer tipo de modelo, com variáveis explicativas aleatórias ou não, independentemente das propriedades estatísticas das variáveis do modelo (incluindo os resíduos).
5. O coeficiente não deve ter a sua aplicabilidade restrita a alguma técnica específica de ajuste.
6. O coeficiente deve fornecer valores comparáveis entre si, quando diferentes modelos forem ajustados para o mesmo conjunto de dados.
7. O coeficiente deve fornecer valores que sejam compatíveis com os obtidos por meio de outras medidas de qualidade de ajuste (ex.: erro padrão de previsão, desvio padrão residual).

8. O coeficiente deve ponderar igualmente resíduos com valores positivos e negativos.

Vale observar que as propriedades de Kvalseth (1985) foram apresentadas num contexto em que o coeficiente R^2 era considerado, primordialmente, como uma ferramenta para mensurar a qualidade geral de ajuste do modelo. Dessa forma, não tangem, especificamente, o aspecto da variabilidade explicada.

Outra formulação de critérios de adequabilidade foi proposta por Mittlböck e Schemper (1996), de forma mais enxuta e adaptada ao contexto específico da regressão logística. Apresenta-se tal proposição a seguir.

Critérios de Mittlböck e Schemper

1. O coeficiente deve ter interpretação clara e intuitiva.
2. O coeficiente não pode ser influenciado por transformações lineares aplicadas sobre as variáveis explicativas, mas apenas por transformações monótonas não lineares, em coerência com a natureza da regressão logística.
3. O coeficiente deve assumir valores entre 0 e 1, sendo o limite inferior correspondente à total falta de previsibilidade e o limite superior correspondente à previsibilidade completa da resposta, dadas as variáveis explicativas.
4. Caso os dados possam ser analisados por meio de um modelo de regressão linear, o coeficiente deve apresentar valor próximo ao do R^2 tradicional.

Os critérios apresentados por Mittlböck e Schemper (1996) diferem dos critérios de Kvalseth (1985) por não requererem que o coeficiente R^2 sirva para medir a qualidade geral de ajuste do modelo, mas apenas o grau de variabilidade explicada. Dessa forma, o espírito geral de tal formulação está mais alinhado com o objetivo deste trabalho.

A diferença de interpretação fica clara quando se compara os critérios de número 3 de ambos os trabalhos. Os autores concordam no que diz respeito ao intervalo de variação do coeficiente de determinação, que está entre 0 e 1; entretanto, para Kvalseth (1985), quanto mais próximo de 1 está o valor de R^2 , maior a qualidade de ajuste do modelo; já para Mittlböck e Schemper (1996), quanto maior o coeficiente de determinação, maior a explicação da resposta a partir das variáveis explicativas.

Por outro lado, Mittlböck e Schemper (1996) deram maior ênfase ao aspecto da facilidade de interpretação dos candidatos a coeficiente de determinação, o que os motivou a formular os

critérios 1 e 4. Este último critério, em particular, estabelece o que os autores denominaram como “consistência numérica” entre a regressão logística e a regressão linear, sempre que fizer sentido avaliar os dados por meio de ambas as técnicas. Em seu trabalho de simulação, assumiram como balizador o valor do R^2 tradicional, com o qual compararam diversos outros coeficientes alternativos calculados para a regressão logística. Conforme observação feita por [Cox e Wermuth \(1992\)](#), tal estratégia de comparação apresenta limitações importantes, já discutidas na Seção 2.4.

Diversos outros estudos de simulação foram conduzidos, anteriormente, sob a ótica da existência de uma variável resposta latente, de natureza contínua, por trás da resposta categorizada. Seguindo essa linha, merecem destaque os trabalhos de [Hagle e Mitchell \(1992\)](#), [Veall e Zimmermann \(1994\)](#) e [Cameron e Windmeijer \(1997\)](#), que lançaram mão de tal premissa e buscaram, para a regressão logística, uma medida que demonstrasse maior proximidade numérica em relação ao coeficiente R^2 usual, obtido pelo modelo de regressão linear ajustado para a variável resposta latente.

Tendo em vista o objetivo deste trabalho, mencionado no Capítulo 1, a tática de considerar como necessária a existência de uma variável resposta contínua não se faz apropriada para todas as ocasiões. Com frequência, as variáveis resposta de interesse são verdadeiramente categóricas nominais, não sendo possível estabelecer analogia com uma resposta contínua latente. Por conseguinte, o critério 4 apresentado por [Mittlböck e Schemper \(1996\)](#) não se aplica ao presente escopo, fazendo-se necessário estabelecer alguma nova estratégia numérica para decidir quais propostas de coeficientes correspondem melhor ao objetivo em questão. Ou seja, deve-se avaliar qual coeficiente de determinação mensura de forma mais correta o grau de variabilidade explicada, tomando como base alguma referência que não seja o R^2 da estimação por mínimos quadrados.

4.4 Considerações Finais do Capítulo

Neste capítulo, discorreu-se a respeito de diversas propostas de coeficientes de determinação, para modelos de regressão logística múltipla. Diversas delas são baseadas na verossimilhança, principal fonte para avaliação de tais modelos, mas medidas alternativas também foram apresentadas. Discutiu-se, ainda, a respeito dos critérios de adequabilidade que devem ser levados em consideração na escolha de um coeficiente de determinação para a regressão logística.

Observa-se grande discordância na literatura sobre qual coeficiente é o mais adequado. Cada um dos autores faz avaliações sob diferentes prismas, ora buscando analogias algébricas em relação ao caso de mínimos quadrados, mediante a suposição de existência de uma variável resposta latente de natureza contínua ([Hagle e Mitchell, 1992](#)); outrora priorizando o aspecto da interpretabilidade do coeficiente ([Menard, 2000](#); [Tjur, 2009](#)).

Adicionalmente, nota-se que a maior parte dos trabalhos anteriores trataram, especialmente, da regressão logística com resposta dicotômica. O caso multinomial foi bem menos explorado, tanto no que diz respeito à avaliação e comparação numérica dos coeficientes do tipo R^2 já existentes, quanto à busca por outras alternativas apropriadas. Faz-se necessário um aprofundamento maior de pesquisas para este contexto.

Capítulo 5

Simulação

Realizou-se um estudo de simulação, com o objetivo de avaliar as propostas de coeficientes de determinação do tipo R^2 apresentadas no Capítulo 4. Considerou-se diversos cenários de tamanho amostral e distribuição das variáveis explicativas.

Os coeficientes foram comparados numericamente entre si, investigando-se relações que favorecessem um julgamento a respeito de sua adequabilidade. Utilizou-se a noção de *sobreposição* para auxiliar a produzir diferentes graus hipotéticos de variabilidade explicada, servindo como referência adicional para avaliação dos coeficientes. Além disso, foram levados em conta os comportamentos das propostas de R^2 em relação a alguns índices de acurácia preditiva (Capítulo 3).

Todo o processo de simulação e as análises de resultados foram realizados por meio da linguagem R, versão 3.6.2, para sistema operacional Windows de 64 bits.

5.1 Esquema de Simulação

A simulação realizada baseou-se no modelo de regressão logística com resposta dicotômica e uma única variável explicativa aleatória X , tal como considerado em trabalhos de simulação anteriores (Hu *et al.*, 2006; Mittlböck e Schemper, 1996; Veall e Zimmermann, 1994). Conforme argumento apresentado por Mittlböck e Schemper (1996), entende-se que todos os diferentes graus de explicabilidade proporcionados por $p - 1$ variáveis explicativas são devidamente representados por meio dos comportamentos extremos de uma variável única na regressão logística simples. Neste cenário, a formulação do modelo apresentado na Equação (2.1) reduz-se a:

$$Y_i \sim \text{Ber}(\pi_1(x_i)), \quad \text{sendo } \log\left(\frac{\pi_1(x_i)}{1 - \pi_1(x_i)}\right) = \beta_0 + \beta_1 x_i, \quad (5.1)$$

para $1 \leq i \leq n$.

De toda forma, avaliar-se-á também o efeito exercido sobre os coeficientes do tipo R^2 quando se inclui uma segunda variável explicativa, potencialmente não correlacionada com a variável resposta e nem com a primeira variável explicativa considerada. Neste caso, a formulação do modelo é dada por:

$$Y_i \sim \text{Ber}(\pi_1(\mathbf{x}_i)), \quad \text{sendo } \log \left(\frac{\pi_1(\mathbf{x}_i)}{1 - \pi_1(\mathbf{x}_i)} \right) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i}, \quad (5.2)$$

para $1 \leq i \leq n$.

Já o modelo basal, sem variáveis explicativas, é dado pela expressão:

$$Y_i \sim \text{Ber}(\bar{\pi}_1), \quad \text{sendo } \log \left(\frac{\bar{\pi}_1}{1 - \bar{\pi}_1} \right) = \beta_0, \quad (5.3)$$

para $1 \leq i \leq n$, em que $\bar{\pi}_1 = \sum_i y_i / n$.

Adiante, na Seção 5.1.2, será introduzido o conceito intuitivo de sobreposição, por meio do qual garantir-se-á que, mesmo considerando apenas uma variável explicativa, todos os cenários podem ser devidamente contemplados; ou seja, desde a contribuição nula até a contribuição máxima possível para explicar a resposta de interesse.

5.1.1 Fatores de variação

Os seguintes fatores foram considerados para variação no processo de simulação:

- **Tamanho amostral** (n). Considerou-se os seguintes valores:

- i) $n = 100$
- ii) $n = 500$
- iii) $n = 1.000$
- iv) $n = 5.000$

Não foram testadas amostras com menos de 100 observações, tendo em vista que tornam inviável a modelagem de eventos de natureza mais rara (por exemplo, $\bar{\pi}_1 \approx 1\%$), para os quais também se tem interesse em estudar o comportamento das propostas de R^2 . Por outro lado, tamanhos amostrais maiores que 5.000 foram testados, mas não se observou diferenças relevantes nos resultados em relação ao cenário com $n = 5.000$, tanto na análise descritiva

univariada de cada R^2 , quanto na avaliação do comportamento desses coeficientes *versus* outras medidas.

- **Distribuição da variável explicativa (X).** Considerou-se os seguintes cenários, a fim de abranger distribuições contínuas com alguns diferentes padrões de assimetria e variabilidade, bem como uma distribuição discreta com duas categorias:
 - i) $X \sim \mathcal{N}(0,1)$ - distribuição normal padrão
 - ii) $X \sim \text{Gama}(n,1)$ - distribuição gama
 - iii) $X \sim \mathcal{U}(0,1)$ - distribuição uniforme
 - iv) $X \sim \text{Ber}(\pi_1)$ - distribuição de Bernoulli
- **Parâmetros do modelo (β_0 e β_1).** Para cada combinação de tamanho amostral (n) e distribuição da variável explicativa X , considerou-se valores β_0 e β_1 que fossem capazes de representar diversos graus de variabilidade explicada da resposta, bem como todos os patamares possíveis de proporção geral de sucesso (π_1). O conceito chave que se propõe para garantir tal diversidade é o de “sobreposição”, que será ilustrado na Seção 5.1.2.

5.1.2 Etapas da simulação

Foram avaliados 16 cenários, correspondentes ao cruzamento dos valores fixados para o tamanho amostral (em 4 níveis) com as distribuições fixadas de X (também em 4 níveis). Para cada cenário, as etapas da simulação foram repetidas 20.000 vezes, totalizando, dessa forma, um total de 320.000 iterações.

A quantidade de 20.000 repetições por cenário foi definida empiricamente, observando a convergência das estimativas dos valores médios dos 12 coeficientes de determinação avaliados. Para a maior parte dos coeficientes, 10.000 repetições já se mostraram suficientes para fornecer estabilidade na 3ª casa decimal das médias, que foi a precisão adotada para exibição e análise dos resultados neste trabalho. O total de repetições foi, então, sendo incrementado gradativamente até 20.000, para garantir a precisão desejada em todos os cenários de tamanho amostral e distribuição de X .

Cada iteração do processo de simulação foi composta pelos 17 passos descritos a seguir.

- 1) Sorteio de um valor inicial π_1 para a proporção geral de sucesso na amostra, de acordo com uma distribuição uniforme entre 0 e 0,5. Valores de π_1 superiores a 0,5 geram cenários análogos, caso se inverta as definições de sucesso e fracasso.

- 2) Geração de um vetor $\mathbf{x}$ com os valores observados da variável explicativa, de acordo com a distribuição fixada para X e o valor fixado de n , e ordenação descendente dos valores dentro deste vetor.

No caso em que X tem distribuição de Bernoulli, a probabilidade associada ao valor 1 foi assumida igual a $\bar{\pi}_1$, sorteado no passo 1, tal como para a variável resposta.

- 3) Criação de um vetor com os valores observados da variável resposta ($\mathbf{y}$), ordenado de forma que os primeiros n_1 valores sejam iguais a 1 e os demais $n - n_1$ valores sejam iguais a 0, sendo n_1 igual ao valor inteiro resultante do arredondamento de $n \cdot \bar{\pi}_1$. Em seguida, por conta de tal arredondamento, atualizou-se o valor da proporção geral de sucesso da amostra para $\bar{\pi}_1 = \bar{y} = \sum y_i / n$.

Neste momento, para cada iteração, um ajuste perfeito seria obtido para o modelo de regressão logística $\mathbf{y} \sim \mathbf{x}$, em vista da ordenação realizada nos vetores $\mathbf{x}$ e $\mathbf{y}$. Trata-se de uma questão denominada “separação completa” (Ryan, 2008, p. 320), que resulta na inexistência dos estimadores dos parâmetros β_0 e β_1 .

O cenário de separação completa corresponde à maior explicação possível da variabilidade da resposta, quando se espera que os coeficientes R^2 assumam valor igual a 1. Por outro lado, caso se altere de forma aleatória a ordem dos valores do vetor $\mathbf{y}$, espera-se gerar uma menor explicação da resposta por meio do vetor ordenado $\mathbf{x}$ e, conseqüentemente, menores valores de R^2 . No limite de uma completa aleatorização da ordem dos valores em $\mathbf{y}$, espera-se chegar a um cenário de total não explicação, em que R^2 deveria assumir, idealmente, o valor mínimo de zero.

