Predições estatísticas para dados politômicos

Guaraci de Lima Requena

Tese apresentada
ao
Instituto de Matemática e Estatística
da
Universidade de São Paulo
para
obtenção do título
de
Doutor em Ciências

Programa: Estatística

Orientador: Prof. Dr. Carlos Alberto de Bragança Pereira

Coorientador: Prof. Dr. Adriano Polpo de Campos

Durante o desenvolvimento deste trabalho o autor recebeu auxílio financeiro da CAPES e CNPq

São Paulo, 5 de setembro 2018

Predições estatísticas para dados politômicos

Esta versão da tese contém as correções e alterações sugeridas pela Comissão Julgadora durante a defesa da versão original do trabalho, realizada em 17/08/2018. Uma cópia da versão original está disponível no Instituto de Matemática e Estatística da Universidade de São Paulo.

Comissão Julgadora:

- Prof. Dr. Carlos Alberto de Bragança Pereira (orientador) - IME-USP
- Prof. Dr. Adriano Polpo de Campos (coorientador) - UFSCar
- Prof^{a}. Dr^{a}. Luzia Aparecida Trinca - UNESP
- Prof^{a}. Dr^{a}. Juliana Belo Diniz - IPq-FMUSP
- Prof. Dr. Junior Barrera - IME-USP
- Prof. Dr. Erlandson Ferreira Saraiva - UFMS

Agradecimentos

Ao Prof. Dr. Carlos Alberto de Bragança Pereira – Professor Carlinhos – pela orientação, ensinamentos e confiança. Por me mostrar a importância da aplicação da estatística na prática, especialmente na área da saúde. Também, por me fazer ver que o estatístico é como um mago que faz previsões sobre o invisível. Certa vez o Prof. Sergio Wechsler me disse que bastava ficar perto do Carlinhos para aprender estatística. Hoje, entendo o que ele queria me dizer.

Ao Prof. Dr. Adriano Polpo, pela coorientação, sempre com opiniões construtivas, visando uma boa discussão, trazendo importantes reflexões sobre a estatística e suas aplicações. Também, pelo incentivo de cursar o doutorado no IME, me mostrando que as mudanças trazem experiência, e que experiência é fundamental na tomada de boas decisões. Agradeço ainda pela grande contribuição para a conclusão deste trabalho.

Ao meu pai, José Aparecido Requena, que desde o início dos meus estudos me deu apoio incomensurável, que sempre me mostrou a importância do saber e de ter uma visão crítica sobre as coisas. Por ter me ensinado, e me ensinar, com sua própria história. À minha mãe, Maria José de Lima Requena, pela confiança, dedicação e disponibilidade incondicionais. Por me mostrar que as verdadeiras conquistas são as que te fazem ser uma pessoa melhor. À minha irmã, Potira de Lima Requena, por compartilhar sua história comigo e por todo o afeto e apoio sempre presentes.

À minha namorada Isabela Bertolini Coelho, pelo companheirismo e carinho enormes, por toda a dedicação nestes anos de doutorado e por sonhar comigo os anos vindouros. Sendo mestre em estatística pelo IME, leu e revisou este texto, e diversas partes foram melhoradas a partir de suas sugestões. Obrigado por construir e compartilhar essa etapa, e por evoluir, junto comigo.

Ao PROTOC e seus membros, por terem compartilhado comigo diversos projetos, participando da minha formação como estatístico. Por me mostrarem a importância da ciência para a sociedade e a importância da estatística para a ciência. Em especial à Prof^{a}. Dr^{a}. Roseli Gedanke Shavitt, por ter confiado em meu trabalho e, também, por me apresentar bons projetos, sendo o primeiro deles o motivador para trabalhar com dados categóricos.

Aos amigos que sempre tornam a caminhada mais agradável. Ao Felippe Colombari, pela confraternidade, que desde os tempos de república esteve presente e por ter propiciado tantas discussões que trouxeram reflexões das mais diversas, que certamente compõem este trabalho. Aos IMEanos: Brian Melo, Helton Graziadei, Henrique Bolfarine, Jaime Curivil, Maicon Pinheiro e Renato Ciani pelas pizzas e por mostrarem que não se trata de colher os frutos, mas sim de fazer o vinho; e também ao Leandro Augusto Ferreira pelos cafés e parcerias.

Aos professores que participaram da minha formação. Em especial, à Prof.^{a} Dr.^{a} Adriana Barbosa Santos, ao Prof. Dr. Fernando Ferrari, ao Prof. Dr. Luiz Carlos Baida (IBILCE-UNESP) e ao Prof. Dr. Carlos Diniz (UFSCar) que me apresentaram a estatística e que me incentivaram por ela trilhar.

Ao IME, seus professores e seus funcionários que formam, e fazem funcionar, toda a estrutura

na qual pude me formar. Também agradeço à Faculdade de Saúde Pública da USP, especialmente aos funcionários da biblioteca, por proporcionarem um ótimo lugar aos estudos.

Aos membros da comissão julgadora deste trabalho, Profª. Drª. Luzia Aparecida Trinca, Profª. Drª. Juliana Belo Diniz, Prof. Dr. Junior Barrera e Prof. Dr. Erlandson Ferreira Saraiva, pelas sugestões, críticas construtivas e por propiciar uma boa discussão, que incorporo nesta versão final.

À CAPES e ao CNPq pelo suporte financeiro.

Resumo

REQUENA, G. de L. **Predições estatísticas para dados politômicos**. 2018. 76 f. Tese (Doutorado) - Instituto de Matemática e Estatística, Universidade de São Paulo, São Paulo, 2018.

Este trabalho generaliza a partição da distribuição de Bernoulli multivariada em distribuições de Bernoulli e como esta partição leva a um modelo de regressão e a um classificador para dados politômicos.

Como ponto de partida, desejamos explicitar a função de ligação para os modelos de regressão multinomial e escrevê-la a partir de funções de distribuição, como feito no caso binomial, a fim de flexibilizá-la para além da logito usual. Para isso, estudamos as fatorações da Bernoulli multivariada em Bernoullis, bem como a multinomial em binomiais, a fim de explicitar como as funções de distribuição podem desempenhar um papel na ligação entre o espaço das covariáveis e o vetor de probabilidades. Basu e Pereira (1982) exploram tais fatorações em um problema de não resposta e Pereira e Stern (2008) as generalizam para uma classe de fatorações.

Este trabalho propõe uma simplificação tanto da regressão multinomial – agregando a flexibilidade do caso binomial –, quanto da classificação politômica, no sentido de decompor o problema politômico em dicotômicos através da generalização da classe de fatorações. Um problema computacional surge pois tal classe pode ter um número muito grande de elementos distintos de acordo com o número de categorias e, assim, duas propostas são feitas para buscar uma que minimiza os riscos de classificação binomial envolvidos, passo-a-passo.

A motivação para este trabalho é apresentada a fim de se estudar as performances de tais modelos de regressão e classificadores. Partimos de um problema da área médica, mais especificamente em transtorno obsessivo-compulsivo, em que desejamos classificar um indivíduo a fim de obter um fenótipo mais puro de tal transtorno e de modelá-lo a fim de buscar as covariáveis que estão relacionadas com tal fenótipo, a partir de um conjunto de dados reais.

Palavras-chave: classificação, regressão multinomial, dados categóricos, fatoração, transtorno obsessivo-compulsivo.

Abstract

REQUENA, G. de L. **Statistical predictions for polytomous data**. 2018. 76 f. Tese (Doutorado) - Instituto de Matemática e Estatística, Universidade de São Paulo, São Paulo, 2018.

This work explores a partition of the multivariate Bernoulli distribution in Bernoulli distributions and how this partition leads to a regression model and to a classifier for polytomous data.

As starting point, we want to make explicit the link function for multinomial regression models and write it from distribution functions, as in the binomial case, in order to flexibilize it beyond the usual logit. For that, we study the factorizations of the multivariate Bernoulli in Bernoullis, as well as the multinomial in binomials, in order to make explicit as the distribution functions may play a role in the linkage between the space of covariates and the vector of probabilities. Basu and Pereira (1982) explore these factorizations in a nonresponse problem and Pereira and Stern (2008) generalize them to a class of factorizations.

Thus, this work proposes a simplification of the multinomial regression – adding the flexibility from the binomial case –, and of the polytomous classification, decomposing de polytomous problem in dichotomous through the generalization of the class of factorizations. At this point, a computational problem arises because the amount of factorizations may be very large according to the number of categories and then we propose two approaches to seek a factorization that minimize the involved binomial classification risks, step-by-step.

The motivation for this work is presented in order to study the performance of such regression models and classifiers. We start from a medical problem, more precisely in obsessive-compulsive disorder, in which we want to classify a patient in order to get a more pure phenotype of such disorder and model it in order to seek the related covariates, from a real dataset.

Keywords: classification, multinomial regression, categorical data, factorization, obsessive-compulsive disorder.

Sumário

Lista de Figuras

Lista de Tabelas

Capítulo 1

Introdução

Há na literatura sobre modelos de regressão para dados categóricos um grande esforço com relação a flexibilizações e generalizações do modelo de regressão binomial logístico (ver Diniz, 2015, para um panorama sobre o assunto), contudo pouco se tem feito no mesmo sentido para dados multinomiais. A função de ligação no contexto multinomial logístico usual, com categoria de referência, não é explicitamente dada através de uma função de distribuição, como no caso binomial via distribuição logística padrão.

Uma solução vai no sentido de reparametrizar o problema multinomial a fim de decompô-lo em binomiais. Basu e Pereira (1982), frente a um problema de não resposta em uma pesquisa que busca investigar uma resposta binária (A_1 ou A_2), exploram tal fatoração (ver também Pereira, 1980). Para isso, escrevem as proporções populacionais como

	R	R'	
A_1	p_{11}	p_{12}	π
A_2	p_{21}	p_{22}	$1-\pi$
	θ_1	$1-\theta_1$	1

em que R é a categoria dos respondedores e R' dos não respondedores. O interesse está em fazer inferências em π. Assim, apresentam a verossimilhança de um modelo trinomial (A_1, A_2 e R') como

$$L(p_{11}, p_{21}, \theta_1 | x_1, x_2, n_2) = p_{11}^{x_1} p_{21}^{x_2} (1-\theta_1)^{n_2},$$

em que, para um indivíduo, $n_2 = 1$ se ele é não respondedor, $x_1 = 1$ se ele é respondedor e pertence a categoria A_1 e $x_2 = 1$ se ele é respondedor e pertence a categoria A_2, uma vez que não há observações especificamente em $R' \cap A_1$ ou $R' \cap A_2$. Dessa forma, para resolver o problema do ponto de vista Bayesiano, propõem a seguinte reparametrização:

$$\theta_1 = p_{11} + p_{21}, \quad \theta_2 = \tfrac{p_{11}}{\theta_1},$$

com a transformação reversa sendo

$$p_{11} = \theta_1 \theta_2, \quad p_{21} = \theta_1(1-\theta_2)$$

e, com isso, reescrevem a verossimilhança como o produto de duas binomiais:

$$L(\theta_1, \theta_2 | n_1, x_1) = \theta_1^{n_1}(1-\theta_1)^{n_2}\theta_2^{x_1}(1-\theta_2)^{x_2},$$

em que $n_1 = x_1 + x_2$, fatorando a verossimilhança trinomial em 2 binomiais, sendo o fator $\theta_1^{n_1}(1-\theta_1)^{n_2}$ referente a R e R' e o fator $\theta_2^{x_1}(1-\theta_2)^{x_2}$ referente a A_1 e A_2, dado R.

Baseando-se nesta ideia, Aitchison *et al.* (2004) afirmam que não há nenhuma forma normal (se referindo a função de ligação probito) diretamente comparável ao caso binomial e discutem a mesma reparametrização de Basu e Pereira (1982) para o contexto de regressão, escrevendo

$$\begin{aligned} p_1 &= \Phi(\eta_1) \\ p_2 &= (1-\Phi(\eta_1))\Phi(\eta_2) \\ p_3 &= (1-\Phi(\eta_1))(1-\Phi(\eta_2)), \end{aligned}$$

em que η_1 e η_2 são preditores lineares reais e Φ é a função de distribuição normal padrão. Para elucidar a ideia, Aitchison *et al.* (2004) sugerem que pode-se considerar a categoria 1 como sendo "não doente" e as categorias 2 e 3 como sendo dois tipos diferentes de doença (mutuamente exclusivas). E assim, modela-se a probabilidade de ser "não doente" a partir de uma probito, considerando a junção das categorias 2 e 3 como "doente" e, em seguida, modela-se a probabilidade para a categoria 2 a partir de uma probito, dado que é "doente".

Pereira e Stern (2008) generalizam tal reparametrização para uma classe específica para um número qualquer de categorias, digamos D, reparametrizando o vetor de parâmetros $\boldsymbol{p}$ por $D-1$ parâmetros de distribuições Bernloullis, θ_j, $j = 1, 2, ..., D-1$:

$$\begin{aligned} p_1 &= \theta_1 \\ p_k &= \theta_k \prod_{l=1}^{k-1}(1-\theta_l), \quad k = 2, 3, ..., D-1 \\ p_D &= \prod_{l=1}^{D-1}(1-\theta_l). \end{aligned}$$

Neste sentido, Caron e Polpo (2009) – que propõem a função de ligação Weibull para o caso binomial – sugerem sua aplicação em regressão multinomial utilizando tal reparametrização.

Utilizando este contexto como motivação estatística para esta tese, ao invés de explorar o comportamento das diversas funções de ligação binomiais para o caso multinomial, através das reparametrizações já propostas, exploramos por outro lado como essa classe de reparametrizações pode ser generalizada. A partir daí, a proposta se dá ao simplificar o problema politômico, seja de regressão e/ou classificação, por problemas dicotômicos.

É preciso, então, como apresentamos no Capítulo 2, abordar o problema através das fatorações da distribuição de Bernoulli multivariada (por vezes chamada de distribuição categórica, ou simplesmente de multinomial de tamanho 1) em Bernoulli's, antes de abordá-lo com relação à regressão e à classificação. A partir disso, vemos que podemos exprimir um resultado politômico através de $D-1$ resultados dicotômicos. Também, que um vetor qualquer assumindo valores no simplex de ordem $D-1$ pode ser unicamente representado por $D-1$ escalares em $(0,1)$. Para explorar estes fatos, definimos um objeto que denominamos de bipartição total, na qual consiste em particionar

Ω em dois subconjuntos não vazios e prosseguir com esse processo de particionamento recursiva e exaustivamente.

Definida bipartição total, exploramos como ela induz uma fatoração da distribuição Bernoulli D-variada em $D-1$ Bernoulli's e então, no Capítulo 3, como tais fatorações levam a construções gerais de modelos de regressão multinomial, bem como de classificadores politômicos, a partir de regressões e classificadores binomiais. Há, contudo, um número bastante grande de diferentes bipartições totais, dependendo do número de categorias e, consequentemente, um número muito grande de modelos multinomiais que nos leva a um problema computacional: buscar uma bipartição total, dentre todas as possíveis, que induza a um "bom" – segundo alguma métrica especificada – modelo de regressão e a um "bom" classificador. Dessa forma, ainda no Capítulo 3, visando a minimização de riscos de classificação binomiais, propomos duas abordagens a partir das observações da variável resposta e das covariáveis.

Por fim, no Capítulo 4, para avaliar as performances deste estudo, em comparação com o modelo de regressão usual e com os classificadores por ele obtidos, aplicamos as abordagens em um conjunto de dados reais, que foi a motivação para a elaboração desta tese. Tal problema pertence às ciências médicas, mais especificamente ao Transtorno Obsessivo-Compulsivo (TOC). O TOC é considerado uma condição neurobiológica heterogênea com relação à apresentação dos sintomas nos indivíduos, podendo passar por sintomas relacionados à simetria ou ordem, bem como por agressão, de cunho sexual ou religioso, dentre outras. Há diversos instrumentos que buscam medir a gravidade do TOC e o mais utilizado dá um escore geral para obsessão e um escore geral para compulsão (Y-BOCS). Por outro lado, há um instrumento alternativo (DY-BOCS) que atribui escores diferentes para o que é denominado dimensões do TOC, resultando em fenótipos mais específicos. Muitos estudos que visam relacionar fatores genéticos com o TOC falham nessa busca e uma das razões se dá pela natureza heterogênea do fenótipo no TOC, uma vez que grande parte desses estudos acessa a gravidade do TOC via Y-BOCS. Assim, há diversos bancos de dados já construídos com informações de Y-BOCS e genética e um método que consiga atribuir informações com relação às dimensões do TOC a partir do Y-BOCS, assim como feito no DY-BOCS, pode ser muito útil nas buscas pela relação fenótipo-genótipo. Assim, podemos considerar que as características observadas via Y-BOCS são as covariáveis e o escore DY-BOCS para uma dimensão específica do TOC (com distribuição Bernoulli multivariada) é a variável resposta e então, a partir de uma amostra com ambas as informações, podemos propor uma regressão e classificação multinomial para lidar com o problema. O problema original, bem como os detalhes pertinentes com relação ao TOC e aos instrumentos de mensuração de gravidade, é apresentada em Shavitt *et al.* (2017).

No seu livro "Introdução à filosofia matemática", Bertrand Russell escreve que: "A matemática é um estudo que, quando partimos de suas partes mais conhecidas, pode ser continuado em uma de duas direções opostas. A direção mais conhecida é construtiva, rumo a uma complexidade gradualmente crescente: de números inteiros para frações, números reais, números complexos; de adição e multiplicação para diferenciação e integração, e adiante, para a matemática superior. A outra direção, menos conhecida, procede, por análise, rumo à abstração e à simplicidade lógica cada vez maiores; em vez de perguntar o que pode ser definido e deduzido do que é inicialmente suposto, perguntamos que ideias e princípios mais gerais podem ser encontrados, de acordo com o que o nosso ponto de partida pode ser definido ou deduzido." e penso que em Estatística há também esta distinção.

Capítulo 2

A distribuição de Bernoulli multivariada e suas fatorações

Neste capítulo exploramos a distribuição Bernoulli multivariada e suas fatorações em Bernoulli's. Vemos que um conjunto finito de categorias (espaço amostral) pode ser redefinido em uma classe de espaços amostrais com duas categorias de diversas formas distintas e que cada classe conduz a uma fatoração diferente, levando também a uma redefinição do espaço paramétrico, ou seja, a uma reparametrização do problema. A extensão para a distribuição multinomial de tamanho n também é explorada. Este capítulo é desenvolvido com base nos trabalhos: Basu e Pereira (1982), Pereira e Stern (2008) e Aitchison (1986), que exploram, respectivamente, a fatoração da distribuição de Bernoulli multivariada aplicada em um problema de não resposta em um contexto Bayesiano; distribuições discretas e suas propriedades e desenvolvimentos em notação vetorial; e o simplex, suas propriedades e partições.