Adaptando uma ideia concebida por Bura e Gastwirth (2001), propõe-se definir uma “proporção de sobreposição” (*overlap*) para aleatorizar a ordem dos elementos do vetor $\mathbf{y}$. Pode-se analisar o seguinte exemplo hipotético, com $X \sim \text{Ber}(0,4)$ e $n = 25$, em que há separação completa:

$$\begin{aligned}\mathbf{x} &= (1,1,1,1,1,1,1,1,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0), \\ \mathbf{y} &= (1,1,1,1,1,1,1,1,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0).\end{aligned}$$

Fazendo a menor intervenção possível sobre a ordem dos valores de $\mathbf{y}$, provendo alto grau de variabilidade explicada, mas evitando a separação completa, alternou-se aleatoriamente a ordem de apenas um valor 1 com um valor 0. Obteve-se:

$$\begin{aligned}\mathbf{x} &= (1,1,1,1,1,1,1,1,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0), \\ \mathbf{y} &= (1,1,0,1,1,1,1,1,1,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0).\end{aligned}$$

Já no cenário de total ausência de explicação, espera-se que as observações em que $y_i = 1$ estejam pulverizadas de forma homogênea na amostra. Dessa forma, entre as $(100 \cdot \bar{\pi}_1)\%$ primeiras observações de $\mathbf{y}$, espera-se que estejam concentrados $(100 \cdot \bar{\pi}_1)\%$ dos casos em que $y_i = 1$. Isso

corresponde a alternar a posição de $(100 \cdot (1 - \bar{\pi}_1))\%$ das observações que possuem $y_i = 1$ com observações que possuem $y_i = 0$, a partir do cenário da separação completa. Por exemplo:

$$\begin{aligned}\mathbf{x} &= (1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0), \\ \mathbf{y} &= (0, 1, 0, 0, 0, 1, 0, 1, 0, 1, 0, 0, 1, 0, 1, 0, 0, 1, 0, 1, 0, 1).\end{aligned}$$

Denotando-se por n_S a quantidade de observações com resposta $y = 1$ que foram alteradas de posição, tem-se que a proporção de sobreposição n_S/n pode variar dentro do seguinte intervalo, representando todos os cenários conceituais de variabilidade explicada:

$$0 \leq \frac{n_S}{n} \leq 1 - \bar{\pi}_1,$$

sendo $\bar{\pi}_1 = \sum y_i/n$.

Levando em conta o limitante superior de n_S , pode-se, por fim, definir a proporção de sobreposição padronizada p_S , que varia entre 0 e 1 (Equação (5.4)). Ela representa, de modo intuitivo, o quanto de aleatorização foi introduzida no vetor $\mathbf{y}$, partindo do cenário de separação completa, até a total aleatoriedade:

$$p_S = \frac{n_S}{n \cdot (1 - \bar{\pi}_1)}. \quad (5.4)$$

O conceito de sobreposição será utilizado nos passos 4 e 5 da simulação, a seguir.

- 4) Sorteio de um valor inteiro para n_S , correspondente à quantidade de valores 1 a serem alterados de posição com valores 0 no vetor $\mathbf{y}$, de acordo com uma distribuição uniforme discreta entre 0 e $\lfloor n \cdot (1 - \bar{\pi}_1) \rfloor$. Em seguida, cálculo da proporção de sobreposição, de acordo com a Equação (5.4).
- 5) Realização de uma troca aleatória (sobreposição) de valores do vetor $\mathbf{y}$, alternando a posição de n_S valores 1 com n_S valores 0.
- 6) Ajuste do modelo de regressão logística, segundo a Equação (5.2), considerando os vetores de valores observados para a variável resposta ($\mathbf{y}$) e a variável explicativa ($\mathbf{x}$), e consequente obtenção dos parâmetros estimados $\hat{\beta}_0$ e $\hat{\beta}_1$.

Tendo em vista que os valores 1 da variável resposta estão mais associados a valores altos da variável explicativa, por conta das ordenações descendentes realizadas nos passos 2 e 3, os valores obtidos para $\hat{\beta}_1$ são necessariamente maiores que zero, indicando associação positiva. O caso de associação negativa seria análogo a este.

É importante ressaltar que a aleatorização nas posições do vetor $\mathbf{y}$, realizada no passo 5, induz uma perturbação que não satisfaz a suposição teórica acerca das probabilidades $\pi_1(\mathbf{x}_i)$ no modelo

de regressão logística (Equação (5.1)). Tal procedimento foi adotado apenas como artifício para obtenção de estimativas dos parâmetros β_0 e β_1 neste passo 6, que serão adotados como valores de partida para o ajuste do modelo teórico nos passos seguintes. Dessa forma, garante-se a utilização de valores β_0 e β_1 que estejam em escalas coerentes com cada cenário de distribuição de X e tamanho amostral, permitindo a representatividade de diferentes graus de variabilidade explicada, entre os dois extremos bem definidos: da total ausência de explicabilidade até a separação completa.

Convém notar que valores pequenos da proporção de sobreposição p_S , estabelecida no passo 4, podem levar à não convergência na estimação dos parâmetros do modelo por máxima verossimilhança, em virtude de uma separação quase completa (Ryan, 2008). Isso ocorre, especialmente, quando o valor do tamanho amostral n e o valor sorteado $\bar{\pi}_1$ também são pequenos (por exemplo, para $n = 100$, $\bar{\pi}_1 = 0,01$ e $p_S = 0,10$). Para tratar a questão, cada iteração do algoritmo só foi finalizada quando houve convergência na estimação; caso contrário, os valores de $\bar{\pi}_1$ e n_S eram sorteados novamente, de forma sucessiva, até que o ajuste fosse bem-sucedido.

- 7) Adoção das estimativas $\hat{\beta}_0$ e $\hat{\beta}_1$ como sendo os valores iniciais dos parâmetros β_0 e β_1 , para o ajuste de um modelo teórico.
- 8) Cálculo do vetor de probabilidades de sucesso associadas a cada observação, em função de β_0 , β_1 e dos valores do vetor $\mathbf{x}$, segundo a Equação (5.1).
- 9) Geração de um novo vetor com valores da variável resposta ($\mathbf{y}$), de acordo com uma distribuição de Bernoulli, cuja probabilidade de sucesso, para cada observação, é definida de acordo com o vetor de probabilidades de sucesso, obtido no passo 8.
- 10) Cálculo da proporção de sobreposição final, atualizada de acordo com os dados simulados para a variável explicativa ($\mathbf{x}$) e para a variável resposta ($\mathbf{y}$). Será denotada por p'_S .

No passo 4, a proporção de sobreposição inicial havia sido controlada entre 0 e 1, por meio do sorteio de n_S dentro do intervalo de valores plausíveis. Agora, a proporção de sobreposição final é “observada”; ou seja, é decorrente da geração de $\mathbf{y}$ de forma aleatória, segundo a expressão do modelo de regressão logística. Por conta disso, houve iterações em que $p'_S > 1$. Em tais casos, a iteração foi interrompida e reiniciada a partir do passo 1.

- 11) Ajuste de dois modelos de regressão logística, considerando o vetor $\mathbf{y}$ de valores simulados para a variável resposta:
 - Modelo completo, considerando o vetor $\mathbf{x}$ de valores da variável explicativa, de acordo com a Equação (5.1). Em seguida, armazenamento dos parâmetros estimados $\hat{\beta}_0$ e $\hat{\beta}_1$;
 - Modelo basal, apenas com intercepto, de acordo com a Equação (5.3).

12) Cálculo dos valores de dois índices de acurácia preditiva:

- Índice de Gini (G), linearmente relacionado com AUC - Equação (3.5);
- Índice de Kolmogorov-Smirnoff (KS) - Equação (3.6).

Foram considerados apenas os índices de acurácia que não apresentam limitações importantes em sua interpretação, conforme críticas apresentadas na Seção 3.2.1; bem como não necessitam da predefinição de um ponto de corte, aspecto que varia a depender da importância que se deseja atribuir à sensibilidade e à especificidade, em cada contexto prático.

13) Cálculo dos valores das 12 propostas de coeficiente de determinação do tipo R^2 , discutidas ao longo desse trabalho, bem como uma proposta de coeficiente de determinação ajustado para o caso da regressão logística múltipla:

- R_{MQ}^2 de mínimos quadrados - Equação (2.9);
- R_{MF}^2 de McFadden (1974) - Equação (4.15);
- R_{MFadj}^2 de McFadden (1974) - Equação (4.16);
- R_{AN}^2 de Aldrich e Nelson (1984) - Equação (4.17);
- R_{HM}^2 de Hagle e Mitchell (1992) - Equação (4.18);
- R_{MA}^2 de Maddala (1983) - Equação (4.19);
- R_N^2 de Nagelkerke (1991) - Equação (4.20);
- R_P^2 da correlação linear de Pearson, de Hosmer *et al.* (2013) - Equação (4.21);
- R_S^2 da correlação linear de Spearman, de Mittlböck e Schemper (1996) - Equação (4.22);
- R_K^2 da correlação linear de Kendall, de Mittlböck e Schemper (1996) - Equação (4.23);
- R_G^2 baseado no índice de Gini, de Haberman (1982) - Equação (4.25);
- R_{BG}^2 de Bura e Gastwirth (2001) - Equação (4.26);
- R_T^2 de Tjur (2009) - Equação (4.27).

14) Ajuste de modelo de regressão logística, considerando uma variável explicativa transformada

$$X' = a + bX,$$

em que $a, b \sim \mathcal{U}(0,2)$, a fim de avaliar o 2º critério de adequabilidade de Mittlböck e Schemper (1996) (Seção 4.3).

15) Novo cálculo dos valores dos 12 coeficientes de determinação, para o modelo com a variável explicativa linearmente transformada.

- 16) Ajuste de modelo de regressão logística, acrescentando uma variável explicativa X_2 , segundo a Equação (5.2). A variável X_1 foi adotada como sendo a variável explicativa do ajuste principal, conduzido no passo 11. Já X_2 foi gerada de forma totalmente aleatória, segundo a mesma distribuição de X_1 , simulando ausência de correlação com a variável resposta.
- 17) Novo cálculo dos valores dos 12 coeficientes de determinação, bem como um coeficiente de determinação ajustado, para o modelo com duas variáveis explicativas.

5.2 Resultados

5.2.1 Análises iniciais

Observa-se, inicialmente, que os valores da proporção de sobreposição inicial (p_S), definida a partir da Equação (5.4), seguem, tal como esperado, uma distribuição aproximadamente uniforme no intervalo de 0 a 1 (Figura 5.1), mas com frequência menor para valores nos extremos da escala.

No caso de valores de p_S próximos de zero, o desbalanceamento na distribuição de frequências ocorre por conta de situações em que houve problema de separação completa ou quase completa durante o ajuste do modelo, o que levou ao descarte de tal valor e reinicialização da iteração. Este fato é observado, especialmente, para amostras menores (Figura A.1). Já para valores de p_S próximos de 1, a menor frequência se deve aos casos em que, após o ajuste do modelo teórico, a proporção de sobreposição final (p'_S) variou de forma a exceder o valor de 1. Nestes casos, conforme mencionado no passo 10, a iteração foi reiniciada e novos valores foram sorteados.

Figura 5.1: Histograma da distribuição dos valores da proporção de sobreposição inicial (p_S), considerando todos os cenários avaliados.

Já quanto à proporção de sobreposição final (p'_S), sua distribuição apresenta assimetria à

esquerda (Figura 5.2). Nota-se que as proporções “reais” de sobreposição são, na verdade, mais altas que as estipuladas inicialmente, por meio das quais se obteve estimativas para β_0 e β_1 e procedeu-se à “correção” da variável resposta, segundo o modelo teórico (Figura 5.3). Este comportamento é observado em todos os cenários de tamanho amostral e para quase todas as distribuições avaliadas para a variável explicativa, com exceção da Bernoulli, para a qual a proporção de sobreposição final mantém o comportamento uniforme inicial (Figura A.2).

Figura 5.2: Histograma da distribuição dos valores da proporção de sobreposição final (p'_S), considerando todos os cenários avaliados.

Figura 5.3: Gráfico de dispersão dos valores das proporções de sobreposição inicial (p_S) versus final (p'_S), considerando todos os cenários avaliados.

Vale ressaltar, também, que foram gerados diferentes valores da proporção geral de sucesso na amostra ($\bar{\pi}_1$), com distribuição aproximadamente uniforme, conforme esperado (Figura 5.4). Tal como se verificou para a proporção de sobreposição, valores de $\bar{\pi}_1$ próximos de zero possuem menor frequência associada, por conta de iterações em que houve separação completa ou quase completa. Isso ocorre de forma mais comum, novamente, para tamanhos amostrais menores (Figura A.3).

Figura 5.4: *Histograma da distribuição dos valores da proporção geral de sucesso ($\bar{\pi}_1$), considerando todos os cenários avaliados.*

Com relação aos valores dos parâmetros β_0 e β_1 , definidos a partir de estimativas $\hat{\beta}_0$ e $\hat{\beta}_1$ provenientes de diversos graus de sobreposição inicial (p_S), observa-se que suas distribuições não variam de maneira relevante para diferentes tamanhos amostrais; entretanto, são bem distintas a depender da distribuição da variável explicativa considerada (Figuras A.4 e A.5). Tal comportamento comprova que, de fato, fez-se razoável o procedimento de ter definido valores para estes parâmetros de forma particularizada para cada distribuição de X , a fim de que fossem produzidos diferentes graus de sobreposição e, conseqüentemente, de variabilidade explicada.

5.2.2 Panorama geral dos coeficientes R^2

Um resumo descritivo a respeito da distribuição dos 12 coeficientes R^2 , independentemente do tamanho amostral e da distribuição da variável explicativa, é apresentado por meio da Figura 5.5 e da Tabela 5.1.

Nota-se que o patamar mediano da maior parte dos coeficientes encontra-se abaixo de 0,20, o que está relacionado à representatividade expressiva de amostras com altas proporções de sobreposição final (p'_S), conforme comentado na Seção 5.2.1. Entretanto, valores altos dos coeficientes de determinação também estão sendo devidamente representados.

Conforme críticas pertinentes apresentadas por Hagle e Mitchell (1992) e Nagelkerke (1991), respectivamente, observa-se que os coeficientes R^2_{AN} e R^2_{MA} demonstraram valores limitantes bem inferiores a 1 (Tabela 5.1), o que representa uma desvantagem de interpretação e uma violação do 3º critério de adequabilidade de Mittlböck e Schemper (1996). Assim sendo, estes dois índices não representam bons candidatos a coeficiente de determinação para a regressão logística.

Por outro lado, o coeficiente R^2_{MQ} assume valores inferiores a zero (Tabela 5.1). Este compor-

Distribuições gerais dos coeficientes R^2

Figura 5.5: Boxplots das distribuições dos coeficientes de determinação do tipo R^2 .

Coeficiente	Média	Desv. Pad.	Mín.	$p_{10\%}$	$p_{25\%}$	$p_{50\%}$	$p_{75\%}$	$p_{90\%}$	Máx.
R^2_{MQ}	0,257	0,251	-0,039	0,009	0,045	0,172	0,414	0,655	1,000
R^2_{MF}	0,243	0,238	0,000	0,009	0,045	0,163	0,385	0,617	1,000
R^2_{AN}	0,166	0,146	0,000	0,008	0,038	0,127	0,270	0,394	0,581
R^2_{HM}	0,338	0,277	0,000	0,019	0,086	0,279	0,555	0,763	1,000
R^2_{MA}	0,193	0,181	0,000	0,008	0,039	0,135	0,309	0,478	0,750
R^2_N	0,310	0,270	0,000	0,015	0,070	0,239	0,508	0,732	1,000
R^2_P	0,257	0,251	0,000	0,009	0,045	0,172	0,414	0,655	1,000
R^2_S	0,213	0,223	0,000	0,008	0,036	0,132	0,329	0,553	1,000
R^2_K	0,171	0,204	0,000	0,005	0,026	0,094	0,242	0,432	1,000
R^2_G	0,258	0,251	0,000	0,009	0,045	0,172	0,414	0,655	1,000
R^2_{BG}	0,436	0,262	0,000	0,096	0,214	0,415	0,643	0,812	1,000
R^2_T	0,257	0,251	0,000	0,009	0,045	0,172	0,414	0,655	1,000

Tabela 5.1: Medidas resumo dos coeficientes de determinação do tipo R^2 , com destaque para as violações ao 3º critério de adequabilidade de [Mittlböck e Schemper \(1996\)](#).

tamento foi observado, de forma um pouco mais expressiva, quando a proporção geral de sucesso na amostra era inferior a 0,02 e a variável explicativa possuía distribuição normal ou gama, sendo que 2,7% das iterações com estas características apresentaram $R^2_{MQ} < 0$. Por meio de uma investigação mais detalhada, verificou-se que tais iterações possuíam em comum o fato de a variável explicativa X apresentar alguns valores *outliers*, para os quais a probabilidade de sucesso predita era igualmente atípica. Por conseguinte, pode ocorrer que o numerador da Equação (2.9), pautado

nas probabilidades $\hat{\pi}_1(x_i)$, assumiu valor maior que o do denominador, baseado em $\bar{\pi}_1$. Por conta de tal possibilidade, R_{MQ}^2 também não atende, necessariamente, ao 3º critério de adequabilidade de Mittlböck e Schemper (1996), mostrando-se inadequado como coeficiente de determinação.

Todos os demais coeficientes variam, de fato, entre os limites esperados de 0 a 1. Merece destaque o coeficiente R_{BG}^2 , que apresenta os valores de explicabilidade mais elevados entre todas as propostas, seguido por R_{HM}^2 e R_N^2 (Figura 5.5).

O coeficiente R_K^2 também se destaca em relação aos demais, devido ao fato de tender a assumir valores mais baixos. Os quartis da distribuição de R_K^2 foram os menores observados entre todos os coeficientes sob avaliação, estando um pouco abaixo dos quartis do coeficiente R_{AN}^2 (Figura 5.5). Tendo em vista a inadequabilidade já comentada deste último coeficiente, que subestima a proporção de variabilidade explicada, ocorre que R_K^2 também parece não ser uma boa escolha de coeficiente de determinação.

É possível observar que alguns dos coeficientes demonstraram grande semelhança em suas distribuições, com destaque para R_{MQ}^2 , R_P^2 , R_G^2 e R_T^2 (Figura 5.5). Comparando os valores assumidos por eles na simulação, dois a dois, verifica-se que todos os pares destes coeficientes apresentaram uma diferença média absoluta inferior a 0,01, conforme é possível verificar na Tabela 5.2. Ou seja, as quatro propostas são numericamente muito próximas entre si.

Coef.	R_{MF}^2	R_{AN}^2	R_{HM}^2	R_{MA}^2	R_N^2	R_P^2	R_S^2	R_K^2	R_G^2	R_{BG}^2	R_T^2
R_{MQ}^2	0,033	0,092	0,081	0,065	0,053	0,000	0,045	0,087	0,005	0,179	0,002
R_{MF}^2	0,000	0,083	0,095	0,065	0,067	0,033	0,061	0,093	0,033	0,193	0,033
R_{AN}^2	0,083	0,000	0,173	0,026	0,144	0,092	0,057	0,069	0,092	0,270	0,092
R_{HM}^2	0,095	0,173	0,000	0,146	0,028	0,081	0,126	0,168	0,081	0,099	0,081
R_{MA}^2	0,065	0,026	0,146	0,000	0,118	0,066	0,041	0,078	0,065	0,244	0,065
R_N^2	0,067	0,144	0,028	0,118	0,000	0,053	0,098	0,140	0,053	0,126	0,053
R_P^2	0,033	0,092	0,081	0,066	0,053	0,000	0,045	0,087	0,005	0,179	0,002
R_S^2	0,061	0,057	0,126	0,041	0,098	0,045	0,000	0,042	0,045	0,223	0,045
R_K^2	0,093	0,069	0,168	0,078	0,140	0,087	0,042	0,000	0,087	0,266	0,087
R_G^2	0,033	0,092	0,081	0,065	0,053	0,005	0,045	0,087	0,000	0,179	0,002
R_{BG}^2	0,193	0,270	0,099	0,244	0,126	0,179	0,223	0,266	0,179	0,000	0,179

Tabela 5.2: Diferenças médias absolutas entre os pares de coeficientes de determinação do tipo R^2 , com destaque para os valores inferiores a 0,01.

Apesar das baixas diferenças médias entre os pares de coeficientes mencionados no parágrafo anterior, a maioria deles apresentou diferenças absolutas superiores a 0,01 em pelo menos 5% das iterações (Tabela 5.3). Logo, entende-se que faz sentido continuar avaliando todos estes coeficientes. A única exceção é o par R_{MQ}^2 e R_P^2 , cuja diferença absoluta de valores é superior a 0,01 em apenas 0,2% das repetições; ou seja, consistem em índices quase equivalentes. O índice R_P^2 , entretanto, apresenta vantagem sobre R_{MQ}^2 por não assumir valores negativos em circunstância alguma.

Coef.	R_{MF}^2	R_{AN}^2	R_{HM}^2	R_{MA}^2	R_N^2	R_P^2	R_S^2	R_K^2	R_G^2	R_{BG}^2	R_T^2
R_{MQ}^2	69%	65%	88%	58%	82%	0,2%	44%	60%	13%	99%	5%
R_{MF}^2	0%	67%	88%	68%	84%	69%	69%	77%	69%	99%	69%
R_{AN}^2	67%	0%	89%	47%	85%	65%	57%	78%	65%	99%	65%
R_{HM}^2	88%	89%	0%	89%	79%	88%	88%	90%	88%	95%	88%
R_{MA}^2	68%	47%	89%	0%	85%	58%	53%	78%	58%	99%	58%
R_N^2	84%	85%	79%	85%	0%	82%	84%	86%	82%	97%	82%
R_P^2	69%	65%	88%	58%	82%	0%	44%	60%	13%	99%	5%
R_S^2	69%	57%	88%	53%	84%	44%	0%	56%	44%	99%	43%
R_K^2	77%	78%	90%	78%	86%	60%	56%	0%	61%	99%	61%
R_G^2	69%	65%	88%	58%	82%	13%	44%	61%	0%	99%	5%
R_{BG}^2	99%	99%	95%	99%	97%	99%	99%	99%	99%	0%	99%

Tabela 5.3: Percentuais de iterações em que as diferenças absolutas entre os pares de coeficientes de determinação do tipo R^2 foram superiores a 0,01, com destaque para os coeficientes destacados anteriormente.

Coef.	R_{MF}^2	R_{AN}^2	R_{HM}^2	R_{MA}^2	R_N^2	R_P^2	R_S^2	R_K^2	R_G^2	R_{BG}^2	R_T^2
R_{MQ}^2	0,98	0,94	0,98	0,94	0,99	1,00	0,93	0,87	1,00	0,94	1,00
R_{MF}^2	1,00	0,89	0,98	0,89	0,99	0,98	0,88	0,82	0,99	0,95	0,98
R_{AN}^2	0,89	1,00	0,95	1,00	0,94	0,94	0,92	0,82	0,94	0,88	0,94
R_{HM}^2	0,98	0,95	1,00	0,94	1,00	0,98	0,90	0,83	0,98	0,97	0,98
R_{MA}^2	0,89	1,00	0,94	1,00	0,94	0,94	0,92	0,82	0,94	0,87	0,94
R_N^2	0,99	0,94	1,00	0,94	1,00	0,99	0,91	0,83	0,99	0,97	0,99
R_P^2	0,98	0,94	0,98	0,94	0,99	1,00	0,93	0,87	1,00	0,94	1,00
R_S^2	0,88	0,92	0,90	0,92	0,91	0,93	1,00	0,97	0,93	0,84	0,93
R_K^2	0,82	0,82	0,83	0,82	0,83	0,87	0,97	1,00	0,87	0,79	0,87
R_G^2	0,99	0,94	0,98	0,94	0,99	1,00	0,93	0,87	1,00	0,94	1,00
R_{BG}^2	0,95	0,88	0,97	0,87	0,97	0,94	0,84	0,79	0,94	1,00	0,94

Tabela 5.4: Correlações lineares de Pearson entre os pares de coeficientes de determinação do tipo R^2 , obtidas na simulação, com destaque para os valores menores que 0,80 ou a partir de 0,95.

De forma geral, observa-se que todos os pares de coeficientes apresentam uma forte associação linear positiva entre si, por meio das correlações de Pearson (Tabela 5.4). A menor correlação observada foi de 0,79, entre os coeficientes R_K^2 e R_{BG}^2 .

Ao transformar linearmente a variável explicativa ($X' = a + bX$, com $a, b \sim \mathcal{U}(0,2)$), todos os coeficientes de determinação mantiveram os valores que assumiram com a variável explicativa original, em 100% das iterações da simulação. Sendo assim, todas as propostas satisfazem o 2º critério de adequabilidade de Mittlböck e Schemper (1996).

5.2.3 Análises por cenário

As relações entre os 12 coeficientes de determinação se mostraram semelhantes em todos os cenários de tamanho amostral (Figura A.6). À medida que o tamanho amostral aumenta, nota-se uma diminuição discreta e semelhante nos valores de todos os coeficientes, o que se pode verificar a partir de seus valores médios e medianos (Tabela A.1 e Figura A.7).

Por outro lado, observa-se relações distintas entre alguns coeficientes quando a variável explicativa segue distribuição discreta, em comparação às distribuições contínuas que foram avaliadas. Quando X tem distribuição de Bernoulli, os coeficientes R_{MQ}^2 , R_P^2 , R_S^2 e R_K^2 são equivalentes entre si do ponto de vista algébrico, tendo apresentado valores idênticos dentro de cada iteração (Tabela A.2 e Figura A.8). Ainda neste cenário, R_G^2 e R_T^2 são quase equivalentes aos quatro coeficientes supracitados, pois assumiram valores iguais até a 6ª casa decimal em todas as iterações.

Nota-se que todos os coeficientes tenderam a assumir valores mais elevados quando a distribuição de X é Bernoulli. Este fato está relacionado à diferença entre os formatos das distribuições da proporção de sobreposição final (p'_S), a depender de se X é variável explicativa contínua ou discreta, conforme mencionado na Seção 5.2.1 (Figura A.2). Entretanto, tal comportamento não traz prejuízo para a análise qualitativa dos coeficientes, tendo em vista que sua adequabilidade não será julgada a depender de se os valores assumidos são mais altos ou mais baixos, de forma isolada; mas, sim, a depender do comportamento de tais coeficientes em relação a algumas medidas de referência, conforme discutir-se-á nas Seções 5.2.4 e 5.2.5.

Quanto à limitação discutida em relação aos coeficientes R_{AN}^2 e R_{MA}^2 na Seção 5.2.2, verifica-se que ambos apresentaram valores máximos inferiores a 1 em todos os cenários de tamanho amostral e de distribuição de X (Figuras A.7 e A.9 e Tabela A.2). O mesmo ocorreu para o coeficiente R_K^2 , que já se mostrou inadequado anteriormente, bem como para R_S^2 , com exceção do caso em que X segue distribuição de Bernoulli. Dessa forma, o coeficiente R_S^2 também demonstra não ser uma opção perfeitamente adequada de coeficiente de determinação.