2.1 A distribuição de Bernoulli multivariada

Seja W uma variável aleatória assumindo valores no conjunto de categorias $\Omega = \{0, 1\}$. Dizemos que W tem distribuição de Bernoulli com parâmetro θ se sua função de probabilidades (fp) pode ser escrita como

$$Pr(W = w|\theta) = \theta^w(1-\theta)^{1-w}, \quad 0 < \theta < 1. \tag{2.1}$$

Agora, seja Y variável aleatória assumindo valores no conjunto de categorias $\Omega = \{1, 2, ..., D\}$. Dizemos que Y tem distribuição de Bernoulli D-variada com parâmetro $\boldsymbol{p} = (p_1, p_2, ..., p_D)$ se sua fp pode ser representada por

$$Pr(Y = y|\boldsymbol{p}) = \prod_{k=1}^{D} p_k^{\mathbb{1}(y=k)}, \tag{2.2}$$

em que, dado que $y \in \Omega$,

$$\mathbb{1}(y = k) = \begin{cases} 1, & \text{se } y = k \\ 0, & \text{se } y \neq k \end{cases} \qquad k = 1, 2, ..., D.$$

É conveniente tratar as categorias por $1, 2, ..., D$, no entanto são consideradas rótulos nominais, e não numéricos. Assim sendo, apesar de tratamos $Y \in \Omega$, a distribuição Bernoulli D-variada pode

ser vista como referente a um vetor de indicadoras das categorias, ou seja $(\mathbb{1}(Y=1), \mathbb{1}(Y=2), ..., \mathbb{1}(Y=D))$. Em outras palavras, a Bernoulli multivariada é a multinomial de tamanho 1 (por vezes também denominada de distribuição categórica). Dessa forma, alternativamente, podemos definir o espaço amostral como sendo o conjunto de colunas da matriz identidade de ordem D e denotar o vetor aleatório por $\boldsymbol{Y} = (Y_1, Y_2, ..., Y_D)$, em que $Y_k = \mathbb{1}(Y = k)$. Assim a fp pode ser escrita por

$$Pr(\boldsymbol{Y} = \boldsymbol{y}|\boldsymbol{p}) = \prod_{k=1}^{D} p_k^{y_k}. \tag{2.3}$$

O parâmetro $\boldsymbol{p}$ é o vetor de probabilidades – ou proporções populacionais – para as D categorias e, dessa forma, deve obedecer as seguintes restrições: $p_k > 0, k = 1, 2, ..., D$ e $\sum_{k=1}^{D} p_k = 1$. Tais restrições definem o espaço paramétrico, denominado simplex de ordem $d = D - 1$, dado por

$$\mathscr{S}^d = \left\{ \boldsymbol{p} = (p_1, p_2, ..., p_D) \in \mathbb{R}^D; p_k > 0, \sum_{k=1}^{D} p_k = 1 \right\}.$$

A ordem é $d = D - 1$ pois $\mathscr{S}^d$ pode ser unicamente representado por um subvetor de ordem d de $\boldsymbol{p}$, digamos, sem perda de generalidade, $\boldsymbol{p}_{-D} = (p_1, p_2, ..., p_d)$, pelo fato de $p_D = 1 - \sum_{k=1}^{d} p_k$. Então podemos escrever $\mathscr{S}^d = \left\{ \boldsymbol{p}_{-D} \in \mathbb{R}^d; p_k > 0, \sum_{k=1}^{d} p_k < 1 \right\}$. Consideramos p_k diferente de 0 e 1 por questões algébricas, mas tais extensões são diretamente incorporadas na teoria.

Tomando a representação vetorial dada pela Equação (2.3), podemos definir o espaço amostral de forma análoga ao paramétrico por

$$\mathscr{I}^d = \left\{ \boldsymbol{y} = (y_1, y_2, ..., y_D); y_k \in \{0, 1\}, \sum_{k=1}^{D} y_k = 1 \right\},$$

e, assim definido geometricamente, $\mathscr{I}^d$ é o conjunto de vértices de $\mathscr{S}^d$, como representado na Figura 2.1. Ainda sob essa notação, temos que

$$E[\boldsymbol{Y}|\boldsymbol{p}] = \boldsymbol{p} \quad \text{e} \quad Cov[\boldsymbol{Y}|\boldsymbol{p}] = \begin{bmatrix} p_1(1-p_1) & -p_1p_2 & \cdots & -p_1p_D \\ -p_1p_2 & p_2(1-p_2) & \cdots & -p_2p_D \\ \vdots & \vdots & \ddots & \vdots \\ -p_1p_D & -p_2p_D & \cdots & p_D(1-p_D) \end{bmatrix}. \tag{2.4}$$

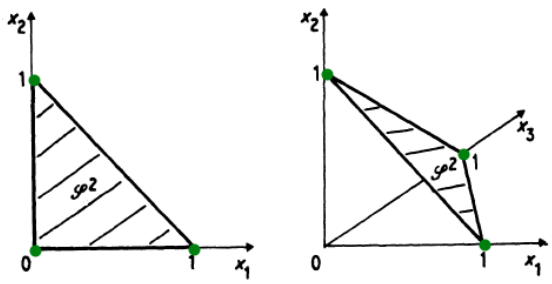


Figura 2.1: *Representações geométricas de $\mathscr{S}^2$ e $\mathscr{I}^2$ (Aitchison, 1986).*

Notação *Denotamos por $(Y|Y \in \Omega; \boldsymbol{p}) \sim BM_D(\boldsymbol{p})$ quando Y tem distribuição de Bernoulli mul-*

tivariada em $\Omega = \{1, 2, ..., D\}$ *com parâmetro* $\boldsymbol{p} = (p_1, p_2, ..., p_D)$ *(Equação (2.2)). Quando não houver ambiguidades sobre o conjunto de categorias considerado, denotamos por* $Y|\boldsymbol{p} \sim BM_D(\boldsymbol{p})$. *A distribuição Bernoulli usual com parâmetro* θ *é denotada por* $B(\theta)$ *(Equação (2.1)).*

2.2 Fatorações

A partir de Ω, podemos agregar categorias ou tomar um subconjunto delas e redefinir o espaço amostral, bem como redefinir a variável aleatória, assumindo valores neste espaço a partir de transformações em Y, o que leva também a uma reparametrização do problema, como tratam a definição e propriedades a seguir.

2.2.1 Bipartições e R-partições de Ω

Definição 2.1 *Seja* A *um subconjunto não vazio de* Ω, *ou seja,* A *um subconjunto de categorias. Denominamos por bipartição de* Ω *a classe* $\Lambda = \{A, A^c\}$, *em que* A^c *denota o conjunto complementar a* A. *Se* $A_1, A_2, ..., A_R$ *são subconjuntos não vazios formando uma partição de* Ω, *denominamos* $\Lambda = \{A_1, A_2, ..., A_R\}$ *por R-partição de* Ω.

Propriedade 2.1 *Sejam* $Y|\boldsymbol{p} \sim BM_D(\boldsymbol{p})$ *e* $\Lambda = \{A, A^c\}$ *uma bipartição de* Ω. *Defina*

$$W = \mathbb{1}(Y \in A) = \begin{cases} 1, & se\ Y \in A \\ 0, & se\ Y \in A^c \end{cases} \tag{2.5}$$

e também

$$q = Pr(Y \in A|\boldsymbol{p}) = \sum_{k \in A} p_k.$$

Então

$$W|q \sim B(q).$$

Propriedade 2.2 *Sejam* $Y|\boldsymbol{p} \sim BM_D(\boldsymbol{p})$ *e* $\Lambda = \{A_1, A_2, .., A_R\}$ *uma R-partição de* Ω. *Defina*

$$W = \begin{cases} 1, & se\ Y \in A_1 \\ 2, & se\ Y \in A_2 \\ \vdots & \\ R, & se\ Y \in A_R \end{cases} \tag{2.6}$$

e $\boldsymbol{q} = (q_1, q_2, ..., q_R)$, *em que*

$$q_l = Pr(Y \in A_l|\boldsymbol{p}) = \sum_{k \in A_l} p_k, \qquad l = 1, 2, ..., R.$$

Então $\boldsymbol{q} \in \mathscr{S}^{R-1}$ *e*

$$W|\boldsymbol{q} \sim BM_R(\boldsymbol{q}).$$

Propriedade 2.3 *Sejam* $Y|\boldsymbol{p} \sim BM_D(\boldsymbol{p})$ *e, sem perda de generalidade,* $A = \{1, 2, ..., M\} \subset \Omega$,

$\boldsymbol{p}_A = (p_1, p_2, ..., p_M)$ *e* $q = \sum_{k\in A} p_k$. *Então* $\frac{1}{q}\boldsymbol{p}_A = \left(\frac{p_1}{q}, \frac{p_2}{q}, ..., \frac{p_M}{q}\right) \in \mathscr{S}^{M-1}$ *e*

$$(Y|Y \in A; \boldsymbol{p}) \sim BM_M\left(\frac{1}{q}\boldsymbol{p}_A\right).$$

2.2.2 Uma classe de fatorações

As 3 propriedades anteriores nos permitem fatorar a distribuição de Bernoulli multivariada em Bernoulli's multivariadas de menores dimensões, como enunciado no teorema e corolários a seguir.

Teorema 2.1 *Sejam* $\Lambda = \{A, A^c\}$ *uma bipartição de* Ω *e* $q = \sum_{k\in A} p_k$, *em que, sem perda de generalidade,* $A = \{1, 2, ..., M\}$. *Considere também* $W = \mathbb{1}(Y \in A)$, *como na Equação (2.5). A fp de* $Y|\boldsymbol{p} \sim BM_D(\boldsymbol{p})$ *se fatora nas fp's de* $W|q \sim B(q)$, $(Y|Y \in A; \boldsymbol{p}_A) \sim BM_M\left(\frac{1}{q}\boldsymbol{p}_A\right)$ *e* $(Y|Y \in A^c; \boldsymbol{p}_{A^c}) \sim BM_{D-M}\left(\frac{1}{1-q}\boldsymbol{p}_{A^c}\right)$, *em que* $\boldsymbol{p}_A = (p_1, p_2, ..., p_M)$ *e* $\boldsymbol{p}_{A^c} = (p_{M+1,M+2}, ..., p_D)$.

Demostração *De fato, como podemos escrever* $\mathbb{1}(y \in A) = \sum_{k\in A} \mathbb{1}(y = k)$, *temos que*

$$\begin{aligned}
Pr(Y = y|\boldsymbol{p}) &= \prod_{k=1}^{D} p_k^{\mathbb{1}(y=k)} \\
&= \left(\frac{q}{q}\right)^{\mathbb{1}(y\in A)} \prod_{k=1}^{M} p_k^{\mathbb{1}(y=k)} \left(\frac{1-q}{1-q}\right)^{\mathbb{1}(y\notin A)} \prod_{k=M+1}^{D} p_k^{\mathbb{1}(y=k)} \\
&= \frac{q^{\mathbb{1}(y\in A)}}{\prod_{k=1}^{M} q^{\mathbb{1}(y=k)}} \prod_{k=1}^{M} p_k^{\mathbb{1}(y=k)} \frac{(1-q)^{\mathbb{1}(y\notin A)}}{\prod_{k=M+1}^{D}(1-q)^{\mathbb{1}(y=k)}} \prod_{k=M+1}^{D} p_k^{\mathbb{1}(y=k)} \\
&= \left[q^w(1-q)^{1-w}\right] \left[\prod_{k=1}^{M}\left(\frac{p_k}{q}\right)^{\mathbb{1}(y=k)}\right] \left[\prod_{k=M+1}^{D}\left(\frac{p_k}{1-q}\right)^{\mathbb{1}(y=k)}\right].
\end{aligned}$$

■

Corolário 2.1 *Sem perda de generalidade, considere* $\Lambda = \{A_1, A_2, ..., A_R\}$ *uma R-partição de* Ω, *em que* $A_1 = \{1, ..., M_1\}$, $A_2 = \{M_1 + 1, M_1 + 2, ..., M_2\}$, ..., $A_R = \{M_{R-1} + 1, M_{R-1} + 2, ..., D\}$. *Defina* $\boldsymbol{q} = (q_1, q_2, ..., q_R)$, *tal que* $q_l = \sum_{k\in A_l} p_k, l = 1, 2, ..., R$. *Defina também* $\boldsymbol{p}_{A_1} = (p_1, ..., p_{M_1})$, $\boldsymbol{p}_{A_2} = (p_{M_1+1}, ..., p_{M_2})$, ... *e* $\boldsymbol{p}_{A_R} = (p_{M_{R-1}+1}, ..., p_D)$. *Tome ainda* $W = \{l \in \{1, 2, ..., R\} : Y \in A_l\}$, *como na Equação (2.6). Então, a fp de* $Y|\boldsymbol{p} \sim BM_D(\boldsymbol{p})$ *se fatora nas fp's de* $W|\boldsymbol{q} \sim BM_R(\boldsymbol{q})$, $(Y|Y \in A_1; \boldsymbol{p}_{A_1}) \sim BM_{M_1}(\boldsymbol{p}_{A_1})$, $(Y|Y \in A_2; \boldsymbol{p}_{A_2}) \sim BM_{M_2-M_1}(\boldsymbol{p}_{A_2})$, ... *e* $(Y|Y \in A_R; \boldsymbol{p}_{A_R}) \sim BM_{D-M_{R-1}}(\boldsymbol{p}_{A_R})$.

Corolário 2.2 *Sem perda de generalidade, seja* $\Lambda = \{\{1\}, \{2, 3, ..., D\}\}$ *uma bipartição de* Ω. *Definindo* $W = \mathbb{1}(Y = 1)$, *temos que a fp de* $Y|\boldsymbol{p} \sim BM_D(\boldsymbol{p})$ *se fatora nas fp's de* $W|p_1 \sim B(p_1)$ *e* $(Y|Y \in \{2, 3, ..., D\}; (p_2, p_3, ..., p_D)) \sim BM_d\left(\frac{1}{1-p_1}(p_2, p_3, ..., p_D)\right)$.

Em particular, podemos fatorar a distribuição de Bernoulli D-variada em d Bernoulli's, como enunciamos a seguir.

Teorema 2.2 *Se* $Y|\boldsymbol{p} \sim BM_D(\boldsymbol{p})$, *então pode-se definir d variáveis aleatórias* $W_1, W_2, ..., W_d$, $W_j \in \{0, 1\}$, *transformações de* Y, *com respectivos parâmetros* $\theta_1, \theta_2, ..., \theta_d$, $0 < \theta_j < 1$, *em que* $W_j|\theta_j \sim$

$B(\theta_j)$, tal que

$$Pr(Y = y|\boldsymbol{p}) = \prod_{j=1}^{d} Pr(W_j = w_j|\theta_j) = \prod_{j=1}^{d} \theta_j^{w_j}(1-\theta_j)^{(1-w_j)}. \tag{2.7}$$

Demostração *De fato, como vimos do Teorema 2.1, considerando $\{A, A^c\}$ uma bipartição de Ω, temos que a fp de $Y|\boldsymbol{p}$ se fatora nas fp's de $W|q$, $(Y|Y \in A; \boldsymbol{p}_A)$ e $(Y|Y \in A^c; \boldsymbol{p}_{A^c})$. Por sua vez, $(Y|Y \in A; \boldsymbol{p}_A) \sim BM_M\left(\frac{1}{q}\boldsymbol{p}_A\right)$ e $(Y|Y \in A^c; \boldsymbol{p}_{A^c}) \sim BM_{D-M}\left(\frac{1}{1-q}\boldsymbol{p}_{A^c}\right)$. Reaplicando o Teorema 2.1 – recursiva e exaustivamente –, demostramos o que queríamos.*

■

O teorema anterior garante a existência de d variáveis com distribuição de Bernoulli tal que a fatoração seja possível, no entanto, nada diz sobre unicidade e tão pouco como construí-las. O corolário que segue explora esses pontos.

Corolário 2.3 *Seja $Y|\boldsymbol{p} \sim BM_D(\boldsymbol{p})$. Defina*

$$W_k = \mathbb{1}(Y = k|Y \in \{k, k+1, ..., D\}) = \begin{cases} 1, & \text{se } Y = k, \text{ dado que } Y \in \{k, k+1, ..., D\} \\ 0, & \text{se } Y \neq k, \text{ dado que } Y \in \{k, k+1, ..., D\} \end{cases}$$

e

$$\theta_k = Pr(Y = k|Y \in \{k, k+1, ..., D\}; \boldsymbol{p}) = \frac{p_k}{\sum_{l=k}^{D} p_l}, \qquad k = 1, 2, ..., d. \tag{2.8}$$

Então, o Teorema 2.2 segue para $W_1|\theta_1, W_2|\theta_2, ...$ e $W_d|\theta_d$.

Demostração *De fato, em um primeiro momento, basta considerarmos $A_1 = \{1\}$ e $A_2 = \{2, 3, ..., D\}$ uma bipartição de Ω e aplicar o Corolário 2.2, obtendo*

$$Pr(Y = y|\boldsymbol{p}) = \theta_1^{w_1}(1-\theta_1)^{(1-w_1)} Pr(Y = y|Y \in A_2; \boldsymbol{p}).$$

Aplicando o Corolário 2.2 novamente em $(Y|Y \in A_2; \boldsymbol{p})$, considerando que $B_1 = \{2\}$ e $B_2 = \{3, 4, ..., D\}$ formam uma bipartição de A_2, obtemos que

$$Pr(Y = y|\boldsymbol{p}) = \theta_1^{w_1}(1-\theta_1)^{(1-w_1)}\theta_2^{w_2}(1-\theta_2)^{(1-w_2)} Pr(Y = y|Y \in B_2; \boldsymbol{p}).$$

Procedendo recursiva e exaustivamente, provamos o corolário.

■

Observação *Ao definirmos $W_k = \mathbb{1}(Y = k|Y \in \{k, k+1, ..., D\}), k = 1, 2, ..., d$, note que se $y = 1$, então $w_1 = 1$, mas w_2 não está definida, isto é, não assume 0 nem 1. Podemos concluir que se $W_1 = 1$, a distribuição de W_2 não está definida. É como se observação w_2 de W_2 só ocorresse se observássemos $w_1 = 0$ para W_1. Assim, devemos considerar que, se $w_1 = 1$, então $\theta_2^{w_2}(1-\theta_2)^{1-w_2} = 1$. De maneira geral, para escrever a Equação (2.7), devemos considerar que $\theta_j^{w_j}(1-\theta_j)^{1-w_j} = 1$ sempre que w_j não está definida neste sentido.*

Note que, da Equação (2.8), obtemos a transformação inversa, dada por

$$\begin{aligned} p_1 &= \theta_1, \\ p_2 &= (1-\theta_1)\theta_2 \\ &\vdots \\ p_d &= (1-\theta_1)(1-\theta_2)...(1-\theta_{d-1})\theta_d \\ p_D &= (1-\theta_1)(1-\theta_2)...(1-\theta_d). \end{aligned}$$

Considerando que se $k=1$ então $\prod_{l=1}^{k-1}(1-\theta_l)=1$, e definindo $\theta_D = Pr(Y=D|Y\in\{D\};\boldsymbol{p}) = 1$, temos que, de uma forma geral, como apresentado em Pereira e Stern (2008),

$$p_k = \theta_k \prod_{l=1}^{k-1}(1-\theta_l), \qquad k=1,2,...,D. \tag{2.9}$$

Tal relação se preserva entre Y e $W_1, W_2, ..., W_d$, isto é, considerando que se $k=1$ então $\prod_{l=1}^{k-1}(1-W_l)=1$, e definindo $W_D = \mathbb{1}(Y=D|Y\in\{D\})=1$, temos que

$$\mathbb{1}(Y=k) = W_k \prod_{l=1}^{k-1}(1-W_l), \qquad k=1,2,...,D. \tag{2.10}$$

Note que, se $y=k$, então $w_k, w_{k-1}, w_{k-2}, ...$ e w_1 estão definidas e são, respectivamente, $1,0,0,...$ e 0. No entanto, $w_{k+1}, w_{k+2}, ...$ e w_d não estão definidas.