5.2.4 Relação com a proporção de sobreposição final (p'_S)

Avaliou-se o comportamento de cada coeficiente R^2 versus a proporção de sobreposição final (p'_S), definida na Equação (5.4). Devido à grande quantidade de iterações da simulação, os gráficos de dispersão referenciados nesta e na próxima seção foram gerados a partir de amostras aleatórias de 20.000 pontos, seja na avaliação geral ou nas avaliações por cenário. Adicionalmente, para observar de forma mais clara a relação entre os coeficientes R^2 e p'_S , são apresentadas curvas suavizadas a partir do método *loess* (Cleveland e Devlin, 1988), utilizando parâmetro de suavização $\alpha = 0,2$; ou seja, cada ponto da curva polinomial foi estimado a partir de uma vizinhança local correspondente a 20% dos pontos disponíveis.

A partir da Figura 5.6, percebe-se que todos os coeficientes R^2 apresentam relações inversamente proporcionais com p'_S , tal como esperado. Observa-se, ainda, que diversos coeficientes se aproximam fortemente de 1 à medida que p'_S se aproxima de zero, o que corresponde ao cenário de separação quase completa. A exceção se dá para os coeficientes R^2_{AN} , R^2_{MA} , R^2_S e R^2_K , cuja inadequabilidade já foi comentada anteriormente. Este problema pode ser identificado em todos os cenários de tamanho amostral (Figuras A.10 a A.13) e, no caso de R^2_S e R^2_K , apenas quando a variável explicativa segue distribuição contínua (Figuras A.14 a A.17).

Quando p'_S se aproxima de 1, o que indica total ausência de explicabilidade da variável resposta por meio da variável explicativa, espera-se que os valores de R^2 sejam próximos de zero. Esta relação é notada para quase todos os coeficientes, com exceção de R^2_{BG} , que apresenta grande dispersão em sua distribuição, mesmo em situações com alta sobreposição associada (Figura 5.6). Isso ocorre em todos os cenários de tamanho amostral (Figuras A.10 a A.13) e, especialmente, para variáveis explicativas com distribuição contínua (Figuras A.14 a A.17). Dessa forma, o coeficiente R^2_{BG} não parece mensurar o grau de variabilidade explicada de forma satisfatória.

Convém notar que a maior parte dos coeficientes R^2 apresentam relações não lineares com p'_S , bem como alta concentração de valores menores que $1 - p'_S$, o que corresponde à área triangular inferior dos gráficos da Figura 5.6. Tal comportamento mostra-se razoável, pois, intuitivamente, é de se esperar que uma sobreposição de $(100 \cdot p'_S)\%$ nas posições do vetor $\mathbf{y}$ (ou, equivalentemente, de $\mathbf{x}$), em relação ao cenário de separação completa, contribua com uma perda de, no mínimo, $(100 \cdot p'_S)\%$ no grau de variabilidade explicada. Nesse sentido, o coeficiente R^2_{BG} é o que se destaca de forma mais negativa, tendo assumido valores menores que $1 - p'_S$ em apenas 21,8% das iterações da simulação.

Restringindo a avaliação apenas às propostas de R^2 que demonstraram comportamento razoável até então, é possível perceber que, à medida que se incrementa o tamanho amostral, todos os co-

Relações entre os coeficientes R^2 e o índice p'_S

Figura 5.6: Gráficos de dispersão representando as relações entre os coeficientes de determinação R^2 e a proporção de sobreposição final (p'_S).

eficientes tendem a apresentar uma associação mais forte com o índice p'_S . Este comportamento é notável, graficamente, a partir da diminuição gradual da dispersão em torno das curvas *loess* (Figuras A.10 a A.13). A partir da Tabela 5.5, chama-se a atenção, de forma especial, para os coeficientes R_P^2 , R_G^2 e R_T^2 , cujos pontos nos gráficos de dispersão apresentaram menor variabilidade média em relação às curvas *loess*, para todos os cenários de tamanho amostral e de distribuição de X . Dessa forma, estes coeficientes são os que apresentam relação mais estreita com a proporção de sobreposição final.

Coeficiente		R_{MF}^2	R_{HM}^2	R_N^2	R_P^2	R_G^2	R_T^2
Geral		0,040	0,041	0,040	0,025	0,026	0,025
Por tamanho amostral	$n = 100$	0,050	0,057	0,054	0,040	0,044	0,042
	$n = 500$	0,038	0,038	0,037	0,022	0,024	0,023
	$n = 1.000$	0,037	0,036	0,035	0,020	0,021	0,021
	$n = 5.000$	0,037	0,034	0,034	0,018	0,018	0,018
Por distribuição de X	$X \sim \mathcal{N}(0,1)$	0,036	0,032	0,031	0,019	0,022	0,020
	$X \sim \text{Gama}(n,1)$	0,031	0,028	0,027	0,020	0,022	0,020
	$X \sim \mathcal{U}(0,1)$	0,044	0,042	0,041	0,019	0,022	0,021
	$X \sim \text{Ber}(\bar{\pi}_1)$	0,029	0,021	0,022	0,020	0,020	0,020

Tabela 5.5: Diferenças médias absolutas entre os valores obtidos e os valores suavizados de alguns coeficientes, a partir das curvas *loess* geradas em relação ao índice p'_S . Em destaque, os coeficientes com menores diferenças médias, em cada cenário.

5.2.5 Relações com os índices de acurácia preditiva

Cada coeficiente de determinação do tipo R^2 foi comparado com o índice KS de Kolmogorov-Smirnoff, definido na Equação (3.6), que se faz uma das alternativas mais razoáveis para mensuração da acurácia na previsão da variável resposta dicotômica. Avaliou-se, ainda, o comportamento dos coeficientes *versus* o índice de Gini, definido na Equação (3.5), que consiste numa adaptação da área abaixo da curva ROC para a escala de 0 a 1.

De forma semelhante ao que se observou para a proporção de sobreposição final, alguns coeficientes demonstram comportamentos inapropriados em relação ao índice KS , levando em conta que podem assumir valores baixos, mesmo em situações em que KS assume valores altos. Tal incoerência ocorre, de maneira mais comum, para os coeficientes R_{AN}^2 , R_{MA}^2 , R_S^2 e R_K^2 , o que pode ser notado a partir da alta concentração de pontos no canto inferior direito dos respectivos gráficos de dispersão (Figura 5.7). Também é observada para R_{MQ}^2 , R_P^2 , R_G^2 e R_T^2 , especialmente em amostras menores (Figuras A.18 a A.21) e para distribuições contínuas de X (Figuras A.22 a A.25).

Relações entre os coeficientes R^2 e o índice KS

Figura 5.7: Gráficos de dispersão representando as relações entre os coeficientes de determinação R^2 e o índice de Kolmogorov-Smirnov (KS).

Sendo assim, o grau de variabilidade explicada segundo todos esses coeficientes não apresenta total conformidade com o conceito de acurácia preditiva, quando mensurada por meio de KS .

Adicionalmente, a partir da Tabela 5.6, observa-se que as diferenças absolutas entre os valores de R^2 e KS foram superiores a 0,50 em pelo menos 1,8% das iterações, para vários dos coeficientes avaliados. Diferenças de tal magnitude apontam dissonâncias importantes entre os valores dos coeficientes de determinação e o índice KS . As exceções foram os coeficientes R_{MF}^2 , R_{HM}^2 , R_N^2 e R_{BG}^2 , que apresentaram tal situação problemática com menor frequência; entretanto, vale ressaltar que R_{BG}^2 já havia se mostrado inadequado nas análises relacionadas a p'_S .

Coef.	% de iterações com $ R^2 - KS > 0,5$
R_{MQ}^2	1,9%
R_{MF}^2	0,3%
R_{AN}^2	9,0%
R_{HM}^2	0,1%
R_{MA}^2	7,1%
R_N^2	0,2%
R_P^2	1,9%
R_S^2	7,1%
R_K^2	10,9%
R_G^2	1,8%
R_{BG}^2	0%
R_T^2	1,9%

Tabela 5.6: Percentual de iterações em que a diferença absoluta entre o índice KS e cada coeficiente de determinação foi superior a 0,50. Em destaque, os coeficientes que se mostraram mais apropriados nesta avaliação.

Com exceção de R_{HM}^2 , R_N^2 e R_{BG}^2 , o índice KS assumiu valores maiores que os dos demais coeficientes em 100% das iterações (Tabela 5.7), parecendo se comportar como um limitante superior para os valores de diversas propostas de coeficiente de determinação. Esse comportamento mostra-se razoável, tendo em vista que a avaliação de assertividade da resposta dicotômica, tal como realizada pelo KS , é menos restritiva do que a quantificação da variabilidade explicada, que é realizada pelos coeficientes do tipo R^2 . Enquanto o índice KS leva em conta apenas as medidas de sensibilidade e especificidade, ou seja, a capacidade de o modelo prever corretamente os casos de sucesso e fracasso dado um ponto de corte específico, as propostas de R^2 baseiam-se em cálculos que envolvem diretamente as probabilidades preditas de sucesso, mensurando de forma “contínua” o quão distantes elas estão dos extremos 0 e 1 (cenário de separação completa).

Coef.	% de iterações com $KS > R^2$
R_{MQ}^2	100%
R_{MF}^2	100%
R_{AN}^2	100%
R_{HM}^2	96,8%
R_{MA}^2	100%
R_N^2	99,4%
R_P^2	100%
R_S^2	100%
R_K^2	100%
R_G^2	100%
R_{BG}^2	90,5%
R_T^2	100%

Tabela 5.7: Percentual de iterações em que o índice KS apresentou valor superior ao de cada coeficiente de determinação. Em destaque, os coeficientes que se mostraram mais apropriados nesta avaliação.

Dessa forma, R_{MF}^2 foi o coeficiente que demonstrou maior coerência com o índice KS , com base nas avaliações que foram realizadas. Quando comparado a R_{HM}^2 e R_N^2 , o coeficiente R_{MF}^2 foi, ainda, o que apresentou menor variabilidade média em relação às curvas *loess* dos gráficos de dispersão *versus* KS , para todos os cenários avaliados (Tabela 5.8).

Coeficiente		R_{MF}^2	R_{HM}^2	R_N^2
Geral		0,031	0,047	0,043
Por tamanho amostral	$n = 100$	0,039	0,058	0,055
	$n = 500$	0,028	0,046	0,042
	$n = 1.000$	0,026	0,042	0,039
	$n = 5.000$	0,024	0,039	0,036
Por distribuição de X	$X \sim \mathcal{N}(0,1)$	0,039	0,058	0,055
	$X \sim \text{Gama}(n,1)$	0,028	0,046	0,042
	$X \sim \mathcal{U}(0,1)$	0,026	0,042	0,039
	$X \sim \text{Ber}(\bar{\pi}_1)$	0,024	0,039	0,036

Tabela 5.8: Diferenças médias absolutas entre os valores obtidos e os valores suavizados de alguns coeficientes, a partir das curvas *loess* geradas em relação ao índice KS . Em destaque, as menores diferenças médias, em cada cenário.

Já com base no índice de Gini (G), observa-se, tal como para KS , que diversos coeficientes do tipo R^2 assumem valores relativamente baixos em ocasiões em que G assume valores altos,

Relações entre os coeficientes R^2 e o índice de Gini

Figura 5.8: Gráficos de dispersão representando as relações entre os coeficientes de determinação R^2 e o índice de Gini (G).

conforme apresentado na Tabela 5.9 e Figura 5.8.

Coef.	% de iterações com $ R^2 - G > 0,5$
R_{MQ}^2	4,1%
R_{MF}^2	0,8%
R_{AN}^2	16,2%
R_{HM}^2	0,2%
R_{MA}^2	11,8%
R_N^2	0,4%
R_P^2	4,1%
R_S^2	12,3%
R_K^2	23,3%
R_G^2	3,9%
R_{BG}^2	0%
R_T^2	4,0%

Tabela 5.9: Percentual de iterações em que a diferença absoluta entre o índice de Gini (G) e cada coeficiente de determinação foi superior a 0,50. Em destaque, os coeficientes que se mostraram mais apropriados nesta avaliação.

Os coeficientes que se comportam de forma mais coerente com o índice G são R_{MF}^2 , R_{HM}^2 , R_N^2 e R_{BG}^2 , para todos os tamanhos amostrais (Figuras A.26 a A.29) e distribuições da variável explicativa X (Figuras A.30 a A.33). Particularmente, vale ressaltar que R_{BG}^2 é algebricamente idêntico a KS (Figura A.25) e a G (Figura A.33), no caso em que a variável explicativa segue

Coeficiente		R_{MF}^2	R_{HM}^2	R_N^2
Geral		0,034	0,046	0,044
Por tamanho amostral	$n = 100$	0,036	0,047	0,045
	$n = 500$	0,034	0,047	0,045
	$n = 1.000$	0,034	0,046	0,044
	$n = 5.000$	0,033	0,046	0,044
Por distribuição de X	$X \sim \mathcal{N}(0,1)$	0,018	0,042	0,038
	$X \sim \text{Gama}(n,1)$	0,018	0,040	0,036
	$X \sim \mathcal{U}(0,1)$	0,017	0,041	0,037
	$X \sim \text{Ber}(\bar{\pi}_1)$	0,020	0,012	0,012

Tabela 5.10: Diferenças médias absolutas entre os valores obtidos e os valores suavizados de alguns coeficientes, a partir das curvas loess geradas em relação ao índice de Gini (G). Em destaque, os coeficientes com menores diferenças médias, em cada cenário.

distribuição de Bernoulli.

Novamente, R_{MF}^2 demonstrou maior grau de associação não linear com o índice de Gini do que os coeficientes R_{HM}^2 e R_N^2 , mensurada por meio da diferença média absoluta entre os pontos e a curva *loess* dos respectivos gráficos de dispersão. Este fato ocorreu para todos os tamanhos amostrais e todas as distribuições contínuas de X (Tabela 5.10).

5.2.6 Acréscimo de variável explicativa

Incorporando ao modelo uma segunda variável explicativa, gerada de forma independente da variável resposta e da explicativa já presente, os valores dos coeficientes R_{MF}^2 , R_{AN}^2 , R_{HM}^2 , R_{MA}^2 e R_N^2 foram maiores ou iguais aos do ajuste inicial em 100% das iterações (Tabela 5.11). Este resultado indica que tais coeficientes carecem, de alguma maneira, de uma correção para o caso da regressão logística múltipla, de forma que não sejam superestimados por conta da presença de variáveis com baixa ou nenhuma explicabilidade da resposta.

Coeficiente	Média	Desv. Pad.	Mín.	$p_{1\%}$	$p_{25\%}$	$p_{50\%}$	$p_{75\%}$	$p_{99\%}$	Máx.
R_{MQ}^2	0,005	0,020	-0,119	-0,005	<0,001	0,001	0,004	0,066	0,219
R_{MF}^2	0,005	0,016	0	<0,001	<0,001	0,001	0,004	0,061	0,173
R_{AN}^2	0,003	0,007	0	<0,001	<0,001	0,001	0,003	0,035	0,066
R_{HM}^2	0,007	0,018	0	<0,001	<0,001	0,002	0,006	0,081	0,192
R_{MA}^2	0,004	0,008	0	<0,001	<0,001	0,001	0,003	0,040	0,076
R_N^2	0,006	0,017	0	<0,001	<0,001	0,001	0,005	0,074	0,174
R_P^2	0,005	0,020	-0,116	-0,005	<0,001	0,001	0,004	0,066	0,219
R_S^2	0,004	0,021	-0,388	-0,006	<0,001	0,001	0,003	0,044	0,371
R_K^2	0,003	0,022	-0,388	-0,004	<0,001	0,001	0,002	0,030	0,472
R_G^2	0,005	0,018	-0,011	<0,001	<0,001	0,001	0,004	0,060	0,181
R_{BG}^2	-0,193	0,160	-0,931	-0,695	-0,278	-0,155	-0,072	0,010	0,078
R_T^2	0,005	0,019	-0,045	-0,001	<0,001	0,001	0,004	0,062	0,202

Tabela 5.11: Medidas resumo da variação nos valores dos coeficientes de determinação do tipo R^2 , após inclusão de uma segunda variável explicativa, independente em relação à variável resposta.

Entre os demais coeficientes avaliados, todos admitem a possibilidade de diminuição de valor quando se acrescenta uma variável que não contribui com a explicabilidade. Entretanto, nota-se que R_S^2 , R_K^2 e R_{BG}^2 podem, eventualmente, reduzir ou aumentar de forma exacerbada (Tabela 5.11). Em particular, R_{BG}^2 é o único coeficiente cuja diferença média após acréscimo da variável é substancialmente menor que zero (-0,193). Dessa forma, esses três coeficientes são bastante inadequados para o caso da regressão logística múltipla.

No que tange ao coeficiente R_{MF}^2 , que mostrou-se uma alternativa pertinente a partir das análises realizadas nas seções anteriores, [McFadden \(1974\)](#) sugeriu R_{MFadj}^2 como uma alternativa para o caso múltiplo, tal como apresentado na Equação (4.16). Com efeito, observa-se que este coeficiente pode decrescer, em relação a R_{MF}^2 , mediante inclusão de uma variável explicativa espúria (Tabela 5.12).

Coeficiente	Média	Desv. Pad.	Mín.	$p_{1\%}$	$p_{25\%}$	$p_{50\%}$	$p_{75\%}$	$p_{99\%}$	Máx.
R_{MFadj}^2	-0,004	0,015	-0,179	-0,039	-0,008	-0,002	-0,001	0,036	0,140

Tabela 5.12: Medidas resumo da variação nos valores de R_{MFadj}^2 , após inclusão de uma segunda variável explicativa, independente em relação à variável resposta.

Convém notar que os coeficientes R_{MQ}^2 , R_P^2 , R_G^2 e R_T^2 , apesar de não sofrerem de alguma inadequabilidade séria, ainda tendem a assumir valores maiores que os de R_{MFadj}^2 , quando se inclui a variável sem força de explicabilidade (Tabela 5.11). Houve acréscimo no valor dos coeficientes mencionados em 62% a 72% das iterações, enquanto, para R_{MFadj}^2 , tal comportamento foi registrado em apenas 12,6% das iterações (Tabela 5.13). Dessa forma, R_{MQ}^2 , R_P^2 , R_G^2 e R_T^2 parecem não estar suficientemente “ajustados” para o caso de duas ou mais variáveis preditoras, requerendo algum tipo de correção.

Coef.	% de iterações com acréscimo em R^2 (2 var.)
R_{MQ}^2	62,6%
R_{MF}^2	75,2%
R_{MFadj}^2	12,6%
R_{AN}^2	75,2%
R_{HM}^2	75,2%
R_{MA}^2	75,2%
R_N^2	75,2%
R_P^2	62,5%
R_S^2	56,2%
R_K^2	56,2%
R_G^2	71,7%
R_{BG}^2	2,2%
R_T^2	67,5%

Tabela 5.13: Percentual de iterações em que houve acréscimo no valor dos coeficientes de determinação do tipo R^2 , após inclusão de uma segunda variável explicativa, independente em relação à variável resposta.

5.3 Considerações Finais do Capítulo

Foram apresentados os resultados de um estudo de simulação, com objetivo de analisar os comportamentos das 12 propostas de coeficientes de determinação. Verificou-se os aspectos essenciais de adequabilidade algébrica, que são endereçados por meio dos critérios 2 e 3 de [Mittlböck e Schemper \(1996\)](#), apresentados na Seção [4.3](#). Buscou-se, ainda, uma alternativa de avaliação de adequabilidade associada ao critério 4, por meio de comparações dos coeficientes de determinação com o índice de sobreposição padronizado, cujo conceito foi introduzido na Seção [5.1](#), bem como com índices de acurácia preditiva tradicionais.

Capítulo 6

Conclusões

Tendo em vista as discussões apresentadas na revisão literária (Capítulo 4), bem como os resultados do estudo de simulação (Capítulo 5), resume-se, a seguir, os aspectos de adequabilidade dos 12 coeficientes de determinação avaliados (Tabela 6.1):

Coef.	Critério 1: Interpretação intuitiva	Critério 2: Não sensibilidade a transformações lineares em X	Critério 3: Variação no intervalo de 0 a 1	Critério 4: Conformidade com índices de acurácia preditiva e p'_S
R^2_{MQ}	Atende	Atende	Não atende	Não atende
R^2_{MF}	Atende	Atende	Atende	Atende
R^2_{AN}	Não atende	Atende	Não atende	Não atende
R^2_{HM}	Não atende	Atende	Atende	Atende parcialm.
R^2_{MA}	Não atende	Atende	Não atende	Não atende
R^2_N	Não atende	Atende	Atende	Atende parcialm.
R^2_P	Atende	Atende	Atende	Não atende
R^2_S	Não atende	Atende	Não atende	Não atende
R^2_K	Não atende	Atende	Não atende	Não atende
R^2_G	Não atende	Atende	Atende	Não atende
R^2_{BG}	Atende	Atende	Atende	Não atende
R^2_T	Atende	Atende	Atende	Não atende

Tabela 6.1: Painel resumo acerca da adequabilidade dos 12 coeficientes de determinação.

Tomou-se como base a compilação de critérios de Mittlböck e Schemper (1996) (Seção 4.3), mediante alteração conveniente da 4ª afirmação, que versava sobre a similaridade dos valores R^2 obtidos em relação a um ajuste de regressão linear. Em vez disso, considerou-se a coerência dos valores R^2 em relação à proporção de sobreposição (Seção 5.1.2) e às principais medidas de acurácia

preditiva (Seções 3.2.2 e 3.2.3), que dizem respeito especialmente a modelos de classificação binária, como a regressão logística.

Ao contrário dos critérios 2 a 4, que foram examinados quantitativamente mediante estudo de simulação, a avaliação de interpretabilidade (critério 1) foi realizada de forma qualitativa, com base nos argumentos positivos e negativos apresentados em trabalhos anteriores para cada proposta. No caso de coeficientes como R_{BG}^2 e R_T^2 , que foram pouco debatidos, a não ser pelos autores que os propuseram originalmente, concluiu-se a favor de sua interpretabilidade, por serem os únicos coeficientes com forte apelo gráfico associado à quantificação da variabilidade explicada. Adicionalmente, R_T^2 conta com a vantagem da sobriedade de sua expressão algébrica, que é a mais simples e compreensível entre todas as propostas.

Observa-se que o coeficiente de determinação R_{MF}^2 , introduzido por [McFadden \(1974\)](#), foi o único que atendeu, de forma completa, a todos os critérios estabelecidos. Quanto às objeções apresentadas em relação à sua interpretabilidade, em trabalhos das décadas de 80 e 90, considera-se que todas foram devidamente respondidas no estudo posterior realizado por [Menard \(2000\)](#), conforme discussão da Seção 4.1.2. Adicionalmente, [Ben-Akiva e Lerman \(1985\)](#) propuseram um coeficiente ajustado para o caso da regressão múltipla (R_{MFadj}^2), que mostrou-se mais adequado que todos os demais coeficientes avaliados no estudo de simulação, no que diz respeito ao comportamento quando uma variável explicativa espúria é acrescentada ao modelo.

Assim como R_{MF}^2 , os coeficientes R_{HM}^2 e R_N^2 demonstraram conformidade quanto à proporção de sobreposição final (p'_S), bem como quanto ao índice de Gini (G), de forma contrária às proposições que lhes deram origem (R_{AN}^2 e R_{MA}^2 , respectivamente). Entretanto, esses dois coeficientes apresentaram uma pequena dissonância quanto ao comportamento em relação ao índice KS , frequentemente empregado na avaliação da acurácia de modelos preditivos. Além disso, a interpretabilidade de suas expressões algébricas não é trivial, o que representa uma desvantagem em relação ao 1º critério de [Mittlböck e Schemper \(1996\)](#).

Os coeficientes R_P^2 , R_{BG}^2 e R_T^2 atenderam a três dos quatro critérios de adequabilidade, falhando apenas no que diz respeito ao 4º critério. Quanto a R_{BG}^2 , suas relações com a proporção de sobreposição e o índice KS não são intuitivas, tendo em vista que assume, com muita frequência, valores superiores aos destes índices, quando o esperado seria o oposto. Já no caso de R_P^2 e R_T^2 , proporções baixas de variabilidade explicada são obtidas, eventualmente, mesmo em casos nos quais a acurácia preditiva é alta. A mesma limitação é observada, de forma mais enfática, para os coeficientes R_{MQ}^2 , R_{AN}^2 , R_{MA}^2 , R_S^2 e R_K^2 , que demonstram, também, violações ao 3º critério; ou seja, não variam entre os extremos de 0 a 1 de forma bem definida.

Por fim, os coeficientes R_{MF}^2 e R_{MFadj}^2 consolidaram-se, no âmbito deste estudo, como as alternativas mais adequadas para mensuração da variabilidade explicada nos modelos de regressão logística com resposta nominal, simples e múltipla, respectivamente. Tal conclusão é formulada, de forma especial, em decorrência da conformidade de R_{MF}^2 em relação à proporção de sobreposição e aos principais índices de acurácia preditiva, bem como devido à maior estabilidade de R_{MFadj}^2 ao se acrescentar novas variáveis ao modelo.

Como sugestão para pesquisas futuras, recomenda-se o desenvolvimento de estratégias para avaliação da adequabilidade dos coeficientes de determinação na regressão logística com resposta multinomial e não ordinal. Com efeito, essa extensão é possível para os coeficientes baseados na função de verossimilhança (ex.: R_{MF}^2), mediante definição de outros índices que possam servir como balizadores; por exemplo, por meio de uma generalização da proporção de sobreposição ou dos índices de acurácia preditiva, utilizados neste trabalho. Para diversos outros coeficientes compilados neste trabalho (ex.: R_P^2 , R_G^2 , R_T^2 etc.), não foram estabelecidas, até então, adaptações para o caso multinomial.

Apêndice A

Tabelas e Figuras

Distribuição de p_S , por tamanho amostral e distribuição de X

Figura A.1: Histograma da distribuição dos valores da proporção de sobreposição inicial (p_S), de acordo com o tamanho amostral e a distribuição de X .

Distribuição de p'_S , por tamanho amostral e distribuição de X

Figura A.2: Histograma da distribuição dos valores da proporção de sobreposição final (p'_S), de acordo com o tamanho amostral e a distribuição de X .

Distribuição de $\bar{\pi}_1$, por tamanho amostral e distribuição de X

Figura A.3: Histograma da distribuição dos valores da probabilidade geral de sucesso ($\bar{\pi}_1$), de acordo com o tamanho amostral e a distribuição de X .

Distribuição de $\hat{\beta}_0$, por tamanho amostral e distribuição de X

Figura A.4: Histograma da distribuição dos valores $\hat{\beta}_0$, de acordo com o tamanho amostral e a distribuição de X .

Distribuição de $\hat{\beta}_1$, por tamanho amostral e distribuição de X

Figura A.5: Histograma da distribuição dos valores $\hat{\beta}_1$, de acordo com o tamanho amostral e a distribuição de X .