Exemplo 2.1 *Considere* $\Omega=\{1,2,3\}$. *Podemos aplicar o Teorema 2.2 (ou, neste caso, o Corolário 2.3, de três diferentes formas:*

1. $W_1 = \mathbb{1}(Y=1|Y\in\{1,2,3\})$, $W_2 = \mathbb{1}(Y=2|Y\in\{2,3\})$, *em que* $\theta_1 = p_1$ *e* $\theta_2 = \frac{p_2}{p_2+p_3}$,

ou tomando a primeira categoria por "2" e a segunda por "3"

2. $W_1 = \mathbb{1}(Y=2|Y\in\{1,2,3\})$, $W_2 = \mathbb{1}(Y=1|Y\in\{1,3\})$, *em que* $\theta_1 = p_2$ *e* $\theta_2 = \frac{p_1}{p_1+p_3}$,

ou ainda tomando a primeira categoria por "3" e a segunda por "1"

3. $W_1 = \mathbb{1}(Y=3|Y\in\{1,2,3\})$, $W_2 = \mathbb{1}(Y=1|Y\in\{1,2\})$, *em que* $\theta_1 = p_3$ *e* $\theta_2 = \frac{p_1}{p_1+p_2}$,

Assim, para os 3 *casos,* $W_1|\theta_1 \sim B(\theta_1)$ *e* $W_2|\theta_2 \sim B(\theta_2)$, *e temos que*

$$Pr(Y=y|\boldsymbol{p}) = \theta_1^{w_1}(1-\theta_1)^{(1-w_1)}\theta_2^{w_2}(1-\theta_2)^{(1-w_2)}.$$

Note que, para o caso 1 do exemplo anterior (análogo para os casos 2 e 3), definir $W_1 = 1-\mathbb{1}(Y=1|Y\in\{1,2,3\})$ e/ou $W_2 = 1-\mathbb{1}(Y=2|Y\in\{2,3\})$ nos leva à mesma fatoração, a menos das notações.

Considerando-se uma permutação $\{a_1, a_2, ..., a_D\}$ dos elementos de Ω e definindo $W_j = \mathbb{1}(Y=a_j|Y\in\{a_j, a_{j+1}, ..., a_D\})$ – excluindo-se os casos simétricos, ou seja, aqueles em que definimos $W_j = 1-\mathbb{1}(Y=a_j|Y\in\{a_j, a_{j+1}, ..., a_D\})$, para algum $j=1,2,...,d$ – obtemos uma classe de estruturas ao qual o Teorema 2.2 se aplica, como elucidado no exemplo anterior.

2.2.3 Bipartição total

A estrutura definida pelo Corolário 2.3 é resultado de um processo de bipartição, recursivo e exaustivo, dos conjuntos: $\{1, 2, ..., D\}$ em $\{1\}$ e $\{2, 3, ..., D\}$; $\{2, 3, ..., D\}$ em $\{2\}$ e $\{3, 4, ..., D\}$; e assim por diante até $\{d, D\}$ em $\{d\}$ e $\{D\}$. Entende-se por recursivo o processo de se biparticionar o respectivo subconjunto de categorias de forma análoga à bipartição anterior e por exaustivo o processo de se biparticionar (recursivamente) até que tal feito não seja mais possível. A seguir formalizamos essa ideia explorando algumas propriedades a fim de generalizar a construção feita no corolário anterior.

Definição 2.2 *Dizemos que λ é uma bipartição total de Ω, e denotamos por $\lambda(\Omega)$, se $\lambda(\Omega) = \{\Lambda_1, \Lambda_2, ..., \Lambda_d\}$, em que $\Lambda_1 = \{A_{11}, A_{12}\}$ é uma bipartição de Ω e $\Lambda_j = \{A_{j1}, A_{j2}\}$, $j = 2, 3, ..., d$, é uma bipartição de algum elemento $A_{k_1k_2}$ pertencente a algum $\Lambda_l, l < j$ (e que para cada elemento $A_{k_1k_2}$ haja somente uma bipartição Λ_j). De forma geral, qualquer permutação de $\Lambda_j, j = 1, 2, ..., d$, em $\lambda(\Omega)$ é considerada uma bipartição total.*

Exemplo 2.2 *Alguns exemplos são:*

1. $\Omega = \{1, 2, 3\}$ *com* $\lambda(\Omega) = \{\Lambda_1, \Lambda_2\}$ *tal que* $\Lambda_1 = \big\{\{1\}, \{2, 3\}\big\}$ *e* $\Lambda_2 = \big\{\{2\}, \{3\}\big\}$.
2. $\Omega = \{1, 2, 3, 4\}$ *com* $\lambda(\Omega) = \{\Lambda_1, \Lambda_2, \Lambda_3\}$ *tal que* $\Lambda_1 = \big\{\{1, 2\}, \{3, 4\}\big\}$, $\Lambda_2 = \big\{\{1\}, \{2\}\big\}$ *e* $\Lambda_3 = \big\{\{3\}, \{4\}\big\}$.
3. $\Omega = \{1, 2, ..., D\}$ *com* $\lambda(\Omega) = \{\Lambda_1, \Lambda_2, ..., \Lambda_d\}$ *tal que* $\Lambda_j = \big\{\{j\}, \{j + 1, j + 2, ..., D\}\big\}, j = 1, 2, ..., d$.

A formulação de bipartição total pode ser representada de diversas maneiras, como, por exemplo, em um problema de "colchetear" (em inglês: *"bracketing problem"*) categorias $1, 2, ..., D$ de forma que dentro de um par de colchetes haja exatamente 2 categorias e, uma vez que 2 categorias estão "colcheteadas", elas são consideradas como uma categoria só. O "colcheteamento" prossegue até a exaustão, fazendo uso de d pares de colchetes (e ainda que dentro de cada par de colchetes não haja ordem entre as duas categorias). Por exemplo, considerando $D = 4$, podemos "colchetear" a as categorias $1, 2, 3, 4$ de 15 diferentes formas: $\big[[12][34]\big]$; $\big[[13][24]\big]$; $\big[[14][23]\big]$; $\Big[1\big[2[34]\big]\Big]$; $\Big[1\big[3[24]\big]\Big]$; $\Big[1\big[4[23]\big]\Big]$; $\Big[2\big[1[34]\big]\Big]$; $\Big[2\big[3[14]\big]\Big]$; $\Big[2\big[4[13]\big]\Big]$; $\Big[3\big[1[24]\big]\Big]$; $\Big[3\big[2[14]\big]\Big]$; $\Big[3\big[4[12]\big]\Big]$; $\Big[4\big[1[23]\big]\Big]$; $\Big[4\big[2[13]\big]\Big]$ e $\Big[4\big[3[12]\big]\Big]$. O *bracketing problem* quando fixamos uma ordem para as categorias (por exemplo, $1, 2, 3..., D$) é conhecido como "problema de Catalan" devido a Catalan (1838) (mais detalhes em Comtet (1974)). Também podemos utilizar uma representação de bolas e urnas para uma bipartição total, em que Λ_j é ora considerado uma urna com duas bolas, ora considerado uma bola dentro de alguma outra urna Λ_l. Uma representação gráfica pode ser dada através de árvores binárias completas enraizadas com folhas rotuladas, que apresentaremos mais adiante.

Uma bipartição total de Ω se aplica naturalmente ao se contar de quantas formas podemos retirar n bolas com reposição de uma urna $\Omega = \{1, 2, ..., D\}$, de tal modo que tenhamos n_1 bolas $\{1\}$, n_2 bolas $\{2\}$, ... e n_D bolas $\{D\}$, em que $n = n_1 + n_2 + ... + n_D$. Tal quantidade é o coeficiente multinomial, podendo ser vista como (aplicando o princípio fundamental da contagem): o número de formas de se retirar n_1 bolas $\{1\}$ em n retiradas; e o número de formas de se retirar n_2 bolas

$\{2\}$ em $n - n_1$ retiradas; ...; e o número de formas de se retirar n_d bolas $\{d\}$ em $n_d + n_D$ retiradas, isto é,

$$\begin{pmatrix} n \\ n_1, n_2, ..., n_D \end{pmatrix} = \begin{pmatrix} n \\ n_1 \end{pmatrix} \begin{pmatrix} n - n_1 \\ n_2 \end{pmatrix} ... \begin{pmatrix} n_d + n_D \\ n_d \end{pmatrix} = \frac{n!}{n_1! n_2! ... n_D!}. \tag{2.11}$$

A definição do coeficiente multinomial em binomiais é discutida em Feller (1968). Neste caso, fazendo uma analogia com o item 3 do exemplo anterior, podemos representar a urna Ω através das urnas $\Lambda_j, j = 1, 2, ..., d$, considerando que a urna Λ_j contém duas "bolas": $\{j\}$ e Λ_{j+1}.

2.2.4 Generalizando a classe de fatorações

Uma bipartição total de Ω induz a fatoração da distribuição de Bernoulli multivariada em Bernoulli's, como enunciado no teorema a seguir.

Teorema 2.3 *Sejam $Y|\boldsymbol{p} \sim BM_D(\boldsymbol{p})$ e uma bipartição total $\lambda(\Omega) = \{\Lambda_1, \Lambda_2, ..., \Lambda_d\}$, em que $\Lambda_j = \{A_{j1}, A_{j2}\}$. Defina $W_j = \mathbb{1}(Y \in A_{j1} | Y \in \Lambda_j)$, ou seja,*

$$W_j = \begin{cases} 1, & \text{se } Y \in A_{j1} \\ 0, & \text{se } Y \in A_{j2} \end{cases}, \tag{2.12}$$

e também

$$\theta_j = Pr(Y \in A_{j1} | Y \in \Lambda_j; \boldsymbol{p}) = \frac{\sum_{k \in A_{j1}} p_k}{\sum_{k \in A_{j1} \cup A_{j2}} p_k}. \tag{2.13}$$

Então a fp de $Y|\boldsymbol{p} \sim BM_D(\boldsymbol{p})$ se fatora nas fp's de $W_j|\theta_j \sim B(\theta_j), j = 1, 2, ..., d$, isto é,

$$Pr(Y = y|\boldsymbol{p}) = \prod_{j=1}^{d} Pr(W_j = w_j|\theta_j) = \prod_{j=1}^{d} \theta_j^{w_j} (1 - \theta_j)^{(1-w_j)}.$$

Observações

1. *Por facilidade, escrevemos $Y \in \Lambda_j$ entendendo-se que $Y \in A_{j1} \cup A_{j2}$. No sentido que vimos representando até aqui, $Y \in \Lambda_j$ denotaria que $Y = A_{j1}$ ou $Y = A_{j2}$, mas estamos considerando Y assumindo valores em Ω.*

2. *W_j é definida somente para $Y \in \Lambda_j$ e, assim, se algum w_j não está definido (para um certo y), considera-se que $\theta_j^{w_j}(1 - \theta_j)^{1-w_j} = 1$.*

O teorema anterior é uma versão mais completa do Teorema 2.2 no sentido de não só garantir a existência, mas também de definir uma classe de estruturas, induzidas por $\lambda(\Omega)$, na qual W_j's e os respectivos θ_j's são definidos. Desse modo, a demonstração é análoga a do Teorema 2.2.

Dada uma bipartição total $\lambda(\Omega) = \{\Lambda_1, \Lambda_2, ..., \Lambda_d\}$, em que $\Lambda_j = \{A_{j1}, A_{j2}\}$, podemos escrever $\boldsymbol{p}$ como uma transformação de $\theta_1, \theta_2, ..., \theta_d$ (inversa a Equação (2.13)). Para notar tal fato, suponha que $\Lambda_m = \{A_{m1}, A_{m2}\}$, com $A_{m1} \cup A_{m2} \neq \Omega$, em que $A_{m1} = \{k\}$. Então,

$$p_k = Pr(Y = k|\boldsymbol{p}) = Pr(Y = k|Y \in \Lambda_m; \boldsymbol{p}) = Pr(Y = k|Y \in \Lambda_m; \boldsymbol{p}) Pr(Y \in \Lambda_m|\boldsymbol{p}) = \theta_m Pr(Y \in \Lambda_m|\boldsymbol{p}).$$

Além disso, pela definição de bipartição total, há algum $\Lambda_l = \{A_{l1}, A_{l2}\}, l \neq m$, em que, sem perda de generalidade, $A_{l1} = A_{m1} \cup A_{m2}$, isto é, Λ_m é uma bipartição de A_{l1}. Assim, $Pr(Y \in \Lambda_m|\boldsymbol{p}) = Pr(Y \in A_{l1}|\boldsymbol{p})$ e, portanto, seguindo com a equação anterior,

$$p_k = \theta_m Pr(Y \in A_{l1}|\boldsymbol{p}) = \theta_m Pr(Y \in A_{l1}, Y \in \Lambda_l|\boldsymbol{p}) = \theta_m Pr(Y \in A_{l1}|Y \in \Lambda_l;\boldsymbol{p}) Pr(Y \in \Lambda_l|\boldsymbol{p}) = \theta_m \theta_l Pr(Y \in \Lambda_l|\boldsymbol{p}).$$

Dessa forma, em algum momento, relacionamos p_k com alguns θ_j's através de uma transformação, digamos t_{k_λ}, que generaliza a Equação (2.9). Em outra palavras, existem dois subconjuntos disjuntos de índices de $\{1, 2, ..., d\}$, digamos J_{k1} e J_{k2}, tais que

$$p_k = t_{k_\lambda}(\boldsymbol{\theta}) = \prod_{j \in J_{k1}} \theta_j \prod_{j \in J_{k2}} (1 - \theta_j), \tag{2.14}$$

em que $\boldsymbol{\theta} = (\theta_1, \theta_2, ..., \theta_d)$, e se J_{k1} ou J_{k2} forem vazios, definimos o respectivo produtório igual a 1.

Especificamente, J_{k1} e J_{k2} são unicamente determinados por $\lambda(\Omega)$ como

$$J_{k1} = \{j \in \{1, 2, ..., d\}; k \in A_{j1}\} \quad \text{e} \quad J_{k2} = \{j \in \{1, 2, ..., d\}; k \in A_{j2}\}, \tag{2.15}$$

e generalizando a Equação (2.9), podemos escrever

$$p_k = \prod_{j \in J_k} \theta_j^{w_j} (1 - \theta_j)^{1 - w_j}, \qquad k = 1, 2, ..., D,$$

em que $J_k = J_{k1} \cup J_{k2} = \{j \in \{1, 2, ..., d\}; k \in \Lambda_j\}$. Note que $w_j = 1$ para $j \in J_{k1}$, $w_j = 0$ para $j \in J_{k2}$ e w_j não está definida para $j \notin J_k$.

Por outro lado, generalizando a Equação (2.10), a estrutura hierárquica se preserva para $W_j, j = 1, 2, ..., d$, e podemos escrever Y através de uma transformação, digamos v_{k_λ}, por

$$Y = k, \text{ se } v_{k_\lambda}(\boldsymbol{W}) = \prod_{j \in J_{k1}} W_j \prod_{j \in J_{k2}} (1 - W_j) = 1, \tag{2.16}$$

em que $\boldsymbol{W} = (W_1, W_2, ..., W_d)$.

Isto é, de forma geral, podemos definir $\boldsymbol{v}_\lambda$ e $\boldsymbol{t}_\lambda$ como

$$Y \;=\; \boldsymbol{v}_\lambda(\boldsymbol{W}) = \begin{cases} 1, & \text{se } v_{1_\lambda}(\boldsymbol{W}) = 1 \\ 2, & \text{se } v_{2_\lambda}(\boldsymbol{W}) = 1 \\ \vdots & \\ D, & \text{se } v_{D_\lambda}(\boldsymbol{W}) = 1 \end{cases} \tag{2.17}$$

e

$$\boldsymbol{p} \;=\; \boldsymbol{t}_\lambda(\boldsymbol{\theta}) = (t_{1_\lambda}(\boldsymbol{\theta}), t_{2_\lambda}(\boldsymbol{\theta}), ..., t_{D_\lambda}(\boldsymbol{\theta})). \tag{2.18}$$

2.2.5 Extensões para multinomial e verossimilhança

Antes de representar graficamente uma bipartição total e as transformações em Y e $\boldsymbol{p}$ induzidas por ela, exploramos a seguir a fatoração da distribuição multinomial com D categorias em d binomiais. Consequentemente, a verossimilhança do modelo de Bernoulli D-variado também se fatora.

Propriedade 2.4 *Seja $\boldsymbol{N} = (N_1, N_2, ..., N_D)$ um vetor aleatório com distribuição multinomial com parâmetros n e $\boldsymbol{p} = (p_1, p_2, ..., p_D)$, ou seja, com fp dada por*

$$Pr(\boldsymbol{N} = (n_1, n_2, ..., n_D)|n, \boldsymbol{p}) = \left(\begin{array}{c} n \\ n_1, n_2, ..., n_D \end{array} \right) \prod_{k=1}^{D} p_k^{n_k},$$

em que $n = \sum_{k=1}^{D} n_k$. Então, a distribuição de $(\boldsymbol{N}|n, \boldsymbol{p})$ se fatora em d distribuições binomiais, isto é, existem d variáveis aleatórias $X_1, X_2, ..., X_d$ com parâmetros $m_1, m_2, ..., m_d$ e $\theta_1, \theta_2, ..., \theta_d$, respectivamente, tal que

$$Pr(\boldsymbol{N} = (n_1, n_2, ..., n_D)|n, \boldsymbol{p}) = \prod_{j=1}^{d} \left(\begin{array}{c} m_j \\ x_j \end{array} \right) \theta_j^{x_j} (1 - \theta_j)^{m_j - x_j}.$$

Demostração *De fato. Primeiro, note que se $Y_1, Y_2, ..., Y_n$ são independentes, dado $\boldsymbol{p}$, tais que $Y_i|\boldsymbol{p} \sim BM_D(\boldsymbol{p}), i = 1, 2, ..., n$, então $\boldsymbol{N} = (N_1, N_2, ..., N_D)$ tal que $N_k = \sum_{i=1}^{n} \mathbb{1}(Y_i = k)$ tem distribuição multinomial com parâmetros n e $\boldsymbol{p}$. Note também que se $W_{j1}, W_{j2}, ..., W_{jm_j}$ são independentes, dado θ_j, tal que $W_{ji} \sim B(\theta_j), i = 1, 2, ..., m_j$, então $X_j = \sum_{i=1}^{m_j} W_{ji}$ tem distribuição binomial com parâmetros m_j e θ_j, para $j = 1, 2, ..., d$.*

Assim, usando o Teorema 2.3, temos que

$$\begin{aligned} \prod_{k=1}^{D} p_k^{n_k} &= \prod_{k=1}^{D} p_k^{\sum_{i=1}^{n} \mathbb{1}(y_i = k)} = \prod_{i=1}^{n} \prod_{k=1}^{D} p_k^{\mathbb{1}(y_i = k)} \\ &= \prod_{i=1}^{n} \prod_{j=1}^{d} \theta_j^{w_{ji}} (1 - \theta_j)^{1 - w_{ji}} = \prod_{j=1}^{d} \theta_j^{\sum_{i=1}^{n} w_{ji}} (1 - \theta_j)^{\sum_{i=1}^{n} (1 - w_{ji})}. \end{aligned}$$

No entanto, há somente m_j termos na soma $\sum_{i=1}^{n} w_{ji}$ – bem como em $\sum_{i=1}^{n} (1 - w_{ji})$ – que estão definidos. Isto é, considerando uma bipartição total $\lambda(\Omega) = \{\Lambda_1, \Lambda_2, ..., \Lambda_d\}$, com $\Lambda_j = \{A_{j1}, A_{j2}\}$, temos que

$$m_j = \sum_{i=1}^{n} \mathbb{1}(y_i \in \Lambda_j), \qquad j = 1, 2, ..., d. \tag{2.19}$$

Assim, definindo o conjunto de índices

$$\begin{aligned} I_j &= \{i \in \{1, 2, ..., n\}; w_{ji} \text{ está definido}\} \\ &= \{i \in \{1, 2, ..., n\}; y_i \in \Lambda_j\}, \end{aligned} \tag{2.20}$$

podemos escrever

$$x_j = \sum_{i=1}^{n} w_{ji} = \sum_{i \in I_j} w_{ji} \qquad e \qquad m_j - x_j = \sum_{i=1}^{n} (1 - w_{ji}) = \sum_{i \in I_j} (1 - w_{ji}).$$

Assim sendo,

$$\begin{aligned}\prod_{k=1}^{D} p_k^{n_k} &= \prod_{j=1}^{d} \theta_j^{\sum_{i=1}^{n} w_{ji}} (1-\theta_j)^{\sum_{i=1}^{n}(1-w_{ji})} \\ &= \prod_{j=1}^{d} \theta_j^{\sum_{i\in I_j} w_{ji}} (1-\theta_j)^{\sum_{i\in I_j}(1-w_{ji})} \\ &= \prod_{j=1}^{d} \theta_j^{x_j} (1-\theta_j)^{(m_j-x_j)}.\end{aligned}$$

Por outro lado, como elucidado na Equação (2.11), temos que

$$\begin{pmatrix} n \\ n_1, n_2, ..., n_D \end{pmatrix} = \begin{pmatrix} m_1 \\ x_1 \end{pmatrix} \begin{pmatrix} m_2 \\ x_2 \end{pmatrix} ... \begin{pmatrix} m_d \\ x_d \end{pmatrix} = \frac{n!}{n_1! n_2! ... n_D!},$$

e então,

$$\begin{pmatrix} n \\ n_1, n_2, ..., n_D \end{pmatrix} \prod_{k=1}^{D} p_k^{n_k} = \prod_{j=1}^{d} \begin{pmatrix} m_j \\ x_j \end{pmatrix} \theta_j^{x_j} (1-\theta_j)^{m_j - x_j}.$$

■

Propriedade 2.5 *Sejam $y_1, y_2, ..., y_n$ observações independentes, dado $\boldsymbol{p}$, de $Y|\boldsymbol{p} \sim BM_D(\boldsymbol{p})$. Então a verossimilhança do modelo de Bernoulli multivariado se fatora em d verossimilhanças do modelo de Bernoulli.*

Demostração *De fato, como a verossimilhança do modelo Bernoulli multivariado é dada por*

$$L(\boldsymbol{p}|y_1, y_2, ..., y_n) = \prod_{i=1}^{n} \prod_{k=1}^{D} p_k^{\mathbb{1}(y_i=k)} = \prod_{k=1}^{D} p_k^{n_k},$$

em que $n_k = \sum_{i=1}^{n} \mathbb{1}(y = k)$, vimos na propriedade anterior, definindo I_j como na Equação (2.20), que

$$\prod_{k=1}^{D} p_k^{n_k} = \prod_{j=1}^{d} \theta_j^{\sum_{i\in I_j} w_{ji}} (1-\theta_j)^{\sum_{i\in I_j}(1-w_{ji})},$$

e, sendo $L_j(\theta_j|\{w_{ji} : i \in I_j\})$ as verossimilhanças dos modelos Bernoulli's, temos que

$$L(\boldsymbol{p}|y_1, y_2, ..., y_n) = \prod_{k=1}^{D} p_k^{n_k} = \prod_{j=1}^{d} \prod_{i\in I_j} \theta_j^{w_{ji}} (1-\theta_j)^{(1-w_{ji})} = \prod_{j=1}^{d} L_j(\theta_j|\{w_{ji} : i \in I_j\}).$$

■

Das duas propriedades anteriores, vemos que uma bipartição total $\lambda(\Omega)$ induz também uma transformação das observações $y_1, y_2, ..., y_n$ de $Y|\boldsymbol{p}$ para $w_{j1}, w_{j2}, ..., w_{jn}$ de $W_j|\theta_j, j = 1, 2, ..., d$. Isto é, denotando por $\mathcal{D} = \{y_1, y_2, ..., y_n\}$, definimos $\mathcal{D}_\lambda = \{\mathcal{D}_{j_\lambda}, j = 1, 2, ..., d\}$ tal que

$$\mathcal{D}_{j_\lambda} = \{w_{ji} : i \in I_j\}, \tag{2.21}$$

em que I_j é dado pela Equação (2.20) e da propriedade anterior segue que

$$L(\boldsymbol{p}|\mathcal{D}) = \prod_{j=1}^{d} L_j(\theta_j|\mathcal{D}_{j_\lambda}). \tag{2.22}$$

Exemplo 2.3 *Seja* $\Omega = \{1,2,3,4\}$ *e* $\lambda(\Omega) = \{\Lambda_1, \Lambda_2, \Lambda_3\}$ *tal que* $\Lambda_1 = \{\{1\},\{2,3,4\}\}$, $\Lambda_2 = \{\{2\},\{3,4\}\}$ *e* $\Lambda_3 = \{\{3\},\{4\}\}$. *Suponha que temos* 16 *observações de* $Y|\boldsymbol{p} \sim BM_4(\boldsymbol{p})$*:* $\mathcal{D} = \{1,1,1,1,\ 2,2,2,2,\ 3,3,3,3,\ 4,4,4,4\}$*, então*

$$\begin{aligned}
\mathcal{D}_{1_\lambda} &= \{w_{11}; w_{12}; w_{13}; w_{14}, w_{15}; w_{16}; w_{17}; w_{18}, w_{19}; w_{1,10}; w_{1,11}; w_{1,12}, w_{1,13}; w_{1,14}; w_{1,15}; w_{1,16}\} \\
&= \{1,1,1,1,0,0,0,0,0,0,0,0,0,0,0,0\}; \\
\mathcal{D}_{2_\lambda} &= \{w_{25}; w_{26}; w_{27}; w_{28}, w_{29}; w_{2,10}; w_{2,11}; w_{2,12}, w_{2,13}; w_{2,14}; w_{2,15}; w_{2,16}\} \\
&= \{1,1,1,1,0,0,0,0,0,0,0,0\}; \\
\mathcal{D}_{1_\lambda} &= \{w_{39}; w_{3,10}; w_{3,11}; w_{3,12}, w_{3,13}; w_{3,14}; w_{3,15}; w_{3,16}\} \\
&= \{1,1,1,1,0,0,0,0\},
\end{aligned}$$

podendo ser representado, com um certo abuso de notação, por

$$\mathcal{D}_\lambda = \left\{ \begin{array}{c} \mathcal{D}_{1_\lambda} \\ \mathcal{D}_{1_\lambda} \\ \mathcal{D}_{1_\lambda} \end{array} \right\} = \left\{ \begin{array}{cccccccccccccccc} 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ & & & & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ & & & & & & & & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \end{array} \right\}.$$

Propriedade 2.6 *A log-verossimilhança do modelo de Bernoulli multivariado é a soma de d log-verossimilhanças do modelo de Bernoulli, isto é,*

$$l(\boldsymbol{p}|\mathcal{D}) = \sum_{j=1}^{d} l_j(\theta_j|\mathcal{D}_{j_\lambda}).$$

2.3 Representações gráficas

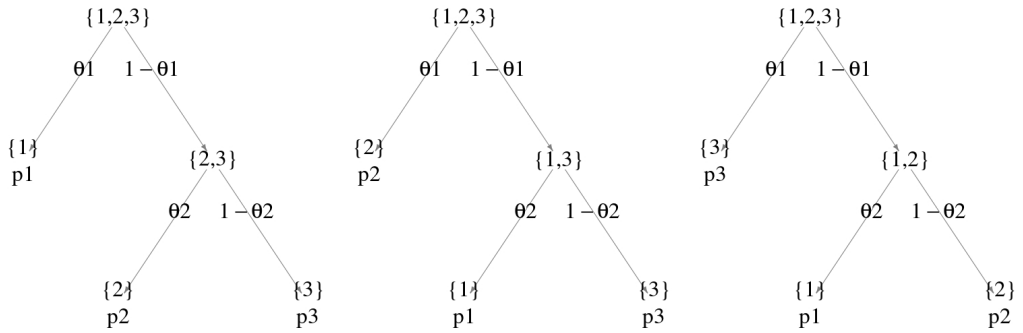
Podemos utilizar grafos para representar uma bipartição total qualquer de Ω – mais especificamente, árvores binárias completas enraizadas com folhas rotuladas não ordenadas –, em que:

- a raiz representa o conjunto de categorias $\Omega = \{1,2,...,D\}$;
- os ramos representam $\Lambda_j = \{A_{j1}, A_{j2}\}$, isto é, como o nó $A_{j1} \cup A_{j2}$ está sendo biparticionado nos nós A_{j1} e A_{j2};
- os nós – sendo Ω a raiz – representam os conjuntos A_{j1} e A_{j2} que compõem Λ_j, em que:
  - os d nós internos são aqueles que serão biparticionados em um próximo nível e
  - os D nós terminais (as folhas) são $\{1\},\{2\},...,\{D\}$, que por suas vezes não podem ser biparticionados.

Também podemos representar nas árvores a relação entre θ_j's e p_k's (Equação (2.18)), ou seja, podemos imprimir θ_j ou $1-\theta_j$ no par de ramos que representa Λ_j, quando o respectivo ramo parte de $A_{j1} \cup A_{j2}$ para A_{j1} ou para A_{j2}, respectivamente, contemplando todos os ramos, estruturados

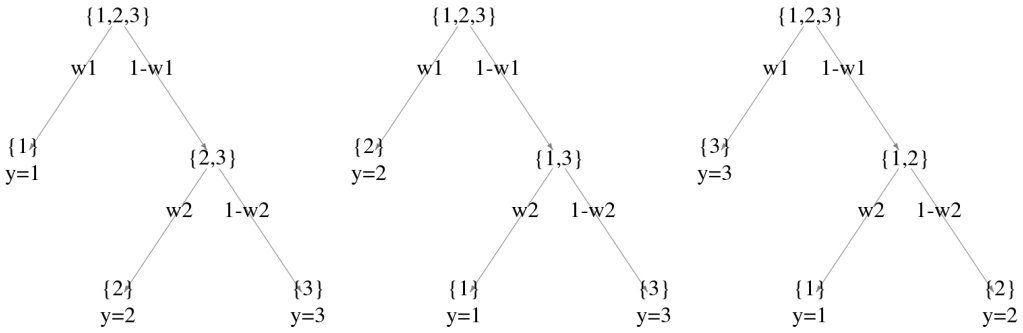
da raiz até às folhas e que, nelas, podemos imprimir os respectivos p_k's. Analogamente, podemos representar a relação entre W_j's e Y (Equação (2.17)).

Nas 2 figuras a seguir, revisitamos o Exemplo 2.1 e apresentamos tais representações, respectivamente. Para o caso 1 da Figura 2.2 (em que a bipartição total é dada no Exemplo 2.2.1), vê-se que θ_1 é a probabilidade condicional de $Y = 1$ dado que $Y \in \Omega$, bem como que θ_2 é a probabilidade condicional de $Y = 2$ dado que $Y \in \{2, 3\}$. Ainda para o caso 1, também vê-se a representação de $\boldsymbol{t}_\lambda$, ou seja, p_k como o produto das probabilidades condicionais (θ_j ou $1 - \theta_j$) através dos ramos que partem da raiz Ω à folha $\{k\}$: $p_1 = \theta_1$, $p_2 = (1 - \theta_1)\theta_2$ e $p_3 = (1 - \theta_1)(1 - \theta_2)$. Analogamente, para o caso 1 da Figura 2.3, vê-se que W_1 é a indicadora de $Y = 1$, dado que $Y \in \Omega$ e que W_2 é a indicadora de $Y = 2$, dado que $Y \in \{2, 3\}$. Também vê-se a representação de $\boldsymbol{v}_\lambda$, ou seja, $(Y = k)$ se o produto das variáveis W_j's ou $1 - W_j$'s, através dos ramos que partem da raiz Ω à folha $\{k\}$ é igual a 1: $Y = 1$ se $W_1 = 1$, $Y = 2$ se $(1 - W_1)W_2 = 1$ e $Y = 3$ se $(1 - W_1)(1 - W_2) = 1$.



(a) Caso 1. (b) Caso 2. (c) Caso 3.

Figura 2.2: *Representações das estruturas das bipartições totais para o Exemplo 2.1: θ_j's e $\boldsymbol{p}$.*



(a) Caso 1. (b) Caso 2. (c) Caso 3.

Figura 2.3: *Representações das estruturas das bipartições totais para o Exemplo 2.1: W_j's e Y.*

Utilizemos por um momento, como elucidação, a seguinte nomenclatura: nó ancestral para aquele conjunto que está sendo biparticionado; e nós descendentes os 2 gerados da bipartição do respectivo nó ancestral. Assim, generalizando, um θ_j qualquer (ou $1 - \theta_j$) é a probabilidade condicional de, dado o nó ancestral, observarmos o respectivo nó descendente. E para o nó descendente $\{k\}$ (que por sua vez não tem descendentes), p_k pode ser escrito por θ_j's através de sua hereditariedade: do ancestral primordial Ω seguindo até o ancestral de $\{k\}$ e por fim à $\{k\}$. Para as variáveis W_j a analogia é a mesma. Assim, nas representações seguintes, utilizamos somente θ_j's nos ramos ao

invés de W_j's. Aqui, o conjunto de índices J_k também fica representado, sendo todos os pares de ramos – que representam Λ_j – que contém a categoria k.

Além disso note que, ainda utilizando o primeiro caso na Figura 2.2, a representação para $W_1 = 1 - \mathbb{1}(Y = 1|Y \in \{1,2,3\})$ e/ou $W_2 = 1 - \mathbb{1}(Y = 2|Y \in \{2,3\})$ é a mesma, a menos de rotações nos nós. A Figura 2.4 mostra tal simetria para o caso 1, que é análoga para os casos 2 e 3. Já a Figura 2.5 mostra o caso abordado no Corolário 2.3 em que a bipartição total é dada no Exemplo 2.2.3.

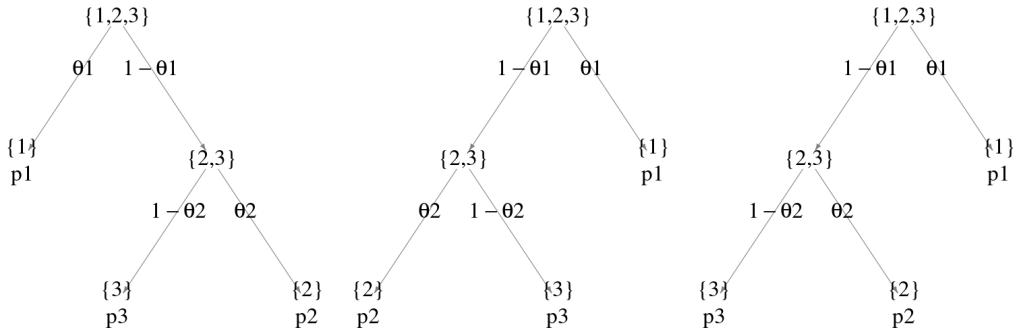


Figura 2.4: *Representações simétricas entre si e ao caso* 1 *da Figura 2.2.*

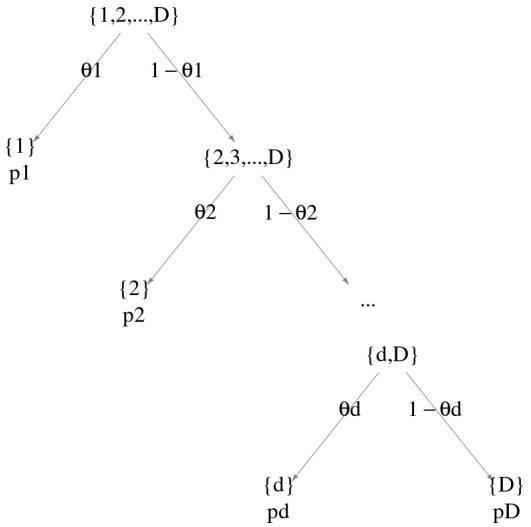


Figura 2.5: *Representação para o caso abordado no Corolário 2.3 e no Exemplo 2.2.3.*

Exemplo 2.4 *Considere* $\Omega = \{1,2,3,4\}$. *A Figura 2.6 mostra duas diferentes estruturas – onde há um total de* 15 *possíveis – em que a estrutura na Figura 2.6(a) é análoga a outras* 11 *e a na Figura 2.6(b) é análoga a outras* 2.