Medida	n	R_{MQ}^2	R_{MF}^2	R_{AN}^2	R_{HM}^2	R_{MA}^2	R_N^2	R_P^2	R_S^2	R_K^2	R_G^2	R_{BG}^2	R_T^2
Mínimo	$n = 100$	-0,037	0	0	0	0	0	0	0	0	0	0	0
	$n = 500$	-0,017	0	0	0	0	0	0	0	0	0	0	0
	$n = 1.000$	-0,026	0	0	0	0	0	0	0	0	0	0	0
	$n = 5.000$	-0,012	0	0	0	0	0	0	0	0	0	0	0
Máximo	$n = 100$	1	1	0,581	1	0,750	1	1	1	1	1	1	1
	$n = 500$	1	1	0,581	1	0,750	1	1	1	1	1	1	1
	$n = 1.000$	1	1	0,578	1	0,746	1	1	1	1	1	1	1
	$n = 5.000$	1	1	0,580	1	0,749	1	1	1	1	1	1	1
Média	$n = 100$	0,268	0,252	0,176	0,353	0,204	0,323	0,268	0,224	0,179	0,269	0,447	0,268
	$n = 500$	0,258	0,244	0,166	0,339	0,193	0,311	0,258	0,213	0,171	0,258	0,438	0,258
	$n = 1.000$	0,255	0,242	0,164	0,335	0,190	0,308	0,255	0,210	0,169	0,255	0,434	0,255
	$n = 5.000$	0,251	0,238	0,160	0,329	0,185	0,302	0,251	0,206	0,166	0,251	0,428	0,251
Mediana	$n = 100$	0,190	0,177	0,142	0,303	0,152	0,261	0,191	0,149	0,108	0,191	0,432	0,190
	$n = 500$	0,173	0,165	0,126	0,280	0,134	0,241	0,173	0,131	0,094	0,173	0,418	0,173
	$n = 1.000$	0,167	0,160	0,122	0,273	0,130	0,234	0,167	0,126	0,090	0,167	0,412	0,167
	$n = 5.000$	0,161	0,154	0,117	0,264	0,124	0,226	0,161	0,121	0,087	0,160	0,404	0,161
Desvio padrão	$n = 100$	0,249	0,237	0,145	0,276	0,180	0,268	0,249	0,222	0,203	0,250	0,258	0,249
	$n = 500$	0,251	0,237	0,146	0,277	0,181	0,269	0,251	0,224	0,204	0,251	0,261	0,251
	$n = 1.000$	0,253	0,239	0,147	0,278	0,182	0,271	0,253	0,225	0,205	0,253	0,263	0,253
	$n = 5.000$	0,253	0,240	0,147	0,279	0,181	0,271	0,253	0,225	0,206	0,253	0,265	0,253

Tabela A.1: Medidas resumo dos coeficientes de determinação do tipo R^2 , por tamanho amostral.

Distribuições dos coeficientes R^2 , de acordo com o tamanho amostral

Figura A.6: *Boxplots das distribuições dos coeficientes de determinação do tipo R^2 , por tamanho amostral.*

Distribuições dos coeficientes R^2 , de acordo com o tamanho amostral

Figura A.7: Boxplots das distribuições dos coeficientes de determinação do tipo R^2 , por tamanho amostral.

Medida	Distrib. de X	R_{MQ}^2	R_{MF}^2	R_{AN}^2	R_{HM}^2	R_{MA}^2	R_N^2	R_P^2	R_S^2	R_K^2	R_G^2	R_{BG}^2	R_T^2
Mínimo	Normal	-0,004	0	0	0	0	0	0	0	0	0	0	0
	Gama	-0,037	0	0	0	0	0	0	0	0	0	0	0
	Uniforme	0	0	0	0	0	0	0	0	0	0	0	0
	Bernoulli	0	0	0	0	0	0	0	0	0	0	0	0
Máximo	Normal	1	1	0,581	1	0,750	1	1	0,984	0,981	1	1	1
	Gama	1	1	0,579	1	0,747	1	1	0,980	0,962	1	1	1
	Uniforme	1	1	0,574	1	0,740	1	1	0,979	0,961	1	1	1
	Bernoulli	1	1	0,581	1	0,750	1	1	1	1	1	1	1
Média	Normal	0,227	0,224	0,156	0,317	0,179	0,289	0,227	0,172	0,116	0,227	0,409	0,227
	Gama	0,245	0,231	0,157	0,324	0,180	0,296	0,246	0,161	0,108	0,245	0,405	0,245
	Uniforme	0,219	0,218	0,158	0,313	0,183	0,285	0,220	0,181	0,122	0,220	0,423	0,220
	Bernoulli	0,339	0,304	0,195	0,401	0,228	0,373	0,339	0,339	0,339	0,339	0,510	0,339
Mediana	Normal	0,149	0,148	0,116	0,256	0,123	0,218	0,150	0,112	0,075	0,150	0,386	0,149
	Gama	0,172	0,154	0,118	0,265	0,125	0,226	0,172	0,096	0,064	0,171	0,376	0,172
	Uniforme	0,135	0,145	0,117	0,251	0,124	0,214	0,136	0,118	0,079	0,136	0,409	0,136
	Bernoulli	0,259	0,224	0,161	0,361	0,175	0,316	0,259	0,259	0,259	0,259	0,509	0,259
Desvio padrão	Normal	0,226	0,221	0,139	0,265	0,171	0,256	0,226	0,173	0,117	0,226	0,247	0,226
	Gama	0,233	0,227	0,139	0,268	0,171	0,259	0,233	0,171	0,115	0,233	0,252	0,233
	Uniforme	0,226	0,218	0,143	0,265	0,177	0,255	0,226	0,180	0,122	0,227	0,247	0,226
	Bernoulli	0,296	0,274	0,159	0,302	0,199	0,297	0,296	0,296	0,296	0,296	0,285	0,296

Tabela A.2: Medidas resumo dos coeficientes de determinação do tipo R^2 , por distribuição de X .

Distribuições dos coeficientes R^2 , de acordo com a distribuição de X

Figura A.8: *Boxplots das distribuições dos coeficientes de determinação do tipo R^2 , por distribuição de X .*

Distribuições dos coeficientes R^2 , de acordo com a distribuição de X Figura A.9: Boxplots das distribuições dos coeficientes de determinação do tipo R^2 , por distribuição de X .

Relações entre os coeficientes R^2 e o índice p'_S , para $n = 100$

Figura A.10: Gráficos de dispersão representando as relações entre os coeficientes de determinação do tipo R^2 e a proporção de sobreposição final (p'_S), para $n = 100$.

Relações entre os coeficientes R^2 e o índice p'_S , para $n = 500$

Figura A.11: Gráficos de dispersão representando as relações entre os coeficientes de determinação do tipo R^2 e a proporção de sobreposição final (p'_S), para $n = 500$.

Relações entre os coeficientes R^2 e o índice p'_S , para $n = 1.000$

Figura A.12: Gráficos de dispersão representando as relações entre os coeficientes de determinação do tipo R^2 e a proporção de sobreposição final (p'_S), para $n = 1.000$.

Relações entre os coeficientes R^2 e o índice p'_S , para $n = 5.000$

Figura A.13: Gráficos de dispersão representando as relações entre os coeficientes de determinação do tipo R^2 e a proporção de sobreposição final (p'_S), para $n = 5.000$.

Relações entre os coeficientes R^2 e o índice p'_S , para $X \sim \text{Normal}$

Figura A.14: Gráficos de dispersão representando as relações entre os coeficientes de determinação do tipo R^2 e a proporção de sobreposição final (p'_S), para $X \sim \mathcal{N}(0,1)$.

Relações entre os coeficientes R^2 e o índice p'_S , para $X \sim \text{Gama}$

Figura A.15: Gráficos de dispersão representando as relações entre os coeficientes de determinação do tipo R^2 e a proporção de sobreposição final (p'_S), para $X \sim \text{Gama}(n,1)$.

Relações entre os coeficientes R^2 e o índice p'_S , para $X \sim \text{Uniforme}$

Figura A.16: Gráficos de dispersão representando as relações entre os coeficientes de determinação do tipo R^2 e a proporção de sobreposição final (p'_S), para $X \sim \mathcal{U}(0,1)$.

Relações entre os coeficientes R^2 e o índice p'_S , para $X \sim \text{Bernoulli}$

Figura A.17: Gráficos de dispersão representando as relações entre os coeficientes de determinação do tipo R^2 e a proporção de sobreposição final (p'_S), para $X \sim \text{Ber}(\pi_1)$.

Relações entre os coeficientes R^2 e o índice KS, para $n = 100$

Figura A.18: Gráficos de dispersão representando as relações entre os coeficientes de determinação do tipo R^2 e o índice de Kolmogorov-Smirnoff (KS), para $n = 100$.

Relações entre os coeficientes R^2 e o índice KS, para $n = 500$

Figura A.19: Gráficos de dispersão representando as relações entre os coeficientes de determinação do tipo R^2 e o índice de Kolmogorov-Smirnoff (KS), para $n = 500$.

Relações entre os coeficientes R^2 e o índice KS, para $n = 1.000$

Figura A.20: Gráficos de dispersão representando as relações entre os coeficientes de determinação do tipo R^2 e o índice de Kolmogorov-Smirnoff (KS), para $n = 1.000$.

Relações entre os coeficientes R^2 e o índice KS, para $n = 5.000$

Figura A.21: Gráficos de dispersão representando as relações entre os coeficientes de determinação do tipo R^2 e o índice de Kolmogorov-Smirnoff (KS), para $n = 5.000$.

Relações entre os coeficientes R^2 e o índice KS, para $X \sim \text{Normal}$

Figura A.22: Gráficos de dispersão representando as relações entre os coeficientes de determinação do tipo R^2 e o índice de Kolmogorov-Smirnoff (KS), para $X \sim \mathcal{N}(0,1)$.

Relações entre os coeficientes R^2 e o índice KS, para $X \sim \text{Gama}$

Figura A.23: Gráficos de dispersão representando as relações entre os coeficientes de determinação do tipo R^2 e o índice de Kolmogorov-Smirnoff (KS), para $X \sim \text{Gama}(n,1)$.

Relações entre os coeficientes R^2 e o índice KS, para $X \sim \text{Uniforme}$

Figura A.24: Gráficos de dispersão representando as relações entre os coeficientes de determinação do tipo R^2 e o índice de Kolmogorov-Smirnoff (KS), para $X \sim \mathcal{U}(0,1)$.

Relações entre os coeficientes R^2 e o índice KS, para $X \sim \text{Bernoulli}$

Figura A.25: Gráficos de dispersão representando as relações entre os coeficientes de determinação do tipo R^2 e o índice de Kolmogorov-Smirnoff (KS), para $X \sim \text{Ber}(\pi_1)$.

Relações entre os coeficientes R^2 e o índice de Gini, para $n = 100$

Figura A.26: Gráficos de dispersão representando as relações entre os coeficientes de determinação do tipo R^2 e o índice de Gini (G), para $n = 100$.

Relações entre os coeficientes R^2 e o índice de Gini, para $n = 500$

Figura A.27: Gráficos de dispersão representando as relações entre os coeficientes de determinação do tipo R^2 e o índice de Gini (G), para $n = 500$.

Relações entre os coeficientes R^2 e o índice de Gini, para $n = 1.000$

Figura A.28: Gráficos de dispersão representando as relações entre os coeficientes de determinação do tipo R^2 e o índice de Gini (G), para $n = 1.000$.

Relações entre os coeficientes R^2 e o índice de Gini, para $n = 5.000$

Figura A.29: Gráficos de dispersão representando as relações entre os coeficientes de determinação do tipo R^2 e o índice de Gini (G), para $n = 5.000$.

Relações entre os coeficientes R^2 e o índice de Gini, para $X \sim \text{Normal}$

Figura A.30: Gráficos de dispersão representando as relações entre os coeficientes de determinação do tipo R^2 e o índice de Gini (G), para $X \sim \mathcal{N}(0,1)$.

Relações entre os coeficientes R^2 e o índice de Gini, para $X \sim \text{Gama}$

Figura A.31: Gráficos de dispersão representando as relações entre os coeficientes de determinação do tipo R^2 e o índice de Gini (G), para $X \sim \text{Gama}(n,1)$.

Relações entre os coeficientes R^2 e o índice de Gini, para $X \sim \text{Uniforme}$

Figura A.32: Gráficos de dispersão representando as relações entre os coeficientes de determinação do tipo R^2 e o índice de Gini (G), para $X \sim \mathcal{U}(0,1)$.

Relações entre os coeficientes R^2 e o índice de Gini, para $X \sim \text{Bernoulli}$

Figura A.33: Gráficos de dispersão representando as relações entre os coeficientes de determinação do tipo R^2 e o índice de Gini (G), para $X \sim \text{Ber}(\bar{\pi}_1)$.

Apêndice B

Códigos da Simulação

(em linguagem R)

```
##### Pacotes necessarios

# library(tidyverse)
# library(testit)
# library(ROCit)
# library(Rmpfr)
# library(cutpointr)

##### Funcao auxiliar: geracao de variavel explicativa com distrib. de Bernoulli

cat = function(n, p = NA) {

  perc = ifelse(is.na(p), runif(1, min = 0, max = 0.5), p)

  return(c(rep(1, round(n * perc)), rep(0, n - round(n * perc))))

}

##### Simulacao

# Semente aleatoria

set.seed(12345)

# Parametros da simulacao

n_values = c(100,
             500,
             1000,
             5000)

x_values = c("rnorm(n)",
             "rgamma(n,1)",
             "runif(n)",
             "cat(n,overall_p_pre)")

x_labels = c("Normal",
             "Gama",
             "Uniforme",
             "Bernoulli")

repetitions = 20000

# Inicializando base de resultados
```

```

results = data.frame()

# Laco principal da simulacao

scenario = 1
i0 = 1

for (i1 in 1:length(n_values)) {

  for (i2 in 1:length(x_values)) {

    for (rep in 1:repetitions) {

      # Inicializando objetos

      n = n_values[i1]
      x = rep(0, n)
      y = rep(0, n)
      success_fit = !has_warning(glm(y ~ x,
                                     family = binomial(link = 'logit')))

      acceptable_pos = FALSE

      while (sum(y) == 0 |
             acceptable_pos == FALSE) {

        # Sorteando a taxa geral de sucesso

        overall_p_pre = runif(1, min = 0, max = 0.5)

        # Gerando a variavel explicativa X

        x = eval(parse(text = x_values[i2]))
        x = sort(x, decreasing = TRUE)

        # Sorteando a proporcao de sobreposicao inicial

        n_response = round(n * overall_p_pre)
        perc_overlap_pre = runif(1, min = 0, max = 1 - overall_p_pre)
        n_overlap = ceiling(n_response * perc_overlap_pre)

        perc_overlap_pre = min(1, n_overlap / (n_response * (1 - n_response / n)))

        # Gerando a variavel resposta

        y = c(sample(c(rep(1, n_response - n_overlap),
                       rep(0, n_overlap))),
              sample(c(rep(0, n - n_response - n_overlap),
                       rep(1, n_overlap))))

        # Avaliando a realizacao efetiva do ajuste

        success_fit_pre = !has_warning(glm(y ~ x,
                                           family = binomial(link = 'logit')))

        # Ajustando um novo modelo, segundo definicoes teoricas

        if (sum(y) != 0) {

          # Definindo parametros b0 e b1

          b0 = glm(y ~ x,
                   family = binomial(link = 'logit'))$coefficients[1]
          b1 = glm(y ~ x,