Além disso, para a estrutura na Figura 2.6(b), temos que

$$\begin{aligned} W_1 &= \mathbb{1}(Y \in \{1,2\}) \\ W_2 &= \mathbb{1}(Y = 1|Y \in \{1,2\}) \\ W_3 &= \mathbb{1}(Y = 3|Y \in \{3,4\}) \end{aligned}$$

com respectivos parâmetros dados por

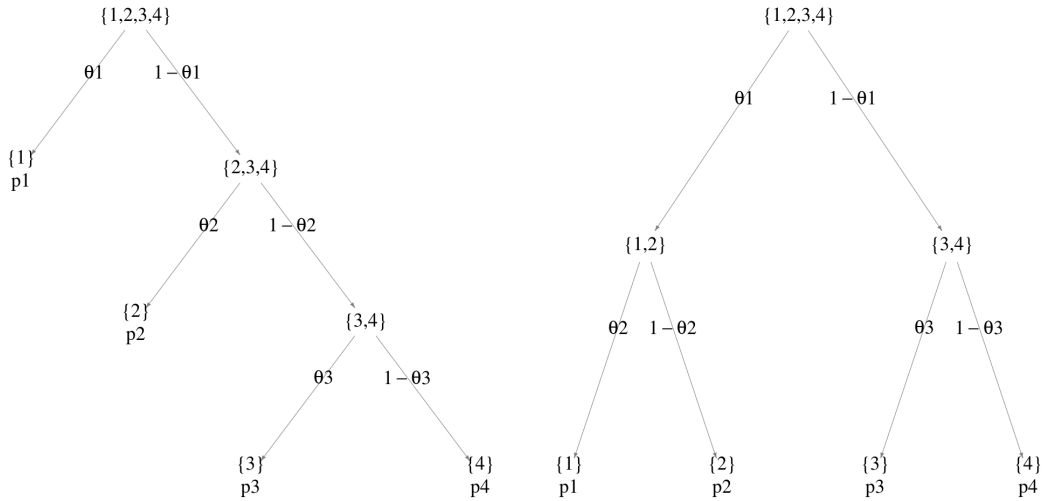
$$\begin{aligned}\theta_1 = &\quad Pr(Y \in \{1,2\} | Y \in \Omega; \boldsymbol{p}) &= \frac{p_1 + p_2}{p_1 + p_2 + p_3 + p_4} \\ \theta_2 = &\quad Pr(Y = 1 | Y \in \{1,2\}; \boldsymbol{p}) &= \frac{p_1}{p_1 + p_2} \\ \theta_3 = &\quad Pr(Y = 3 | Y \in \{3,4\}; \boldsymbol{p}) &= \frac{p_3}{p_3 + p_4}.\end{aligned}$$

Sendo assim, $Y = \boldsymbol{v}_\lambda(\boldsymbol{W})$ *e* $\boldsymbol{p} = \boldsymbol{t}_\lambda(\boldsymbol{p})$*, elucidando as Equações (2.17) e (2.18), são dadas por, respectivamente,*

$$Y = \begin{cases} 1, & se\ W_1 W_2 = 1 \\ 2, & se\ W1(1 - W_2) = 1 \\ 3, & se\ (1 - W_1)W_3 = 1 \\ 4, & se\ (1 - W_1)(1 - W_3) = 1 \end{cases}$$

e

$$\begin{aligned} p_1 &= \theta_1\theta_2 \\ p_2 &= \theta_1(1 - \theta_2) \\ p_3 &= (1 - \theta_1)\theta_3 \\ p_4 &= (1 - \theta_1)(1 - \theta_3). \end{aligned}$$



(a) (b)

Figura 2.6: *Duas representações para* $D = 4$*.*

2.3.1 Níveis das representações

Podemos estar interessados em certas bipartições totais que compartilham de uma mesma característica, por exemplo, aquelas que têm o mesmo número de níveis.

Propriedade 2.7 *Seja* c *o número de níveis em uma representação em árvore de uma bipartição*

total de Ω. *Então,* $c \in [c_{min}, c_{max}]$, *em que* $c_{min} = \lceil \log_2(D) \rceil$, *em que* $\lceil x \rceil$, $x \in \mathbb{R}$, *é o menor número inteiro maior do que* x *(caso elucidado nas Figuras (2.2) e (2.6(b)), e* $c_{max} = d$ *(caso representado na Figura 2.5).*

Sendo $\lambda(\Omega) = \{\Lambda_1, \Lambda_2, ..., \Lambda_d\}$, em que $\Lambda_j = \{A_{j1}, A_{j2}\}$, e sabendo que D dos $2d$ conjuntos $A_{jh}, j = 1, 2, ..., d, h = 1, 2$, são da forma $\{k\}, k = 1, 2, ..., D$, temos que o número de níveis das representações em árvore, digamos c_λ, é dado por

$$c_\lambda = \#\{\Lambda_j \in \lambda(\Omega); A_{jh} = \{k\}, \text{ para algum } h = 1, 2 \text{ e algum } k = 1, 2, ..., D\} - 1.$$

Dessa forma, dado um certo λ, podemos determinar o número de níveis c_λ. Por outro lado, alguém pode estar interessado em, fixado um número de níveis c, explorar somente aqueles λ's de tal modo que o número de níveis seja igual a c. Em outras palavras, sendo $\mathscr{L}$ a classe de todas as bipartições totais de Ω, podemos tomar uma subclasse $\mathscr{L}_c = \{\lambda \in \mathscr{L}; c_\lambda = c\}$.

2.3.2 Contando o número de bipartições totais

A propriedade seguinte explora a quantidade de elementos em $\mathscr{L}$.

Propriedade 2.8 *Seja* $\mathscr{L}$ *a classe que contém todas as bipartições totais de* Ω. *Então* $\mathscr{L}$ *contém* $(2d-1)!!$ *elementos.*

Demostração *Contar o número de bipartições totais é equivalente a contar o número de árvores que as representam. Denote tal quantidade por* $a(D)$. *Para seguir a demostração, vamos estabelecer uma relação de recorrência. Considere, então, as árvores com d folhas. Qualquer árvore com d folhas tem* $d-1$ *nós internos (contando a raiz). Então,* $2d-1$ *nós ao todo (contando as folhas). Ao inserir a categoria* $\{D\}$ *em* Ω, *há (em uma árvore particular com d folhas)* $2d-1$ *possíveis nós para inseri-la e, então, biparticionar tal "novo" nó em* 2*:* $\{D\}$ *e o "antigo" nó. Para cada diferente nó "antigo" que inserimos* $\{D\}$, *teremos uma árvore diferente. Então como há* $a(d)$ *árvores com d folhas, temos que,*

$$a(D) = (2d-1)a(d),$$

e portanto

$$a(D) = (2d-1)(2d-3)...3.1 = (2d-1)!!$$

■

Stanley (1999) traz uma demostração alternativa através de funções geradoras em análise combinatória e também uma ideia com bolas e urnas. Uma discussão mais detalhada sobre árvores binárias completas é encontrada em Murtagh (1984) e Moon (1970).

A fim de elucidar tal demostração, suponha que estamos contando o número bipartições totais com $D = 4$. Então, consideramos em um primeiro momento todas as árvores com 3 folhas (Figura 2.2). Para cada árvore, há 5 possíveis nós para inserir a categoria $\{4\}$. Considere o caso 1 apresentado na Figura 2.2(a). Podemos inserir a categoria $\{4\}$ em $\{1, 2, 3\}$, $\{1\}$, $\{2, 3\}$, $\{2\}$ e $\{3\}$ (mostrado, respectivamente, na Figura 2.7). Para cada um destes casos, teremos uma árvore diferente (e então,

uma bipartição total diferente). Considerando os outros 2 casos (Figuras 2.2(b) e 2.2(c)), temos 15 possíveis bipartições totais (assim como quando representamos uma bipartição total através do *braketing problem*).

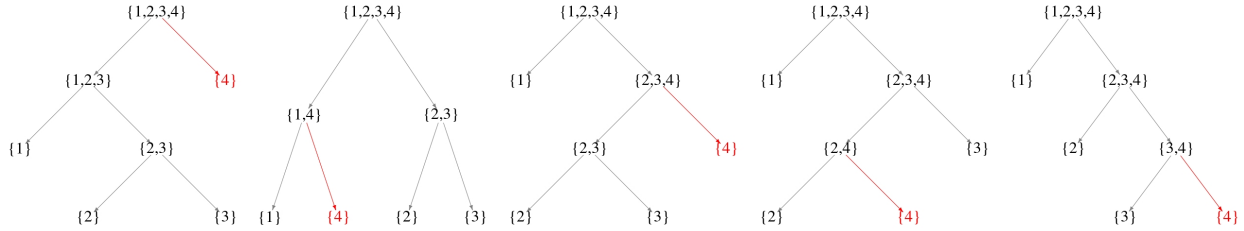


Figura 2.7: *Elucidação da desmostração da Propriedade 2.8.*

Apesar de $(2d-1)!!$ diferentes fatorações da distribuição de Bernoulli D-variada em d Bernoulli's, a propriedade a seguir mostra que a estimativa amostral do parâmetro $\boldsymbol{p}$ (igual a de máxima verossimilhança) pode ser obtida através de qualquer fatoração através das respectivas estimativas amostrais de $\theta_1, \theta_2, ..., \theta_d$ (iguais as de máxima verossimilhança).

Propriedade 2.9 *Seja $\mathcal{D} = \{y_i, i = 1, 2, ..., n\}$ observações independentes de $Y|\boldsymbol{p} \sim BM_D(\boldsymbol{p})$ e λ uma bipartição total qualquer de Ω. Seja também $\hat{\boldsymbol{p}} = (\hat{p}_1, \hat{p}_2, ..., \hat{p}_D)$ o vetor de proporções amostrais, ou seja,*

$$\hat{p}_k = \frac{\sum_{i=1}^{n} \mathbb{1}(y_i = k)}{n}, \qquad k = 1, 2, ..., D.$$

Seja, agora, $\hat{\boldsymbol{\theta}} = (\hat{\theta}_1, \hat{\theta}_2, ..., \hat{\theta}_d)$ o vetor de proporções amostrais dos respectivos dados induzidos por $\lambda(\Omega)$: $\mathcal{D}_{j\lambda} = \{w_{ji} : i \in I_j\}$, dados pela Equação (2.21), isto é,

$$\hat{\theta}_j = \frac{\sum_{i \in I_j} w_{ji}}{m_j}, \qquad j = 1, 2, ..., d,$$

em que m_j é dado pela Equação (2.19). Então,

$$\hat{\boldsymbol{p}} = \boldsymbol{t}_\lambda(\hat{\boldsymbol{\theta}}).$$

Exemplo 2.5 *Revisitando o Exemplo 2.3, temos que $\hat{\theta}_1 = \frac{1}{4}$, $\hat{\theta}_2 = \frac{1}{3}$ e $\hat{\theta}_3 = \frac{1}{2}$. Então, de $\boldsymbol{t}_\lambda$ como na Equação (2.9), temos que*

$$\begin{aligned}
\hat{p}_1 &= \hat{\theta}_1 = \frac{1}{4} \\
\hat{p}_2 &= (1 - \hat{\theta}_1)\hat{\theta}_2 = \frac{1}{4} \\
\hat{p}_3 &= (1 - \hat{\theta}_1)(1 - \hat{\theta}_2)\hat{\theta}_3 = \frac{1}{4} \\
\hat{p}_4 &= (1 - \hat{\theta}_1)(1 - \hat{\theta}_2)(1 - \hat{\theta}_3) = \frac{1}{4}.
\end{aligned}$$

Desse ponto de vista, a fatoração não traz qualquer vantagem. No entanto, ao inserirmos covariáveis, tendo um problema de classificação e regressão multinomial, cada bipartição total nos levará a diferentes problemas de classificação binomial e a diferentes estimações para $\boldsymbol{p}$. Este é o ponto que discutimos na próximo capítulo.

Capítulo 3

Regressão e classificação para dados politômicos

No capítulo anterior vimos que a distribuição de Bernoulli D-variada se fatora de diversas formas em d Bernoulli's. Contudo, quando temos um problema politômico sem covariáveis, tais fatorações não trazem qualquer vantagem do ponto de vista da estimação de máxima verossimilhança. Neste capítulo exploramos a classificação e regressão para dados politômicos no contexto das fatorações e como cada bipartição total conduz a uma regressão e classificação diferentes. Discutimos a classificação em tal contexto de uma forma geral e em seguida a regressão multinomial como uma generalização da usual, logística com categoria de referência.

3.1 Classificação

O problema geral da classificação para dados politômicos é construir uma função que, com base em características observadas de um determinado indivíduo, o prediz em uma das categorias em $\Omega = \{1, 2, ..., D\}$. Tais características são chamadas de covariáveis (ou variáveis preditoras ou também de variáveis independentes).

Definição 3.1 *Sejam $\Omega = \{1, 2, ..., D\}$ e $\boldsymbol{x}$ um vetor de covariáveis para um indivíduo assumindo valores em $\mathcal{X} \subset \mathbb{R}^m$. Definimos por função de classificação (ou simplesmente de classificador) multinomial (ou politômica), a função δ dada por*

$$\begin{array}{rccc} \delta: & \mathcal{X} & \to & \Omega \\ & \boldsymbol{x} & \mapsto & \hat{y} = \delta(\boldsymbol{x}) \end{array}.$$

Como δ é uma função, então para qualquer $\boldsymbol{x} \in \mathcal{X}$, há um único valor $\delta(\boldsymbol{x}) \in \Omega$. Dessa forma, a transformação inversa define regiões de classificação em $\mathcal{X}$, isto é,

$$\mathcal{X}_k = \delta^{-1}(k) = \{\boldsymbol{x} \in \mathcal{X}; \delta(\boldsymbol{x}) = k\}, \quad k = 1, 2, ..., D.$$

E como, por definição, δ é uma função em $\mathcal{X}$, então $\mathcal{X}_1, \mathcal{X}_2, ..., \mathcal{X}_D$ forma uma partição de $\mathcal{X}$. Por outro lado, definir uma partição $\mathcal{X}_1, \mathcal{X}_2, ..., \mathcal{X}_D$ determina unicamente um classificador em Ω por

$$\delta(\boldsymbol{x}) = k, \text{ se } \boldsymbol{x} \in \mathcal{X}_k, \quad k = 1, 2, ..., D.$$

Exemplo 3.1 *Suponha que temos somente uma covariável $x \in \mathbb{R}$. Podemos definir d constantes reais $a_1 < a_2 < \cdots < a_d$ e assim, uma partição de $\mathcal{X} = \mathbb{R}$ como*

$$\begin{aligned} \mathcal{X}_1 &= (-\infty, a_1] \\ \mathcal{X}_k &= (a_{k-1}, a_k], \quad k = 2, 3, ..., d \\ \mathcal{X}_D &= (a_d, \infty). \end{aligned}$$

E portanto, $\delta(x) = k$, se $x \in \mathcal{X}_k, k = 1, 2, ..., D$.

3.1.1 Fatorações em classificadores dicotômicos

Do capítulo anterior, da Equação (2.17), vimos que podemos decompor – através de uma bipartição total λ – uma variável que assume valores em Ω por d variáveis que assumem valores em $\{0, 1\}$. Desse modo, podemos escrever

$$\hat{y} = \boldsymbol{v}_\lambda(\hat{\boldsymbol{w}}), \tag{3.1}$$

em que $\hat{\boldsymbol{w}} = (\hat{w}_1, \hat{w}_2, ..., \hat{w}_d)$, sendo $\hat{w}_j = \gamma_j(\boldsymbol{x}), j = 1, 2, ..., d$, em que γ_j é um classificador binomial (ou dicotômico) qualquer, ou seja, assumindo valores em $\{0, 1\}$. Isto é,

$$\begin{array}{rccc} \gamma_j: & \mathcal{X} & \to & \{0,1\} \\ & \boldsymbol{x} & \mapsto & \hat{w}_j = \gamma_j(\boldsymbol{x}) \end{array}.$$

Dessa forma, podemos definir um classificador multinomial através de d classificadores binomiais, ou seja, $\delta(\boldsymbol{x}) = \boldsymbol{v}_\lambda(\boldsymbol{\gamma}(\boldsymbol{x}))$, em que $\boldsymbol{\gamma}(\boldsymbol{x}) = (\gamma_1(\boldsymbol{x}), \gamma_2(\boldsymbol{x}), ..., \gamma_d(\boldsymbol{x}))$.

Pela relação na Equação (3.1), λ induz também uma partição em $\mathcal{X}$. Para notar tal fato, primeiro suponha que $\lambda(\Omega) = \{\Lambda_1, \Lambda_2, ..., \Lambda_d\}$, em que $\Lambda_j = \{A_{j1}, A_{j2}\}$. Desse modo, da Equação (2.16), temos que

$$\delta(\boldsymbol{x}) = k, \text{ se } \prod_{j \in J_{k1}} \gamma_j(\boldsymbol{x}) \prod_{j \in J_{k2}} (1 - \gamma_j(\boldsymbol{x})) = 1,$$

em que $J_{k1} = \{j \in \{1, 2, ..., d\}; k \in A_{j1}\}$ e $J_{k2} = \{j \in \{1, 2, ..., d\}; k \in A_{j2}\}$. E assim, temos que

$$\mathcal{X}_k = \bigcap_{j \in J_{k1}} \mathcal{X}_{j1} \bigcap_{j \in J_{k2}} \mathcal{X}_{j2}, \tag{3.2}$$

em que

$$\begin{aligned} \mathcal{X}_{j1} &= \gamma_j^{-1}(1) = \{\boldsymbol{x} \in \mathcal{X} : \gamma_j(\boldsymbol{x}) = 1\} \\ \mathcal{X}_{j2} &= \gamma_j^{-1}(0) = \{\boldsymbol{x} \in \mathcal{X} : \gamma_j(\boldsymbol{x}) = 0\}, \end{aligned} \tag{3.3}$$

considerando que $\bigcap_{l \in \emptyset} \mathcal{X}_l = \mathcal{X}$.

Exemplo 3.2 *As representações em árvores exploradas na seção anterior são diretamente aplicáveis a $\hat{\boldsymbol{w}}$ e $\hat{y}$ e, para um determinado indivíduo, podemos pensar que a classificação se dá de forma hierárquica. Tome como exemplo o caso 1 da Figura 2.3. Se, para um certo indivíduo, $\hat{w}_1 = 1$, então o classificamos como $\hat{y} = 1$, independentemente se $\hat{w}_2 = 0$ ou $\hat{w}_2 = 1$. Agora, se $\hat{w}_1 = 0$, então se $\hat{w}_2 = 1$, o classificamos como $\hat{y} = 2$ e, se $\hat{w}_2 = 0$, como $\hat{y} = 3$. Neste caso, as 3 regiões de classificação em $\mathcal{X}$ são determinadas como segue. Primeiro note que $\lambda = \{\Lambda_1, \Lambda_2\}$, em que $\Lambda_1 = \{A_{11}, A_{12}\}$*

e $\Lambda_2 = \{A_{21}, A_{22}\}$*, sendo* $A_{11} = \{1\}, A_{12} = \{2,3\}, A_{21} = \{2\}$ *e* $A_{22} = \{3\}$*. Disso: para* $k = 1$ *temos que* $J_{11} = \{1\}$ *e* $J_{12} = \emptyset$*; para* $k = 2$ *temos que* $J_{21} = \{2\}$ *e* $J_{22} = \{1\}$*; e para* $k = 3$ *temos que* $J_{31} = \emptyset$ *e* $J_{32} = \{1,2\}$*. Da Equação (3.2), temos que*

$$\begin{aligned} \mathcal{X}_1 &= \mathcal{X}_{11} \\ \mathcal{X}_2 &= \mathcal{X}_{21} \cap \mathcal{X}_{12} \\ \mathcal{X}_3 &= \mathcal{X}_{12} \cap \mathcal{X}_{22}, \end{aligned}$$

sendo $\mathcal{X}_{11} = \{\boldsymbol{x} \in \mathcal{X}; \hat{w}_1 = 1\}$*,* $\mathcal{X}_{12} = \{\boldsymbol{x} \in \mathcal{X}; \hat{w}_1 = 0\}$*,* $\mathcal{X}_{21} = \{\boldsymbol{x} \in \mathcal{X}; \hat{w}_2 = 1\}$ *e* $\mathcal{X}_{22} = \{\boldsymbol{x} \in \mathcal{X}; \hat{w}_2 = 0\}$ *(Equação (3.3)). Dessa forma, obtemos*

$$\begin{aligned} \mathcal{X}_1 &= \{\boldsymbol{x} \in \mathcal{X}; \hat{w}_1 = 1\} \\ \mathcal{X}_2 &= \{\boldsymbol{x} \in \mathcal{X}; \hat{w}_1 = 0 \text{ e } \hat{w}_2 = 1\} \\ \mathcal{X}_3 &= \{\boldsymbol{x} \in \mathcal{X}; \hat{w}_1 = 0 \text{ e } \hat{w}_2 = 0\}. \end{aligned}$$

Na seção anterior tivemos que dar atenção ao fato de que se $w_1 = 1$*, então* w_2 *não estava bem definida. Contudo, podemos observar o vetor de covariáveis de um determinado indivíduo, digamos* $\boldsymbol{x}$*, em* $\mathcal{X}_{11} \cap \mathcal{X}_{21}$ *e assim* $\hat{w}_1 = 1$ *e* $\hat{w}_2 = 1$*. Mas não há ambiguidades em casos como este, pois* $\boldsymbol{x} \in \mathcal{X}_{11} \cap \mathcal{X}_{21} \Rightarrow \boldsymbol{x} \in \mathcal{X}_{11} \Rightarrow \boldsymbol{x} \in \mathcal{X}_1 \Rightarrow \hat{y} = 1$*.*

Propor classificadores binomiais, em geral, é tarefa mais fácil do que multinomiais. Para uma classe de classificadores lineares por exemplo, suponha que temos duas covariáveis reais x_1 e x_2. No caso dicotômico, podemos propor uma reta $\eta_1^* = \beta_{10} + \beta_{11}x_1^* + \beta_{12}x_2^*$, sendo β_{10}, β_{11} e β_{12} coeficientes da reta, tal que $\mathcal{X}_{11} = \{(x_1, x_2) \in \mathbb{R}^2; \eta_1 \leq \eta_1^*\}$ e $\mathcal{X}_{12} = \{(x_1, x_2) \in \mathbb{R}^2; \eta_1 > \eta_1^*\}$, em que $\eta_1 = \beta_{10} + \beta_{11}x_1 + \beta_{12}x_2$. No caso politômico, devemos propor $\mathcal{X}_1, \mathcal{X}_2, ..., \mathcal{X}_D$ conjuntamente, através de d retas (ou semirretas, ou segmentos de reta). A vantagem do que vimos discutindo até aqui, vai no sentido de que, de forma independente de η_1^*, podemos propor uma segunda reta η_2^* tal que $\mathcal{X}_{21} = \{(x_1, x_2) \in \mathbb{R}^2; \eta_2 \leq \eta_2^*\}$ e $\mathcal{X}_{12} = \{(x_1, x_2) \in \mathbb{R}^2; \eta_2 > \eta_2^*\}$, e assim por diante até $\mathcal{X}_{d1}$ e $\mathcal{X}_{d2}$, e então obter uma proposta de classificador multinomial via Equação (3.2).

A literatura com repeito a classificação binomial é bem mais vasta do que a multinomial. Em modelos de regressão, por exemplo, – que é o foco do nosso trabalho na próxima seção – desenvolvimentos como a flexibilização nas funções de ligação são quase sempre feitos para Y seguindo uma distribuição de Bernoulli (ou binomial), a fim de se modelar a probabilidade de sucesso como função de $\boldsymbol{x}$ e, a partir de estimativas de tal função, propor um classificador dicotômico. Além disso, outras medidas de acurácia como sensibilidade, especificidade, valor preditivo positivo e valor preditivo negativo podem ser incorporados mais diretamente no caso binomial.

Antes de explorarmos os modelos de regressão multinomial através de suas fatorações, discutimos a seguir um procedimento geral para buscar e avaliar a performance de classificadores com base na minimização da função de risco e como a modelagem de $\boldsymbol{p}$, supondo que $(Y|\boldsymbol{x};\boldsymbol{p}) \sim BM_D(\boldsymbol{p})$, é importante na proposta de tais classificadores. Não é nossa intenção explorar os resultados gerais da teoria de classificação e sim como fatorar um problema multinomial em problemas binomiais, o que pode ser vantajoso e também os problemas encontrados nesta fatoração. Alguns trabalhos que exploram a teoria em geral são Izbicki (2018) e literaturas clássicas em *Statistical Learning* como Hastie *et al.* (2016) e Vapnik (1998), dentre várias outras.

3.1.2 Busca e performance de classificadores

A função de risco ao qual desejamos minimizar é definida como a esperança da função de perda. Definimos tais funções a seguir.

Definição 3.2 *Sejam $(Y, \boldsymbol{X})$ assumindo valores em $\Omega \times \mathcal{X}$ e δ um classificador em Ω. Definimos como função de perda (ou simplesmente perda) uma função $\mathcal{L}$ real qualquer, definida em $\Omega \times \Omega$.*

Comumente utiliza-se a função de perda denominada perda $0-1$, definida por

$$\mathcal{L}(Y, \delta(\boldsymbol{X})) = \mathbb{1}(Y \neq \delta(\boldsymbol{X})),$$

ou seja, a função que atribui uma penalização igual a 1 se $Y \neq \delta(\boldsymbol{X})$, independentemente da categoria de Ω. Note que de uma maneira geral, trata-se $\boldsymbol{X}$ como vetor aleatório e $\hat{Y} = \delta(\boldsymbol{X})$ como variável aleatória em $\{0, 1\}$.

Definição 3.3 *Sejam $(Y, \boldsymbol{X})$ assumindo valores em $\Omega \times \mathcal{X}$, δ um classificador em Ω e $\mathcal{L}$ uma função de perda. Definimos por função de risco (ou simplesmente risco) a esperança de $\mathcal{L}$, ou seja,*

$$R(\delta) = \mathbb{E}[\mathcal{L}(Y, \delta(\boldsymbol{X}))].$$

Assim, para a perda 0-1, temos que

$$R(\delta) = \mathbb{E}[\mathbb{1}(Y \neq \delta(\boldsymbol{X}))] = Pr(Y \neq \delta(\boldsymbol{X})).$$

Neste sentido, um bom classificador é aquele que minimiza a probabilidade de erro na classificação.

Teorema 3.1 *A função de classificação δ que minimiza $R(\delta) = \mathbb{E}[\mathbb{1}(Y \neq \delta(\boldsymbol{X}))]$ é*

$$\delta(\boldsymbol{x}) = \arg\max_{k \in \Omega} Pr(Y = k | \boldsymbol{x}).$$

Demostração *De fato. Primeiramente, suponha que a função de densidade marginal de $\boldsymbol{X}$ é denotada por f. Assim,*

$$\begin{aligned} R(\delta) = Pr(Y \neq \delta(\boldsymbol{X})) &= \int_{\mathcal{X}} Pr(Y \neq \delta(\boldsymbol{x}) | \boldsymbol{x}) f(\boldsymbol{x}) d\boldsymbol{x} \\ &= \int_{\mathcal{X}} \left[1 - Pr(Y = \delta(\boldsymbol{x}) | \boldsymbol{x})\right] f(\boldsymbol{x}) d\boldsymbol{x} \\ &= \int_{\mathcal{X}} \left[1 - \sum_{k=1}^{D} \mathbb{1}(\delta(\boldsymbol{x}) = k) Pr(Y = k | \boldsymbol{x})\right] f(\boldsymbol{x}) d\boldsymbol{x} \end{aligned}$$

A terceira igualdade se dá pois $Pr(Y = \delta(\boldsymbol{x})|\boldsymbol{x}) = Pr(Y = k|\boldsymbol{x})$, se $\delta(\boldsymbol{x}) = k$.
Então, buscamos δ que maximize $\sum_{k=1}^{D} \mathbb{1}(\delta(\boldsymbol{x}) = k) Pr(Y = k|\boldsymbol{x})$.
Note que se $\mathbb{1}(\delta(\boldsymbol{x}) = k) = 1$, então $\mathbb{1}(\delta(\boldsymbol{x}) = j) = 0, j \neq k$. Logo,

$$\delta(\boldsymbol{x}) = k, \text{ se } Pr(Y = k|\boldsymbol{x}) > Pr(Y = j|\boldsymbol{x}), j \neq k, k = 1, 2, ..., D.$$

■

Do resultado anterior, é bastante razoável obter estimativas para $Pr(Y = k|\boldsymbol{x})$ a partir de observações de $(Y, \boldsymbol{X})$ e a partir delas propor um classificador. Tal classificador é muitas vezes denominado classificador de Bayes, por se basear na distribuição condicional de Y dado $\boldsymbol{x}$, porém nada tem a ver com o procedimento Bayesiano de estimação. Omitimos no teorema anterior, por facilidade algébrica, a notação que usamos no capítulo anterior, contudo seguimos supondo que $(Y|\boldsymbol{x};\boldsymbol{p}) \sim BM_D(\boldsymbol{p})$ e, dessa forma, a partir de estimativas para $\boldsymbol{p}$, digamos $\hat{\boldsymbol{p}}$ – para um determinado indivíduo – podemos classificá-lo por $\delta(\boldsymbol{x}) = \arg\max_{k\in\Omega} \hat{\boldsymbol{p}}$. Como discutiremos em modelos de regressão adiante, escrevemos $\boldsymbol{p}$ como função de $\boldsymbol{x}$ e buscamos estimativas para tal função.

Contudo, muitas vezes, apesar do resultado do teorema anterior, visamos buscar um classificador com base na minimização de uma estimativa do risco, uma vez que o valor $\mathbb{E}[\mathcal{L}(Y, \delta(\boldsymbol{X}))]$ é desconhecido. Suponha que temos n observações independentes de $(Y, \boldsymbol{X})$, ou seja, $\mathcal{D} = \{(y_i, \boldsymbol{x}_i), i = 1, 2, ..., n\}$. A estimativa amostral do risco para um classificador δ, com uma perda qualquer $\mathcal{L}$, é dada por

$$\hat{R}(\delta) = \frac{1}{n}\sum_{i=1}^{n} \mathcal{L}(y_i, \delta(\boldsymbol{x}_i)), \tag{3.4}$$

e podemos buscar δ que minimiza $\hat{R}(\delta)$, digamos $\hat{\delta}$. Em geral precisamos especificar uma classe de classificadores na qual δ pertence para tornar tal tarefa factível.

No entanto, avaliar a performance de $\hat{\delta}$ com base em $\hat{R}(\delta)$ é otimista, uma vez que $\hat{\delta}$ é justamente aquele – dentro da classe especificada – que minimiza $\hat{R}(\delta)$. Há alguns procedimentos para evitar este problema, dividindo-se a amostra em duas, geralmente denominadas por amostra de treino e amostra de teste. Suponha, sem perda de generalidade, que $I_{tr} = \{1, 2, ..., n_0\}$ e $I_{te} = \{n_0 + 1, n_0 + 2, ..., n\}$ são os indivíduos que compõem a amostra de treino e a de teste, respectivamente, e assim, $\mathcal{D}_{tr} = \{(y_i, \boldsymbol{x}_i), i \in I_{tr}\}$ é a amostra de treino e $\mathcal{D}_{te} = \{(y_i, \boldsymbol{x}_i), i \in I_{te}\}$ a de teste. Assim, podemos buscar $\hat{\delta}$ que minimiza $\frac{1}{n_0}\sum_{i=1}^{n_0} \mathcal{L}(y_i, \delta(\boldsymbol{x}_i))$ e avaliar sua performance através de $\frac{1}{n-n_0}\sum_{i=n_0+1}^{n} \mathcal{L}(y_i, \delta(\boldsymbol{x}_i))$.

Ainda podemos enfrentar o problema de que, por acaso, uma determinada amostra de teste favoreça $\hat{\delta}$ e, a fim de se evitar tal problema, é recomendável que se divida a amostra aleatoriamente um número grande de vezes, digamos B, obtendo-se $(\mathcal{D}_{tr_1}, \mathcal{D}_{te_1})$, $(\mathcal{D}_{tr_2}, \mathcal{D}_{te_2})$, ..., $(\mathcal{D}_{tr_B}, \mathcal{D}_{te_B})$, e para cada $b = 1, 2, ..., B$, buscar $\hat{\delta}$ via $\mathcal{D}_{tr_b}$ que minimiza

$$r_{tr_b} = \frac{1}{\#\{I_{tr_b}\}}\sum_{i\in I_{tr_b}} \mathcal{L}(y_i, \delta(\boldsymbol{x}_i)),$$

e avaliar sua performance via $\mathcal{D}_{te_b}$ por

$$r_{te_b} = \frac{1}{\#\{I_{te_b}\}}\sum_{i\in I_{te_b}} \mathcal{L}(y_i, \delta(\boldsymbol{x}_i)),$$

em que $\#\{I_{tr_b}\}$ e $\#\{I_{te_b}\}$ são, respectivamente, a quantidade de indivíduos em I_{tr_b} e I_{te_b}, obtendo-se assim uma estimativa empírica para o risco, dada por

$$\hat{R}_{te}(\delta) = \frac{1}{B}\sum_{b=1}^{B} r_{te_b}. \tag{3.5}$$

Para a perda 0-1, por exemplo, temos que

$$\hat{R}_{te}(\delta) = \frac{1}{B}\sum_{b=1}^{B}\left[\frac{1}{\#\{I_{te_b}\}}\sum_{i\in I_{te_b}}\mathbb{1}(y_i \neq \delta(\boldsymbol{x}_i))\right]. \tag{3.6}$$

Note que, para cada $b = 1, 2, ..., B$, obtemos $\hat{\delta}_b$, contudo desejamos propor um único classificador. Podemos, após este processo de busca e validação, propor como classificador aquele que minimiza $\hat{R}$ dado na Equação (3.4), utilizando $\hat{R}_{te}$ como estimativa do risco, dada na Equação (3.5).

3.1.3 Outras medidas de acurácia

Podemos utilizar outras medidas de acurácia, como definido a seguir.

Definição 3.4 *Seja $(Y, \boldsymbol{X})$ assumindo valores em $(\Omega, \mathcal{X})$ e δ um classificador. Definimos como acurácia, sensibilidade, especificidade, valor preditivo positivo e valor preditivo negativo, para $k = 1, 2, ..., D$, por, respectivamente,*

$$\begin{aligned}
AC(\delta) &= Pr(Y = \delta(\boldsymbol{X})) \\
SEN_k(\delta) &= Pr(\delta(\boldsymbol{X}) = k | Y = k) \\
ESP_k(\delta) &= Pr(\delta(\boldsymbol{X}) \neq k | Y \neq k) \\
VPP_k(\delta) &= Pr(Y = k | \delta(\boldsymbol{X}) = k) \\
VPN_k(\delta) &= Pr(Y \neq k | \delta(\boldsymbol{X}) \neq k).
\end{aligned}$$

Considerando-se $\Omega = \{0, 1\}$ e γ um classificador dicotômico, note que temos $SEN_1(\gamma) = ESP_0(\gamma)$ (e vice-versa) e também $VPP_1(\gamma) = VPN_0(\gamma)$ (e vice-versa). A nomenclatura "sensibilidade", "especificidade", "valor preditivo positivo" e "valor preditivo negativo" vem principalmente da área médica, em que se propõe um teste médico $\gamma(\boldsymbol{X})$ para detectar a presença de uma doença ($Y = 1$) ou sua ausência ($Y = 0$). No entanto, estamos tratando 0 e 1 como rótulos quaisquer e, assim tal nomenclatura perde o sentido. Contudo, a mantemos no nosso trabalho.

Note também que podemos definir um risco qualquer com base na distância (absoluta, por exemplo) entre AC e 1 (risco via perda 0-1), mas também para qualquer outra medida de acurácia, como por exemplo $R(\delta) = \sum_{k=1}^{D} |SEN_k(\delta) - 1|$, ou ainda levar em consideração mais de uma medida de acurácia, como por exemplo $R(\delta) = |AC(\delta) - 1| + \sum_{k=1}^{D}(|SEN_k(\delta) - 1| + |VPP_k(\delta) - 1|)$. Há diversas generalizações neste ponto, podendo-se incluir diferentes pesos para diferentes medidas de acurácia e diferentes categorias, ou ainda diversas outras medidas de acurácia, por exemplo, podemos tratar de forma diferente $Pr(\delta(\boldsymbol{X}) = k | Y = j_1)$ e $Pr(\delta(\boldsymbol{X}) = k | Y = j_2), j_1 \neq j_2 \neq k$.

Estimativas empíricas – como a Equação (3.6) para AC – para SEN_k, ESP_k, VPP_k e VPN_k

são dadas, respectivamente, por

$$\begin{aligned}
\widehat{SEN_k}(\delta) &= \tfrac{1}{B}\sum_{b=1}^{B}\left[\tfrac{1}{\#\{i\in I_{te_b};y_i=k\}}\sum_{i\in I_{te_b};y_i=k}\mathbb{1}(y_i=\delta(\boldsymbol{x}_i))\right]\\
\widehat{ESP_k}(\delta) &= \tfrac{1}{B}\sum_{b=1}^{B}\left[\tfrac{1}{\#\{i\in I_{te_b};y_i\neq k\}}\sum_{i\in I_{te_b};y_i\neq k}\mathbb{1}(\delta(\boldsymbol{x}_i)\neq k)\right]\\
\widehat{VPP_k}(\delta) &= \tfrac{1}{B}\sum_{b=1}^{B}\left[\tfrac{1}{\#\{i\in I_{te_b};\delta(\boldsymbol{x}_i)=k\}}\sum_{i\in I_{te_b};\delta(\boldsymbol{x}_i)=k}\mathbb{1}(y_i=\delta(\boldsymbol{x}_i))\right]\\
\widehat{VPN_k}(\delta) &= \tfrac{1}{B}\sum_{b=1}^{B}\left[\tfrac{1}{\#\{i\in I_{te_b};\delta(\boldsymbol{x}_i)\neq k\}}\sum_{i\in I_{te_b};\delta(\boldsymbol{x}_i)\neq k}\mathbb{1}(y_i\neq k)\right].
\end{aligned}$$

Propor um classificador multinomial que leve em consideração todas essas medidas pode ser complicado, ainda mais se houver um número grande de categorias. Por outro lado, o caso binomial é relativamente tarefa mais simples. Por este ponto de vista, construir classificadores binomiais baseando-se em tais medidas de acurácia nos leva a construir um classificador multinomial, via $\boldsymbol{v}_\lambda$, que as leve em consideração também.