```



```

    family = binomial(link = 'logit'))$coefficients[2]

# Gerando a variavel resposta, a partir de b0 e b1

y = rep(0, n)
while (sum(y) == 0) {
  p = exp(b0 + b1 * x) / (1 + exp(b0 + b1 * x))
  y = ifelse(runif(n) <= p, 1, 0)
}

# Calculando a proporcao de sobreposicao final

perc_overlap_pos = data.frame(x, y) %>%
  arrange(x) %>%
  slice_tail(n = sum(y)) %>%
  summarize(sum = sum(y)) %>%
  .$sum/sum(y)

perc_overlap_pos = (1 - perc_overlap_pos) / (1 - sum(y)/n)

# Avaliando possiveis limitacoes

acceptable_pos = (sum(y)/n <= 0.5) & (perc_overlap_pos <= 1)

if (acceptable_pos == TRUE) {

  # Avaliando a realizacao efetiva do ajuste

  success_fit_pos = !has_warning(glm(y ~ x,
                                     family = binomial(link = 'logit')))

  # Ajustando modelo basal

  model_0 = glm(y ~ 1,
               family = binomial(link = 'logit'))

  # Ajustando modelo completo

  model_1 = glm(y ~ x,
               family = binomial(link = 'logit'))

  # Ajustando modelo com acrescimo de variavel com explicabilidade fraca/nula

  x_rand = eval(parse(text = x_values[i2]))
  model_rand = glm(y ~ x + x_rand,
                  family = binomial(link = 'logit'))

  # Ajustando modelo com acrescimo de variavel linearmente transformada

  a = runif(1, min = 0, max = 2)
  b = runif(1, min = 0, max = 2)
  x_comp = a + b * x
  model_comp = glm(y ~ x_comp,
                  family = binomial(link = 'logit'))

  # Calculando log-verossimilhanças

  loglik_0 = logLik(model_0)[1]
  loglik_1 = logLik(model_1)[1]
  loglik_rand = logLik(model_rand)[1]
  loglik_comp = logLik(model_comp)[1]

  loglik_0_high_pr = mpfr(loglik_0, 128)

```

```

loglik_1_high_pr      = mpfr(loglik_1, 128)
loglik_rand_high_pr   = mpfr(loglik_rand, 128)
loglik_comp_high_pr   = mpfr(loglik_comp, 128)

# Calculando coeficientes de determinacao:

# Soma de quadrados (R2_MQ)

R2_MQ = 1 - sum((y-model_1$fitted.values) ^ 2) / sum((y-mean(y)) ^ 2)
R2_MQ_rand = 1 - sum((y-model_rand$fitted.values) ^ 2) / sum((y-mean(y)) ^ 2)
R2_MQ_comp = 1 - sum((y-model_comp$fitted.values) ^ 2) / sum((y-mean(y)) ^ 2)

# McFadden (R2_MF)

R2_MF = 1 - loglik_1 / loglik_0
R2_MF_rand = 1 - loglik_rand / loglik_0
R2_MF_comp = 1 - loglik_comp / loglik_0

# Aldrich e Nelson (R2_AN)

R2_AN = (- 2 * loglik_0 + 2 * loglik_1) / (n - 2 * loglik_0 + 2 * loglik_1)
R2_AN_rand = (- 2 * loglik_0 + 2 * loglik_rand) /
  (n - 2 * loglik_0 + 2 * loglik_rand)
R2_AN_comp = (- 2 * loglik_0 + 2 * loglik_comp) /
  (n - 2 * loglik_0 + 2 * loglik_comp)

# Hagle e Mitchell (R2_HM)

R2_HM = R2_AN / ((- 2 * loglik_0) / (n - 2 * loglik_0))
R2_HM_rand = R2_AN_rand / ((- 2 * loglik_0) / (n - 2 * loglik_0))
R2_HM_comp = R2_AN_comp / ((- 2 * loglik_0) / (n - 2 * loglik_0))

# Maddala (R2_MA)

R2_MA = 1 - (exp(loglik_0_high_pr - loglik_1_high_pr)) ^ (2/n)
R2_MA_rand = 1 - (exp(loglik_0_high_pr - loglik_rand_high_pr)) ^ (2/n)
R2_MA_comp = 1 - (exp(loglik_0_high_pr - loglik_comp_high_pr)) ^ (2/n)

R2_MA = as.double(R2_MA)
R2_MA_rand = as.double(R2_MA_rand)
R2_MA_comp = as.double(R2_MA_comp)

# Nagelkerke (R2_N)

R2_N = R2_MA / (1 - exp(loglik_0_high_pr) ^ (2/n))
R2_N_rand = R2_MA_rand / (1 - exp(loglik_0_high_pr) ^ (2/n))
R2_N_comp = R2_MA_comp / (1 - exp(loglik_0_high_pr) ^ (2/n))

R2_N = as.double(R2_N)
R2_N_rand = as.double(R2_N_rand)
R2_N_comp = as.double(R2_N_comp)

# Pearson (R2_P)

R2_P = cor(y, model_1$fitted.values - mean(y),
  method = "pearson") ^ 2
R2_P_rand = cor(y, model_rand$fitted.values - mean(y),
  method = "pearson") ^ 2
R2_P_comp = cor(y, model_comp$fitted.values - mean(y),
  method = "pearson") ^ 2

# Spearman (R2_S)

```

```

R2_S = cor(y, model_1$fitted.values - mean(y),
           method = "spearman") ^ 2
R2_S_rand = cor(y, model_rand$fitted.values - mean(y),
                method = "spearman") ^ 2
R2_S_comp = cor(y, model_comp$fitted.values - mean(y),
                method = "spearman") ^ 2

# Kendall (R2_K)

R2_K = cor(y, model_1$fitted.values - mean(y),
           method = "kendall") ^ 2
R2_K_rand = cor(y, model_rand$fitted.values - mean(y),
                 method = "kendall") ^ 2
R2_K_comp = cor(y, model_comp$fitted.values - mean(y),
                 method = "kendall") ^ 2

# Gini (R2_G)

R2_G = 1 - ((sum(model_1$fitted.values *
                 (1 - model_1$fitted.values))) / (n * (sum(y)/n) * (1 - sum(y)/n)))
R2_G_rand = 1 - ((sum(model_rand$fitted.values *
                       (1 - model_rand$fitted.values))) / (n * (sum(y)/n) * (1 - sum(y)/n)))
R2_G_comp = 1 - ((sum(model_comp$fitted.values *
                       (1 - model_comp$fitted.values))) / (n * (sum(y)/n) * (1 - sum(y)/n)))

# Bura and Gastwirth (R2_BG)

x_s = (log((sum(y)/n) / (1 - sum(y)/n)) -
       as.numeric(model_1$coefficients[1])) / as.numeric(model_1$coefficients[2])
R2_BG = abs((2 * ((sum(y)/n) * (1/n) * sum(x <= x_s) -
              sum(model_1$fitted.values[x <= x_s]) / n)) /
            (2 * (sum(y)/n) * (1 - sum(y)/n)))

x_s_rand = (log((sum(y)/n) / (1 - sum(y)/n)) -
            as.numeric(model_rand$coefficients[1])) /
            as.numeric(model_rand$coefficients[2])
x_best_predictor = model_rand$coefficients[1] +
                  model_rand$coefficients[2] * x +
                  model_rand$coefficients[3] * x_rand
R2_BG_rand = abs((2 * ((sum(y)/n) * (1/n) * sum(x_best_predictor <= x_s_rand) -
              sum(model_rand$fitted.values[x_best_predictor <= x_s_rand]) / n)) /
            (2 * (sum(y)/n) * (1 - sum(y)/n)))

x_s_comp = (log((sum(y)/n) / (1 - sum(y)/n)) -
            as.numeric(model_comp$coefficients[1])) /
            as.numeric(model_comp$coefficients[2])
R2_BG_comp = abs((2 * ((sum(y)/n) * (1/n) * sum(x_comp <= x_s_comp) -
              sum(model_comp$fitted.values[x_comp <= x_s_comp]) / n)) /
            (2 * (sum(y)/n) * (1 - sum(y)/n)))

# Tjur (R2_T)

R2_T = mean(model_1$fitted.values[y == 1]) -
       mean(model_1$fitted.values[y == 0])
R2_T_rand = mean(model_rand$fitted.values[y == 1]) -
            mean(model_rand$fitted.values[y == 0])
R2_T_comp = mean(model_comp$fitted.values[y == 1]) -
            mean(model_comp$fitted.values[y == 0])

# Calculando valores preditos para a variavel resposta dicotomica

cut = cutpointr(x = model_1$fitted.values,
                class = model_1$y,

```

```

        method = maximize_metric,
        metric = youden,
        pos_class = 1,
        direction = ">=")$optimal_cutpoint
model_1$y_pred = ifelse (model_1$fitted.values < cut, 0, 1)

# Calculando indices de acuracia preditiva

n_right_1 = as.numeric(length(model_1$y[model_1$y_pred == 1 & model_1$y == 1]))
n_right_0 = as.numeric(length(model_1$y[model_1$y_pred == 0 & model_1$y == 0]))

accuracy = (n_right_1 + n_right_0) / n
sensitivity = n_right_1 / sum(y)
specificity = n_right_0 / (n-sum(y))
KS = sensitivity + specificity - 1

ROCit_obj = rocit(model_1$fitted.values, y)
AUC = ROCit_obj$AUC
gini = (AUC - 0.5) / 0.5

base_accuracy = (sum(y)/n)^2 + (1 - sum(y)/n)^2
base_sensitivity = (sum(y)/n)^2
base_specificity = (1 - sum(y)/n)^2

# Consolidando os resultados

results[i0, "scenario"] = scenario
results[i0, "rep"] = rep
results[i0, "n"] = n
results[i0, "x"] = x_labels[i2]
results[i0, "overall_p_pre"] = overall_p_pre
results[i0, "overall_p_pos"] = sum(y)/n
results[i0, "perc_overlap_pre"] = perc_overlap_pre
results[i0, "perc_overlap_pos"] = perc_overlap_pos
results[i0, "success_fit_pre"] = success_fit_pre
results[i0, "success_fit_pos"] = success_fit_pos

results[i0, "b0"] = model_1$coefficients[1]
results[i0, "b1"] = model_1$coefficients[2]
results[i0, "loglik_0"] = loglik_0
results[i0, "loglik_1"] = loglik_1

results[i0, "n_right"] = n_right_0 + n_right_1
results[i0, "accuracy"] = accuracy
results[i0, "base_accuracy"] = base_accuracy

results[i0, "n_1"] = sum(y)
results[i0, "n_right_1"] = n_right_1
results[i0, "sensitivity"] = sensitivity
results[i0, "base_sensitivity"] = sum(y)/n

results[i0, "n_0"] = n-sum(y)
results[i0, "n_right_0"] = n_right_0
results[i0, "specificity"] = specificity
results[i0, "base_specificity"] = 1 - sum(y)/n

results[i0, "KS"] = KS
results[i0, "AUC"] = AUC
results[i0, "gini"] = gini

results[i0, "R2_MQ"] = R2_MQ
results[i0, "R2_MF"] = R2_MF
results[i0, "R2_AN"] = R2_AN

```

```

results[i0, "R2_HM"] = R2_HM
results[i0, "R2_MA"] = R2_MA
results[i0, "R2_N"] = R2_N
results[i0, "R2_P"] = R2_P
results[i0, "R2_S"] = R2_S
results[i0, "R2_K"] = R2_K
results[i0, "R2_G"] = R2_G
results[i0, "R2_BG"] = R2_BG
results[i0, "R2_T"] = R2_T

results[i0, "R2_MQ_dif_rand"] = R2_MQ_rand - R2_MQ
results[i0, "R2_MF_dif_rand"] = R2_MF_rand - R2_MF
results[i0, "R2_AN_dif_rand"] = R2_AN_rand - R2_AN
results[i0, "R2_HM_dif_rand"] = R2_HM_rand - R2_HM
results[i0, "R2_MA_dif_rand"] = R2_MA_rand - R2_MA
results[i0, "R2_N_dif_rand"] = R2_N_rand - R2_N
results[i0, "R2_P_dif_rand"] = R2_P_rand - R2_P
results[i0, "R2_S_dif_rand"] = R2_S_rand - R2_S
results[i0, "R2_K_dif_rand"] = R2_K_rand - R2_K
results[i0, "R2_G_dif_rand"] = R2_G_rand - R2_G
results[i0, "R2_BG_dif_rand"] = R2_BG_rand - R2_BG
results[i0, "R2_T_dif_rand"] = R2_T_rand - R2_T

results[i0, "R2_MQ_dif_comp"] = R2_MQ_comp - R2_MQ
results[i0, "R2_MF_dif_comp"] = R2_MF_comp - R2_MF
results[i0, "R2_AN_dif_comp"] = R2_AN_comp - R2_AN
results[i0, "R2_HM_dif_comp"] = R2_HM_comp - R2_HM
results[i0, "R2_MA_dif_comp"] = R2_MA_comp - R2_MA
results[i0, "R2_N_dif_comp"] = R2_N_comp - R2_N
results[i0, "R2_P_dif_comp"] = R2_P_comp - R2_P
results[i0, "R2_S_dif_comp"] = R2_S_comp - R2_S
results[i0, "R2_K_dif_comp"] = R2_K_comp - R2_K
results[i0, "R2_G_dif_comp"] = R2_G_comp - R2_G
results[i0, "R2_BG_dif_comp"] = R2_BG_comp - R2_BG
results[i0, "R2_T_dif_comp"] = R2_T_comp - R2_T

results[i0, "R2_MF_adj"] = 1 - (loglik_1 - 1) / loglik_0
results[i0, "R2_MF_adj_rand"] = 1 - (loglik_rand - 2) / loglik_0
results[i0, "R2_MF_adj_dif_rand"] = (1 - (loglik_rand - 2) / loglik_0) -
    (1 - (loglik_1 - 1) / loglik_0)
    }
  }
}

print(paste (format(i0, scientific = F), "de",
              format(length(n_values) * length(x_values) * repetitions,
                    scientific = F), "repeticoes concluidas!"))

# Iterando
i0 = i0 + 1

}

scenario = scenario + 1
}
}
# Base de resultados final: armazenada no objeto "results"