Note que para o caso binomial, temos que $AC(\gamma) = Pr(Y = \gamma(\boldsymbol{X})\}$, $SEN_1(\gamma) = Pr(\gamma(\boldsymbol{X}) = 1|Y = 1\}$, $SEN_0(\gamma) = Pr(\gamma(\boldsymbol{X}) = 0|Y = 0\}$, $VPP_1(\gamma) = Pr(Y = 1|\gamma(\boldsymbol{X}) = 1\}$ e $VPP_0(\gamma) = Pr(Y = 0|\gamma(\boldsymbol{X}) = 0\}$. Neste sentido, mostramos a seguir que $SEN_k(\delta)$ pode ser escrita através de $SEN_1(\gamma_j)$ e $SEN_0(\gamma_j)$, e analogamente, $VPP_k(\delta)$ através de $VPP_1(\gamma_j)$ e $VPP_0(\gamma_j)$. De fato, notando que W_j é independente de W_l e também que $\gamma_j(\boldsymbol{X})$ é independente de $\gamma_l(\boldsymbol{X})$, para todo $j \neq l$, temos que

$$\begin{aligned}
SEN_k(\delta) &= Pr(\delta(\boldsymbol{X}) = k|Y = k)\\
&= \frac{Pr(\delta(\boldsymbol{X}) = k, Y = k)}{Pr(Y = k)}\\
&= \frac{Pr\left(\{\gamma_j(\boldsymbol{X}) = 1; j \in J_{k1}\}, \{\gamma_j(\boldsymbol{X}) = 0; j \in J_{k2}\}, \{W_j = 1; j \in J_{k1}\}, \{W_j = 0; j \in J_{k2}\}\right)}{Pr\left(\{W_j = 1; j \in J_{k1}\}, \{W_j = 0; j \in J_{k2}\}\right)}\\
&= \prod_{j\in J_{k1}}\frac{Pr(\gamma_j(\boldsymbol{X}) = 1, W_j = 1)}{Pr(W_j = 1)}\prod_{j\in J_{k2}}\frac{Pr(\gamma_j(\boldsymbol{X}) = 0, W_j = 0)}{Pr(W_j = 0)}\\
&= \prod_{j\in J_{k1}}Pr(\gamma_j(\boldsymbol{X}) = 1|W_j = 1)\prod_{j\in J_{k2}}Pr(\gamma_j(\boldsymbol{X}) = 0|W_j = 0)\\
&= \prod_{j\in J_{k1}}SEN_1(\gamma_j)\prod_{j\in J_{k2}}SEN_0(\gamma_j)
\end{aligned}$$

e, analogamente,

$$VPP_k(\delta) = \prod_{j\in J_{k1}}VPP_1(\gamma_j)\prod_{j\in J_{k2}}VPP_0(\gamma_j).$$

Exemplo 3.3 *Considere* $\Omega = \{1, 2, 3\}$ *e* $\lambda(\Omega) = \{\Lambda_1, \Lambda_2\}$ *com* $\Lambda_1 = \{\{1\}, \{2, 3\}\}$ *e* $\Lambda_2 = \{\{2\}, \{3\}\}$. *Defina* $\gamma_1(\boldsymbol{x}) = 0$ *e* $\gamma_2(\boldsymbol{x}) = 1$, *ou seja, classificadores dicotômicos triviais que independentes do valor* $\boldsymbol{x}$ *classifica o respectivo sujeito, respectivamente, em* 0 *e* 1. *Note que temos* $SEN_0(\gamma_1) = 1$ *e* $SEN_1(\gamma_2) = 1$. *Agora a partir de* $\delta(\boldsymbol{x}) = \boldsymbol{v}_\lambda(\gamma_1(\boldsymbol{x}), \gamma_2(\boldsymbol{x}))$, *isto é:* $\delta(\boldsymbol{x}) = 1$ *se* $\gamma_1(\boldsymbol{x}) = 1$; $\delta(\boldsymbol{x}) = 2$ *se* $\gamma_1(\boldsymbol{x}) = 0$ *e* $\gamma_2(\boldsymbol{x}) = 1$; *e* $\delta(\boldsymbol{x}) = 3$ *se* $\gamma_1(\boldsymbol{x}) = 0$ *e* $\gamma_2(\boldsymbol{x}) = 0$; *temos que* $\delta(\boldsymbol{x}) = 2$ *para qualquer* $\boldsymbol{x}$. *Desse modo, notando que* $J_{21} = \{2\}$ *e* $J_{22} = \{1\}$, *temos que* $SEN_2(\delta) = SEN_1(\gamma_2)SEN_0(\gamma_1) = 1$.

3.2 Regressão

O problema geral da regressão para dados politômicos é construir uma função que relacione um vetor de covariáveis observadas $\boldsymbol{x}$ com a esperança condicional da variável resposta Y dado $\boldsymbol{x}$, no nosso caso $(Y|\boldsymbol{x};\boldsymbol{p}) \sim BM_D(\boldsymbol{p})$. Ou seja, da Equação (2.4), temos que a função de regressão ao qual desejamos modelar é dada por

$$E[Y|\boldsymbol{x};\boldsymbol{p}] = \boldsymbol{p}.$$

Ou seja, a partir de observações de $(Y, \boldsymbol{X}) \in \Omega \times \mathcal{X}$, desejamos modelar uma função

$$\begin{array}{rccccl} \boldsymbol{h}: & \mathcal{X} & \to & \mathscr{S}^d & & \\ & \boldsymbol{x} & \mapsto & \boldsymbol{p} & = & \boldsymbol{h}(\boldsymbol{x}) = (h_1(\boldsymbol{x}), h_2(\boldsymbol{x}), ..., h_D(\boldsymbol{x})) \end{array}$$

a fim de se obter uma estimativa $\hat{\boldsymbol{h}}$ e escrever

$$\hat{\boldsymbol{p}} = \hat{\boldsymbol{h}}(\boldsymbol{x}).$$

Observação *De uma forma mais geral, pode-se considerar que* $(Y|\boldsymbol{x};n;\boldsymbol{p})$ *segue uma distribuição Multinomial com parâmetros* n *e* p*, no entanto nos restringimos ao caso Bernoulli.*

Em geral devemos restringir a classe em que $\boldsymbol{h}$ pertence. Tome o caso dicotômico em que, digamos, $(W|\boldsymbol{x};\theta) \sim B(\theta)$ em que buscamos modelar g, tal que $\theta = g(\boldsymbol{x})$. Supomos, em geral, que

$$\theta = F(\eta(\boldsymbol{\beta}, \boldsymbol{x})) \tag{3.7}$$

em que $\eta(\boldsymbol{\beta}, \boldsymbol{x}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_m x_m$, sendo $\boldsymbol{\beta} = (\beta_0, \beta_1, ..., \beta_m) \in \mathbb{R}^{m+1}$ parâmetros da regressão e F uma função de distribuição. Isto é, restringimos g como sendo da forma $F(\eta)$, $\eta \in \mathbb{R}$. O caso mais utilizado, explorado em diversos trabalhos, como em Cox (1958) por exemplo, é quando se supõe que F é a função de distribuição logística padrão, isto é,

$$F(\eta) = \frac{\exp(\eta)}{1 + \exp(\eta)},$$

em que η denota $\eta(\boldsymbol{\beta}, \boldsymbol{x})$, tendo inversa dada pelo log-odds, ou seja

$$\eta = \log\left(\frac{\theta}{1-\theta}\right), \tag{3.8}$$

denominada função de ligação logito. Há trabalhos ainda mais antigos que exploram tal relação logística, como Pearl e Reed (1920) e Reed e Berkson (1929), mas principalmente estudos em *bio-assays* – buscando-se relacionar doses de um medicamento à proporção de resposta dos indivíduos – exploraram este ponto no início de seu desenvolvimento, como Berkson (1944). Entretanto pode-se utilizar outras funções de distribuição, como a normal padrão, denotada por Φ, proposta por Bliss (1935) também em ensaios de dose-resposta, escrevendo-se $\theta = \Phi(\eta)$, com inversa $\eta = \Phi^{-1}(\theta)$, denominada função de ligação probito.

Neste sentido, há uma variedade de funções com características particulares, cada qual com suas vantagens. Trabalhos como Prentice (1976), Aranda-Ordaz (1981), Stukel (1988), Caron e Polpo (2009), Diniz (2015) e Eugenio (2016) exploram alternativas à logito e probito, bem como à com-

plementar log-log explorada em Fisher (1922).

Contudo, para o caso multinomial, não podemos escrever que $p_k = h_k(\boldsymbol{x}) = F_k(\eta(\boldsymbol{\beta}_k, \boldsymbol{x})), k = 1, 2, ..., D$, uma vez que feriríamos a restrição $\sum_{k=1}^{D} p_k = 1$ de $\mathscr{S}^d$.

Uma solução é fruto da nossa discussão na seção anterior, ou seja, dada uma bipartição total $\lambda(\Omega) = \{\Lambda_1, \Lambda_2, ..., \Lambda_d\}$ com $\Lambda_j = \{A_{j1}, A_{j2}\}$, a partir das Equações (2.18) e (3.7) podemos escrever

$$\boldsymbol{p} = \boldsymbol{t}_\lambda(\boldsymbol{\theta}) = \boldsymbol{t}_\lambda(F_1(\eta_1), F_2(\eta_2), ..., F_d(\eta_d)),$$

em que $\eta_j = \eta(\boldsymbol{\beta}_j, \boldsymbol{x}) = \beta_{j0} + \beta_{j1}x_1 + \beta_{j2}x_2 + \cdots + \beta_{jm}x_m$ e F_j uma função de distribuição, e desse modo, da Equação (2.14), $\boldsymbol{h}$ é restrita a seguinte forma

$$p_k = h_k(\boldsymbol{x}) = \prod_{j \in J_{k1}} F_j(\eta_j) \prod_{j \in J_{k2}} (1 - F_j(\eta_j)) \qquad k = 1, 2, ..., D, \tag{3.9}$$

em que $J_{k1} = \{j \in \{1, 2, ..., d\}; k \in A_{j1}\}$ e $J_{k2} = \{j \in \{1, 2, ..., d\}; k \in A_{j2}\}$.

Estimação dos parâmetros

Assim, desejamos obter estimativas $\hat{\boldsymbol{\beta}}_j, j = 1, 2, ..., d$, a fim de se obter

$$\hat{p}_k = \prod_{j \in J_{k1}} F_j(\eta(\hat{\boldsymbol{\beta}}_j, \boldsymbol{x})) \prod_{j \in J_{k2}} [1 - F_j(\eta(\hat{\boldsymbol{\beta}}_j, \boldsymbol{x}))],$$

e visando obter uma estimativa de máxima verossimilhança, podemos proceder como segue. Da Equação (2.21) temos que, dadas n observações independentes $\mathcal{D} = \{(y_i, \boldsymbol{x}_i), i = 1, 2, ..., n\}$, λ induz $\mathcal{D}_{j_\lambda} = \{(w_{ji}, \boldsymbol{x}_i), i \in I_j\}$, em que, como na Equação (2.20), $I_j = \{i \in \{1, 2, ..., n\}; y_i \in \Lambda_j\}$, e assim, como visto na Equação (2.22), podemos escrever que

$$\begin{aligned} L(\boldsymbol{\beta}|\mathcal{D}) &= \prod_{j=1}^{d} L_j(\boldsymbol{\beta}_j|\mathcal{D}_{j_\lambda}) \\ &= \prod_{j=1}^{d} \prod_{i \in I_j} F_j(\eta(\boldsymbol{\beta}_j, \boldsymbol{x}_i))^{w_{ji}} (1 - F_j(\eta(\boldsymbol{\beta}_j, \boldsymbol{x}_i)))^{(1-w_{ji})}, \end{aligned} \tag{3.10}$$

em que $\boldsymbol{\beta}$ é a matriz $(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, ..., \boldsymbol{\beta}_d)$. Maximizando $L(\boldsymbol{\beta}|\mathcal{D})$ obtemos uma estimativa para $\boldsymbol{\beta}$ ou, de forma mais simples, maximizando separadamente cada $L_j(\boldsymbol{\beta}_j|\mathcal{D}_{j_\lambda})$, obtemos estimativas para $\boldsymbol{\beta}_j$, $j = 1, 2, ..., d$. No capítulo seguinte, na aplicação, utilizamos estimativas de máxima verossimilhança para cada regressão binomial.

Regressão multinomial logística com categoria de referência

A forma mais usual de se ligar $\boldsymbol{x}$ a $\boldsymbol{p}$ na literatura é uma forma similar à função logística, como na Equação (3.8), da seguinte forma

$$\log\left(\frac{p_k}{p_D}\right) = \eta(\boldsymbol{\beta}_k, \boldsymbol{x}), \qquad k = 1, 2, ..., d, \tag{3.11}$$

com inversa dada por

$$\begin{aligned} p_k &= \frac{\exp(\eta_k)}{1+\sum_{j=1}^{d}\exp(\eta_j)} \qquad k=1,2,...,d \\ p_D &= \frac{1}{1+\sum_{j=1}^{d}\exp(\eta_j)}. \end{aligned} \tag{3.12}$$

Contudo, representada dessa forma, não existe uma flexibilização natural para a função de ligação probito, por exemplo, como argumentam Aitchison *et al.* (2004) e, neste sentido, a Equação (3.9) pode trazer toda a flexibilização desenvolvida no contexto binomial para a modelagem multinomial.

A ligação da equação anterior ganhou popularidade no contexto de Modelos Lineares Generalizados (MLG), proposto em Nelder e Wedderburn (1972). Definimos a seguir a família exponencial discreta (Andersen, 1970), para a qual MLG é proposto, e mostramos que a Bernoulli multivariada pertence a tal família.

Definição 3.5 *A família exponencial discreta s-dimensional é formada por distribuições cuja função de probabilidade pode ser escrita da forma*

$$Pr(Y=y|\boldsymbol{p}) = \exp\left\{\sum_{k=1}^{s} a_s(\boldsymbol{p})b_s(y) + c(y) + d(\boldsymbol{p})\right\},$$

em que a_k, b_k, $k=1,2,...,s$, c e d são funções reais.

Teorema 3.2 *A distribuição de Bernoulli D-variada pertence a família exponencial discreta d-dimensional.*

Demostração *De fato,*

$$\begin{aligned} Pr(Y=y|\boldsymbol{p}) &= \prod_{k=1}^{D} p_k^{\mathbb{1}(y=k)} \\ &= \exp\left\{\log\left[\prod_{k=1}^{D} p_k^{\mathbb{1}(y=k)}\right]\right\} \\ &= \exp\left\{\sum_{k=1}^{D}\mathbb{1}(y=k)\log p_k\right\} \\ &= \exp\left\{\left[\sum_{k=1}^{d}\mathbb{1}(y=k)\log p_k\right] + \mathbb{1}(y=D)\log p_D\right\} \\ &= \exp\left\{\left[\sum_{k=1}^{d}\mathbb{1}(y=k)\log p_k\right] + \left(1-\sum_{k=1}^{d}\mathbb{1}(y=k)\right)\log p_D\right\} \\ &= \exp\left\{\left[\sum_{k=1}^{d}\log\left(\frac{p_k}{p_D}\right)\mathbb{1}(y=k)\right] + \log p_D\right\}, \end{aligned} \tag{3.13}$$

e, assim, $a_k(\boldsymbol{p}) = \log\left(\frac{p_k}{p_D}\right)$, $b_k(y) = \mathbb{1}(y=k)$, $k=1,2,...,d$, $c(y)=0$ e $d(\boldsymbol{p}) = \log p_D$, como queríamos demostrar.

■

Note, portanto, que o termo $\log\left(\frac{p_k}{p_D}\right)$ aparece de forma "natural" (neste contexto ele é chamado de parâmetro natural). O procedimento usual segue do fato que, supondo-se ainda que $\mathcal{D} = \{(y_i, \boldsymbol{x}_i), i = 1, 2, ..., n\}$ é um conjunto de observações independentes, podemos escrever a verossimilhança, a partir das Equações (3.11), (3.12) e (3.13), por

$$L(\boldsymbol{\beta}|\mathcal{D}) = \prod_{i=1}^{n} \exp\left\{\left[\sum_{k=1}^{d} \eta(\boldsymbol{\beta}_k, \boldsymbol{x}_i)\mathbb{1}(y_i = k)\right] - \log\left[1 + \sum_{k=1}^{d} \exp(\eta(\boldsymbol{\beta}_k, \boldsymbol{x}_i))\right]\right\}.$$

Como descrito em diversos trabalhos elementares, como Agresti (2002) e Hosmer *et al.* (2013), prossegue-se com a maximização de $L(\boldsymbol{\beta}|\mathcal{D})$ a fim de se obter a estimativa de máxima verossimilhança $\hat{\boldsymbol{\beta}}$ e, a partir da Equação (3.12), obter $\hat{\boldsymbol{p}}$ para um determinado indivíduo. Esta formulação é a mais aplicada nas mais diversas áreas, denominada regressão multinomial logística com categoria de referência.

Note, contudo, que a função de distribuição logística padrão está implícita nas Equações (3.11) e (3.12), deixando implícito que também é preciso decompor o problema D-variado d dicotômicos, assim como feito a partir de uma bipartição total. Para notar tal fato, defina

$$\pi_k = \frac{p_k}{p_k + p_D}, \qquad k = 1, 2, ..., d, \tag{3.14}$$

e assim, como na Equação (2.12), podemos definir

$$Z_k = \mathbb{1}(Y = k|Y \in \{k, D\}) = \begin{cases} 1, & \text{se } Y = k \\ 0, & \text{se } Y = D \end{cases},$$

e então, como na Equação (2.13), podemos definir

$$\pi_k = Pr(Y = k|Y \in \{k, D\}; \boldsymbol{p}),$$

e portanto, $Z_k|\pi_k \sim B(\pi_k)$.

Assim, a partir da Equação (3.11), temos que

$$\eta_k = \log\left(\frac{p_k}{p_D}\right) = \log\left(\frac{\pi_k}{1 - \pi_k}\right), \qquad k = 1, 2, ..., d,$$

obtendo que

$$\pi_k = \frac{\exp(\eta_k)}{1 + \exp(\eta_k)}.$$

Desse modo, a formulação usual também requer a decomposição da modelagem multinomial em binomiais. Neste ponto, a flexibilidade com relação a função de ligação também pode ser incorporada na regressão multinomial com categoria de referência, através de

$$\pi_k = F_k(\eta_k),$$

e invertendo a Equação (3.14), podemos escrever, analogamente à Equação (3.9),

$$\begin{aligned} p_k &= \left(\frac{F_k(\eta_k)}{1-F_k(\eta_k)}\right)\frac{1}{1+\sum_{l=1}^{d}\frac{F_k(\eta_l)}{1-F_k(\eta_l)}} \qquad k=1,2,...,d \\ p_D &= \frac{1}{1+\sum_{l=1}^{d}\frac{F_k(\eta_l)}{1-F_k(\eta_l)}}. \end{aligned}$$

Note que tomando F_k como a função de distribuição logística padrão, reduzimos à regressão multinomial logística com categoria de referência. Assim, explicitamos a função de ligação no caso multinomial de duas formas distintas, como uma decomposição do caso binomial e qualquer uma das duas formas pode trazer a generalização e flexibilização para o contexto multinomial.

3.2.1 Da regressão à classificação

Vimos que a partir de d classificadores dicotômicos podemos construir, através de uma bipartição total, um classificador politômico. Podemos construir cada classificador dicotômico $\gamma_j(\boldsymbol{x})$ a partir de estimativas do respectivo vetor de parâmetros da regressão $\boldsymbol{\beta}_j = (\beta_{j0}, \beta_{j1}, ..., \beta_{jm})$, digamos $\hat{\boldsymbol{\beta}}_j$.