```

Referências Bibliográficas

- Agresti(2002)** Alan Agresti. *Categorical Data Analysis*. Wiley, 2^a edição. Citado na pág. [12](#)
- Aldrich e Cnudde(1975)** John H. Aldrich e Charles F. Cnudde. Probing the bounds of conventional wisdom a comparison of regression, probit and discriminant analysis. *American Journal of Political Science*, 19:571–608. Citado na pág. [10](#)
- Aldrich e Nelson(1984)** John H. Aldrich e Forrest D. Nelson. *Linear Probability, Logit, and Probit Models*. Sage University Press, 1^a edição. Citado na pág. [33](#), [34](#), [55](#)
- Ben-Akiva e Lerman(1985)** Moshe Ben-Akiva e Steven R. Lerman. *Discrete Choice Analysis: Theory and Application to Travel Demand*. MIT Press, 1^a edição. Citado na pág. [32](#), [75](#)
- Bishop(2006)** Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 1^a edição. Citado na pág. [14](#)
- Bura e Gastwirth(2001)** Efstathia Bura e Joseph L. Gastwirth. The binary regression quantile plot: Assessing the importance of predictors in binary regression visually. *Biometrical Journal*, 1:5–21. Citado na pág. [ii](#), [iii](#), [40](#), [41](#), [42](#), [52](#), [55](#)
- Cameron e Windmeijer(1997)** Adrian C. Cameron e Frank A. G. Windmeijer. An r-squared measure of goodness of fit for some common nonlinear regression models. *Journal of Econometrics*, 77:329–342. Citado na pág. [27](#), [47](#)
- Chawla et al.(2002)** N. V. Chawla, K. W. Bowyer, L. O. Hall e W. P. Kegelmeyer. Smote: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16:321–357. Citado na pág. [17](#)
- Chicco e Jurman(2020)** Davide Chicco e Giuseppe Jurman. Comparison of the predicted and observed secondary structure of t4 phage lysozyme. *The Advantages of the Matthews Correlation Coefficient (MCC) Over F1 Score and Accuracy in Binary Classification Evaluation*, 21:1–13. Citado na pág. [24](#)
- Chinchor(1992)** Nancy Chinchor. Muc-4 evaluation metrics. Em *MUC4 '92: Proceedings of the 4th Conference on Message Understanding*, páginas 22–29. Citado na pág. [22](#)
- Cichosz(2015)** Pawel Cichosz. *Data Mining Algorithms: Explained Using R*. Wiley, 1^a edição. Citado na pág. [17](#)

- Cleveland e Devlin(1988)** William S. Cleveland e Susan J. Devlin. Locally weighted regression - an approach to regression analysis by local fitting. *Journal of the American Statistical Association*, 83:596–610. Citado na pág. [63](#)
- Cohen(1960)** Jacob Cohen. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20 (1):37–46. Citado na pág. [23](#)
- Cox e Snell(1989)** David R. Cox e E. Joyce Snell. *Analysis of Binary Data*. Chapman and Hall, 2ª edição. Citado na pág. [32](#), [34](#)
- Cox e Wermuth(1992)** David R. Cox e Nanny Wermuth. A comment on the coefficient of determination for binary responses. *The American Statistician*, 46:1–4. Citado na pág. [2](#), [10](#), [11](#), [35](#), [47](#)
- Cramer(2002)** Jan S. Cramer. The origins of logistic regression. *Universidade de Amsterdã e Instituto Tinbergen*. Citado na pág. [1](#)
- DeMaris(1992)** Alfred DeMaris. *Logit Modeling: Practical Applications*. Sage, 1ª edição. Citado na pág. [30](#), [31](#), [32](#), [33](#)
- Dhrymes(1986)** Phoebus J. Dhrymes. *Limited Dependent Variables*. In: *Handbook of Econometrics*, vol. 3. North-Holland, 1ª edição. Citado na pág. [30](#), [32](#)
- Dinov(2018)** Ivo D. Dinov. *Data Science and Predictive Analytics: Biomedical and Health Applications Using R*. Springer, 1ª edição. Citado na pág. [15](#)
- Efron(1978)** Bradley Efron. Regression and anova with zero-one data: Measures of residual variation. *Journal of the American Statistical Association*, 73:113–121. Citado na pág. [9](#)
- Fang e Chen(2019)** Fang Fang e Yuanyuan Chen. A new approach for credit scoring by directly maximizing the kolmogorov-smirnov statistic. *Computational Statistics and Data Analysis*, 133: 180–194. Citado na pág. [21](#)
- Farhat(2003)** Cecilia Aparecida Vaiano Farhat. *Análise de Diagnóstico em Regressão Logística*. Dissertação de Mestrado, Instituto de Matemática e Estatística da Universidade de São Paulo, Brasil. Citado na pág. [2](#)
- Gini(1912)** Corrado Gini. *Variabilità e Mutabilità: Contributo allo Studio delle Distribuzioni e delle Relazioni Statistiche*. Cuppini, 1ª edição. Citado na pág. [39](#)
- Goodman e Kruskal(1954)** L. A. Goodman e W. H. Kruskal. Measures of association for cross classifications. *Journal of the American Statistical Association*, 49:732–764. Citado na pág. [38](#)
- Haberman(1982)** S. J. Haberman. Analysis of dispersion of multinomial responses. *Journal of the American Statistical Association*, 77:568–580. Citado na pág. [38](#), [39](#), [55](#)
- Hagle e Mitchell(1992)** Timothy M. Hagle e Glenn E. Mitchell. Goodness-of-fit measures for probit and logit. *American Journal of Political Science*, 36:762–784. Citado na pág. [2](#), [30](#), [31](#), [33](#), [34](#), [47](#), [55](#), [58](#)

- Hand(2012)** David J. Hand. Assessing the performance of classification methods. *International Statistical Review*, 80:400–414. Citado na pág. [ii](#), [iii](#), [12](#), [20](#), [21](#), [22](#)
- Hand e Christen(2018)** David J. Hand e Peter Christen. A note on using the f-measure for evaluating record linkage algorithms. *Statistics and Computing*, 28:539–547. Citado na pág. [22](#)
- Hastie et al.(2008)** Trevor Hastie, Robert Tibshirani e Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference and Prediction*. Springer, 2^a edição. Citado na pág. [12](#), [15](#)
- Hauser(1977)** John R. Hauser. Testing the accuracy, usefulness and significance of probabilistic choice models: An information theoretic approach. *Operations Research*, 26:406–421. Citado na pág. [33](#)
- Hosmer et al.(2013)** David W. Hosmer, Stanley Lemeshow e Rodney X. Sturdivant. *Applied Logistic Regression*. Wiley, 3^a edição. Citado na pág. [1](#), [4](#), [12](#), [15](#), [18](#), [20](#), [37](#), [55](#)
- Hu et al.(2006)** Bo Hu, Mari Palta e Jun Shao. Properties of r^2 statistics for logistic regression. *Statistics in Medicine*, 25:1383–1395. Citado na pág. [49](#)
- James et al.(2013)** Gareth James, Daniela Witten, Trevor Hastie e Robert Tibshirani. *An Introduction to Statistical Learning with Applications in R*. Springer, 1^a edição. Citado na pág. [15](#)
- Knoke e Burke(1980)** David Knoke e Peter J. Burke. *Log-Linear Models*. Sage, 1^a edição. Citado na pág. [30](#)
- Korn e Simon(1990)** E. L. Korn e R. Simon. Measures of explained variation for survival data. *Statistics in Medicine*, 9:487–503. Citado na pág. [37](#)
- Kuhn e Johnson(2013)** Max Kuhn e Kjell Johnson. *Applied Predictive Modeling*. Springer, 1^a edição. Citado na pág. [16](#)
- Kvalseth(1985)** Tarald O. Kvalseth. Cautionary note about r^2 . *The American Statistician*, 39:279–285. Citado na pág. [34](#), [45](#), [46](#)
- Lave(1970)** Charles A. Lave. The demand for urban mass transportation. *The Review of Economics and Statistics*, 37:320–323. Citado na pág. [9](#)
- Maddala(1983)** Gangadharrao S. Maddala. *Limited-dependent and Qualitative Variables in Econometrics*. Cambridge, 1^a edição. Citado na pág. [34](#), [35](#), [36](#), [55](#)
- Magee(1990)** Lonnie Magee. r^2 measures based on wald and likelihood ratio joint significance tests. *The American Statistician*, 44:250–253. Citado na pág. [34](#), [35](#), [45](#)
- Matthews(1975)** Brian W. Matthews. Comparison of the predicted and observed secondary structure of t4 phage lysozyme. *Biochimica et Biophysica Acta*, 405:442–451. Citado na pág. [24](#)
- McFadden(1974)** Daniel McFadden. *Conditional Logit Analysis of Qualitative Choice Behavior*. In: P. Zarembka (ed.), *Frontiers in Econometrics*. Academic Press, 1^a edição. Citado na pág. [ii](#), [iii](#), [30](#), [32](#), [39](#), [55](#), [72](#), [75](#)

- Menard(2000)** Scott Menard. Coefficients of determination for multiple logistic regression analysis. *The American Statistician*, 54:17–24. Citado na pág. [2](#), [10](#), [12](#), [18](#), [26](#), [30](#), [31](#), [32](#), [33](#), [34](#), [35](#), [36](#), [45](#), [47](#), [75](#)
- Menard(2002)** Scott Menard. *Applied Logistic Regression Analysis*. Sage, 2^a edição. Citado na pág. [11](#)
- Mittlböck e Schemper(1996)** Martina Mittlböck e Michael Schemper. Explained variation for logistic regression. *Statistics in Medicine*, 15. Citado na pág. [2](#), [11](#), [36](#), [37](#), [39](#), [40](#), [46](#), [47](#), [49](#), [55](#), [58](#), [59](#), [60](#), [62](#), [73](#), [74](#), [75](#)
- Nagelkerke(1991)** N. J. D. Nagelkerke. A note on a general definition of the coefficient of determination. *Biometrika*, 78. Citado na pág. [35](#), [36](#), [55](#), [58](#)
- Nakamura(2013)** Karina Gernhardt Nakamura. *Multicolinearidade em Modelos de Regressão Logística*. Dissertação de Mestrado, Instituto de Matemática e Estatística da Universidade de São Paulo, Brasil. Citado na pág. [2](#)
- Nakas e Yiannoutsos(2004)** Christos Nakas e Constantin Yiannoutsos. Ordered multiple-class roc analysis with continuous measurements. *Statistics in Medicine*, 23:3437–3449. Citado na pág. [20](#)
- Neter et al.(2004)** John Neter, Michael H. Kutner e Chris J. Nachtsheim. *Applied Linear Statistical Models*. McGraw-Hill/Irwin, 5^a edição. Citado na pág. [1](#)
- Okura(2008)** Roberta Irie Sumi Okura. *Modelos de Regressão para Variáveis Categóricas Ordinais com Aplicações ao Problema de Classificação*. Dissertação de Mestrado, Instituto de Matemática e Estatística da Universidade de São Paulo, Brasil. Citado na pág. [2](#)
- Powers(2011)** David Powers. Evaluation from precision, recall and f-measure to roc, informedness, markedness correlation. *Journal of Machine Learning Technologies*, 2:37–63. Citado na pág. [22](#), [23](#), [24](#)
- Ryan(2008)** Thomas P. Ryan. *Modern Regression Methods*. Wiley, 2^a edição. Citado na pág. [2](#), [9](#), [10](#), [20](#), [35](#), [52](#), [54](#)
- Schemper(2003)** M. Schemper. Predictive accuracy and explained variation. *Statistics in Medicine*, 22:2299–2308. Citado na pág. [26](#)
- Scurfield(1996)** Brian Scurfield. Multiple-event forced-choice tasks in the theory of signal detectability. *Journal of Mathematical Psychology*, 40:253–269. Citado na pág. [20](#)
- Shannon(1948)** C. E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27:379–423, 623–656. Citado na pág. [38](#)
- Somers(1962)** R. H. Somers. A new asymmetric measure of association for ordinal variables. *American Sociological Review*, 27:799–811. Citado na pág. [38](#)
- Tjur(2009)** Tue Tjur. Coefficients of determination in logistic regression models - a new proposal: The coefficient of discrimination. *The American Statistician*, 63:366–372. Citado na pág. [35](#), [42](#), [44](#), [45](#), [47](#), [55](#)

- Tufféry(2011)** Stéphane Tufféry. *Data Mining and Statistics for Decision Making*. Wiley, 1ª edição. Citado na pág. [16](#)
- van Belle et al.(2006)** Chengjie Xiong and Gerald van Belle, J. Philip Miller e John Morris. Measuring and estimating diagnostic accuracy when there are three ordinal diagnostic groups. *The American Statistician*, 25:1251–1273. Citado na pág. [20](#)
- Veall e Zimmermann(1994)** Michael R. Veall e Klaus F. Zimmermann. Evaluating pseudo- r^2 's for binary probit models. *Quality and Quantity*, 28:151–164. Citado na pág. [2](#), [33](#), [34](#), [47](#), [49](#)
- Youden(1950)** William J. Youden. Index for rating diagnostic tests. *Cancer*, 3:32–35. Citado na pág. [20](#)