Dessa forma, para um determinado indivíduo, com vetor de covariáveis dadas por $\boldsymbol{x} = (x_1, x_2, ..., x_m)$, podemos escrever

$$\hat{\boldsymbol{p}} = \boldsymbol{t}_\lambda(\hat{\boldsymbol{\theta}}) = \boldsymbol{t}_\lambda(F_1(\hat{\eta}_1), F_2(\hat{\eta}_2), ..., F_d(\hat{\eta}_d)),$$

em que

$$\hat{\eta}_j = \hat{\beta}_{j0} + \hat{\beta}_{j1}x_1 + \cdots + \hat{\beta}_{jm}x_m, \quad j = 1, 2, ..., d,$$

e F_j são funções de distribuição para os modelos binomiais, obtendo-se $\hat{\theta}_j = F_j(\hat{\eta}_j)$.

Aqui, há pelo menos dois possíveis caminhos para a classificação. Primeiro, podemos partir de $\hat{\boldsymbol{p}} = (\hat{p}_1, \hat{p}_2, ..., \hat{p}_D)$ e tomar

$$\hat{\delta}(\boldsymbol{x}) = \arg\max_{k\in\Omega} \hat{p}_k.$$

Ou podemos definir

$$\hat{\gamma}_j(\boldsymbol{x}) = \begin{cases} 1, & \text{se } \hat{\theta}_j > \frac{1}{2} \\ 0, & \text{se } \hat{\theta}_j \leq \frac{1}{2} \end{cases}$$

e a partir da Equação (3.1) obter $\delta(\boldsymbol{x})$.

De uma forma geral, podemos escrever

$$\hat{\gamma}_j(\boldsymbol{x}) = \begin{cases} 1, & \text{se } \hat{\theta}_j > \theta_j^* \\ 0, & \text{se } \hat{\theta}_j \leq \theta_j^* \end{cases},$$

a fim de obter melhores classificadores com base nas estimativas de riscos – como discutido na seção anterior –, além da proporção de erros, considerar sensibilidades, especificidades, valores preditivos positivos e negativos, em que θ_j^* deve ser estimado na amostra de treino a fim de se evitar uma avaliação otimista da performance do classificador proposto.

Toda a teoria elaborada com relação a este ponto pode ser estendida ao caso multinomial. Por exemplo, além das medidas de acurácia descritas, curvas ROC podem ser utilizadas a fim de se obter bons classificadores binomiais.

3.2.2 Escore de Brier

Uma medida de acurácia com relação à classificação a partir de estimações do vetor de probabilidades $\boldsymbol{p}$ é considerar uma distância entre o vetor estimado $\hat{\boldsymbol{p}}$ e o que realmente foi observado, $y = k$. Ou seja, desejamos medir uma distância entre o vetor de observações dado por $(\mathbb{1}(y = 1), \mathbb{1}(y = 2), ..., \mathbb{1}(y = D))$ – isto é, o vetor indicador de qual categoria foi observada – e o vetor de probabilidades estimado $\hat{\boldsymbol{p}}$. Brier (1950), no contexto de fazer previsões com relação ao clima, propõe uma distância quadrática média entre o vetor $\hat{\boldsymbol{p}}_i$ e o vetor $(\mathbb{1}(y_i = 1), \mathbb{1}(y_i = 2), ..., \mathbb{1}(y_i = D))$, para n observações, $i = 1, 2, ..., n$, dada por

$$BS(\delta) = \frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{D}(\hat{p}_{ki} - \mathbb{1}(y_i = k))^2,$$

em que δ é o classificador multinomial.

Desse modo, tal medida, denominada escore de Brier, tem valor mínimo igual a 0, se a classificação é perfeita, ou seja, se $\hat{\boldsymbol{p}}_i = \boldsymbol{y}_i$, para todo $i = 1, 2, ..., n$, e valor máximo igual a D se o oposto ocorre, ou seja, se para aquelas observações em que $y_i = k$, temos que $\hat{p}_{ki} = 0$ e para todo $l \neq k$, obtendo que $y_i \neq l$, temos que $\hat{p}_{li} = 1$, para todo $i = 1, 2, ..., n$. Desse modo, grandes valores de BS indicam uma performance ruim com relação à classificação via regressão multinomial, enquanto baixos valores de BS indicam uma boa performance.

Desse modo, podemos utilizar tal escore para comparar o classificador proposto, via regressão multinomial, através de estimativas de $\boldsymbol{p}$, com outros classificadores propostos, e como uma referência podemos comparar com o classificador aleatório, ou seja, supondo equiprobabilidade entre as categorias, isto é, tomando $\hat{p}_{ki} = \frac{1}{D}$, para todo $k = 1, 2, ..., D$ e $i = 1, 2, ..., n$.

Em outras palavras, denote por BS^* sendo o escore de Brier para o classificador aleatório. Temos que

$$\begin{aligned}
BS^* &= \frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{D}\left(\frac{1}{D} - \mathbb{1}(y_i = k)\right)^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}\left[\left(\frac{1}{D} - 1\right)^2 + \sum_{l=1}^{d}\left(\frac{1}{D} - 0\right)^2\right] \\
&= \left(\frac{d}{D}\right)^2 + \frac{d}{D^2} = \frac{d(1+d)}{D^2} = \frac{dD}{D^2} \\
&= \frac{d}{D}.
\end{aligned}$$

Assim, por mais que o valor máximo do escore de Brier seja igual a D, se o valor determinado for maior do que $\frac{d}{D}$, o classificador proposto é pior do que classificar um sujeito lançando um dado honesto de D faces.

Podemos tratar $BS(\delta)$ como uma medida de acurácia, assim como aquelas apresentadas na Seção 3.1: $AC(\delta)$, $SEN_k(\delta)$, $ESP_k(\delta)$, $VPP_k(\delta)$ e $VPN_k(\delta)$, $k = 1, 2, ..., D$.

3.3 Buscando uma bipartição total

Toda a discussão feita nas seções anteriores supõe que fixemos uma bipartição total e desse modo tanto a classificação quanto a regressão são estendíveis do caso binomial ao multinomial. Contudo, vimos no capítulo anterior que há $(2d-1)!!$ maneiras de se propor uma bipartição total. A princípio poderíamos sugerir que ajustemos todos os $(2d-1)!!$ modelos multinomiais possíveis, um para cada bipartição total diferente, obtendo $(2d-1)!!$ classificadores multinomiais e a partir das avaliações de suas performances, propor o melhor deles. Entretanto, o número de modelos a serem ajustados é extremamente grande e se torna infactível computacionalmente. A tabela a seguir mostra o número de modelos a serem ajustados, digamos $a(D)$, de acordo com o número de categorias, D.

Tabela 3.1: *Número de possíveis modelos de acordo com o número de categorias.*

D =	1	2	3	4	5	6	7	8	9	10
$a(D)$ =	1	1	3	15	105	945	10.395	135.135	2.027.025	34.459.425

D =	15	16	17	18	19	20	25	30	40	50
$a(D) \approx$	2×10^{14}	6×10^{15}	10^{17}	10^{18}	10^{20}	10^{21}	10^{30}	10^{38}	10^{47}	10^{76}

Desse modo, podemos utilizar os dados e a partir das estimativas dos riscos binomiais, nível a nível, propor uma bipartição de acordo com a minimização de tais riscos, a fim de se obter um classificador multinomial passo-a-passo. A seguir apresentamos duas alternativas para se buscar uma bipartição total em $\mathscr{L}$ com base nos dados, mas antes discutimos uma alternativa direta quando o problema aplicado nos oferece uma.

3.3.1 Direta

Muitas vezes o próprio problema que se deseja resolver pode sugerir uma bipartição total e, assim, os resultados ficam interpretáveis diretamente de acordo com cada classificação e regressão binomial. Apresentamos dois exemplos em que tal abordagem poderia ser interessante.

Exemplo 3.4 *Considere que desejamos prever a resposta de um indivíduo com relação à sua decisão de viajar nas suas próximas férias. Suponha que*

$$\Omega = \{1, 2, 3, 4\},$$

em que as categorias $1, 2, 3$ *e* 4 *são, respectivamente:*

1. *não viaja;*
2. *viaja dentro do próprio estado;*
3. *viaja fora do próprio estado, mas dentro do próprio país;*
4. *viaja para fora do próprio país.*

Podemos propor, de acordo com as perguntas feitas pelo pesquisador, a seguinte bipartição total:

$$\lambda(\Omega) = \{\Lambda_1, \Lambda_2, \Lambda_3\},$$

em que

$$\begin{aligned}\Lambda_1 &= \{\{1\},\{2,3,4\}\} = \{\text{“não viaja”}, \text{“viaja”}\}\\ \Lambda_2 &= \{\{2,3\},\{4\}\} = \{\text{“viagem nacional”}, \text{“viagem internacional”}\}\\ \Lambda_3 &= \{\{2\},\{3\}\} = \{\text{“viagem nacional dentro do estado”}, \text{“viagem nacional fora do estado”}\}.\end{aligned}$$

Exemplo 3.5 *Suponha que desejamos estudar as características que levam a população da cidade de São Paulo escolher seu principal meio de transporte no dia-a-dia. Considere que*

$$\Omega = \{1,2,3,4,5,6\},$$

representando, respectivamente,

1. *a pé;*
2. *bicicleta;*
3. *ônibus, trem ou metrô;*
4. *táxi e similares ou carona;*
5. *moto;*
6. *carro particular.*

De acordo com as perguntas feitas pelo pesquisador, podemos propor a seguinte bipartição total:

$$\lambda(\Omega) = \{\Lambda_1, \Lambda_2, \Lambda_3, \Lambda_4, \Lambda_5\},$$

em que

$$\begin{aligned}\Lambda_1 &= \{\{1,2\},\{3,4,5,6\}\} = \{\text{“não motorizado”}, \text{“motorizado”}\}\\ \Lambda_2 &= \{\{1\},\{2\}\} = \{\text{“a pé”}, \text{“bicicleta”}\}\\ \Lambda_3 &= \{\{3\},\{4,5,6\}\} = \{\text{“transporte público”}, \text{“transporte privado”}\}\\ \Lambda_4 &= \{\{4\},\{5,6\}\} = \{\text{“transporte privado terceirizado”}, \text{“transporte próprio”}\}\\ \Lambda_5 &= \{\{5\},\{6\}\} = \{\text{“moto”}, \text{“carro”}\}.\end{aligned}$$

3.3.2 *One-versus-one* passo-a-passo

Há, principalmente na literatura de *Statistical Learning* (ver, por exemplo, Bishop, 2011), duas abordagens relativamente simples com relação a extensão de classificadores dicotômicos para politômicos, chamadas de *one-versus-one* (um contra um) e *one-versus-rest* (um contra o resto). Vamos explorar brevemente tais abordagens para motivar duas propostas para se buscar uma bipartição total em $\mathscr{L}$.

One-versus-one Constroem-se todos os possíveis classificadores dicotômicos para uma categoria versus outra categoria, ou seja, sendo $\Omega = \{1,2,...,D\}$, temos $\binom{D}{2}$ possíveis classificadores em $\big\{\{1\},\{2\}\big\}$; $\big\{\{1\},\{3\}\big\}$;...; $\big\{\{1\},\{D\}\big\}$; $\big\{\{2\},\{3\}\big\}$;...; $\big\{\{2\},\{D\}\big\}$;...; $\big\{\{d\},\{D\}\big\}$. A partir disso, para um determinado indivíduo, o classificamos naquela categoria que, dentre as $\binom{D}{2}$ classificações

binárias, teve o maior número de "votos", ou seja, em que o indivíduo foi classificado o maior número de vezes.

One-versus-rest A ideia é muito similar ao *one-versus-one*, em que se constroem todos os possíveis classificadores dicotômicos para uma categoria versus qualquer outra categoria, ou seja, temos D possíveis classificadores em $\big\{\{1\},\{2,3,...,D\}\big\}$; $\big\{\{2\},\{1,3,...,D\}\big\}$;...; $\big\{\{D\},\{1,2,...,d\}\big\}$. Propõe-se o classificador politômico usando o mesmo sistema de votos e classificando o sujeito na categoria que recebeu o maior número de votos.

Exemplo 3.6 *Considere* $D = 4$. *Assim, para o caso one-versus-one temos* 6 *diferentes classificadores, sendo que* 3 *deles envolvem cada uma das categorias de* $\Omega = \{1, 2, 3, 4\}$. *Suponha que dentre as* 6 *classificações, para um determinado indivíduo, obtemos que a categoria* 1 *foi classificada três vezes, enquanto as categorias* $2, 3$ *e* 4 *foram classificadas uma vez cada. Já para o caso one-versus-rest, temos* 4 *diferentes classificadores, sendo que todos envolvem todas as categorias de* Ω. *Suponha que dentre as* 4 *classificações obtemos que a categoria* 1 *recebeu três votos enquanto as categorias* $2, 3$ *e* 4 *apenas um. Portanto, para ambos os casos, o classificador multinomial – para este indivíduo – classifica tal sujeito na categoria* 1.

Contudo, para ambas as abordagens, pode haver casos em que há empates e então regiões em $\mathcal{X}$ em que $\delta(\boldsymbol{x})$ não é função, ou seja, em que há mais do que 1 possível valor para $\delta(\boldsymbol{x})$. Neste sentido, algumas adaptações devem ser feitas para se evitar tais regiões (ver, por exemplo, Friedman, 1996; Hastie e Tibshirani, 1998).

Com relação à modelagem de regressão, essas duas abordagens também falham no seguinte sentido. Para o caso *one-versus-one* (análogo para o caso *one-versus-rest*), se desejamos modelar os $\binom{D}{2}$ parâmetros binomiais $\pi_{lm} = Pr(Y = m|Y \in \{l, m\}; \boldsymbol{p}) = \frac{p_l}{p_l + p_m}$, a fim de obter estimativas $\hat{\pi}_{lm}$ para se obter $\hat{\boldsymbol{p}}$, pode haver mais do que um valor para $\hat{p}_k$.

Nossa intenção nesta subseção não é explorar tais abordagens mas sim propor uma busca de uma bipartição total com base nessas ideias, como segue. Desejamos, desse modo, obter $\lambda(\Omega) = \{\Lambda_1, \Lambda_2, ..., \Lambda_d\}$ de tal modo que $\Lambda_j = \{A_{j1}, A_{j2}\}$. Denominamos de *one-versus-one* passo-a-passo a seguinte proposta de para a construção de λ:

1. Construindo $\Lambda_1 = \{A_{11}, A_{12}\}$, em que $A_{11} \cup A_{12} = \Omega$:

   (a) Ajustamos todos os $\binom{D}{2}$ modelos binomiais *one-versus-one* e escolhemos aquele cujo classificador dicotômico minimiza o risco desejado. Suponha que o modelo escolhido foi em $\big\{\{a\},\{b\}\big\}$. Dessa forma, sem perda de generalidade, fixamos $\{a\} \subset A_{11}$ e $\{b\} \subset A_{12}$. Contudo, $\{A_{11}, A_{12}\}$ não forma uma bipartição de Ω, o que nos leva ao próximo passo.

   (b) A partir das categorias de Ω que não foram selecionadas no passo anterior, digamos $k \in \Omega \backslash \{a, b\}$, ajustamos todos os possíveis modelos binomiais *one-versus-one* da forma $\big\{\{a, k\},\{b\}\big\}$ e $\big\{\{a\},\{b, k\}\big\}$ e escolhemos aquele cujo classificador dicotômico minimiza o risco desejado. Suponha sem perda de generalidade que o modelo escolhido é em $\big\{\{a, c\},\{b\}\big\}$. Fixamos, portanto, $\{a, c\} \subset A_{11}$ e $\{b\} \subset A_{12}$. Possivelmente, ainda, $\{A_{11}, A_{12}\}$ não forma uma bipartição de Ω, o que nos leva ao próximo passo.

   (c) Repetimos o passo anterior até que $\{A_{11}, A_{12}\}$ seja uma bipartição de Ω.

2. Em um próximo nível, repetimos o item 1. para A_{11} e A_{12} (desde que tenham mais do que 1 categoria), construindo então suas respectivas bipartições, digamos, Λ_2 e Λ_3.

3. De forma geral, repetimos este processo recursiva e exaustivamente até obter uma bipartição total $\lambda(\Omega) = \{\Lambda_1, \Lambda_2, ..., \Lambda_d\}$ em que $\Lambda_j = \{A_{j1}, A_{j2}\}$.

Dessa forma, podemos obter qualquer bipartição total em que, para cada bipartição, minimizamos o risco de classificação binomial desejado. Aplicando o que discutimos nas seções anteriores, construímos um modelo de regressão e um classificador multinomial.

3.3.3 *One-versus-rest* passo-a-passo

A proposta anterior pode ter um número grande de passos a fim de se obter um modelo multinomial. Como alternativa, nesta subseção propomos a seguinte construção para λ:

1. Construindo $\Lambda_1 = \{A_{11}, A_{12}\}$, em que $A_{11} \cup A_{12} = \Omega$:

   (a) Ajustamos todos os D modelos binomiais *one-versus-rest* e escolhemos aquele cujo classificador dicotômico minimiza o risco desejado. Suponha, sem perda de generalidade que o modelo escolhido foi em $\big\{\{1\}, \{2, 3, ..., D\}\big\}$. Dessa forma, fixamos $\{1\} \subset A_{11}$ e $\{2, 3, ..., D\} \subset A_{12}$, obtendo uma bipartição de Ω.

2. Em um próximo nível, repetimos o item 1. para A_{12}, construindo então Λ_2.

3. De forma geral, repetimos este processo recursiva e exaustivamente até obter uma bipartição total $\lambda(\Omega) = \{\Lambda_1, \Lambda_2, ..., \Lambda_d\}$ em que $\Lambda_j = \{A_{j1}, A_{j2}\}$.

Dessa forma, obteremos bipartições totais dentro de uma subclasse de $\mathscr{L}$, digamos $\mathscr{L}_d$, em que o número de níveis é igual a d.

Capítulo 4

Aplicação – TOC

Esta aplicação está inserida no contexto médico, mais especificamente em Transtorno Obsessivo-Compulsivo (TOC). O conjunto de dados foi cedido pelo Programa Transtorno do Espectro Obsessivo-Compulsivo (PROTOC), do Instituto de Psiquiatria da Faculdade de Medicina da Universidade de São Paulo (IPq/FMUSP).

Nesta aplicação, utilizamos em todos os casos a função de ligação logística e estimativas de máxima verossimilhança para os parâmetros das regressões e estimativas amostrais para os riscos, como descritas no capítulo anterior.

O TOC é uma doença heterogênea no que diz respeito a manifestação dos sintomas dos pacientes e existem algumas escalas que procuram mensurar a gravidade de tais sintomas para um determinado indivíduo. Dada tal heterogeneidade, um instrumento que tem sido utilizado na literatura de TOC, que faz uma tentativa de extrair escores dimensionais mais específicos, lidando neste sentido com tal heterogeneidade, é o denominado *Dimensional Yale-Brown Obsessive-Compulsive Scale* (DY-BOCS), que agrupa certos sintomas de TOC em cinco dimensões homogêneas mais uma dimensão heterogênea. As cinco dimensões homogêneas são, em suma:

1. Agressão;
2. Sexual ou Religioso;
3. Simetria ou ordem;
4. Limpeza ou contaminação;
5. Colecionismo.

Contudo, um dos instrumentos mais utilizados a fim de se mensurar a gravidade dos sintomas é o denominado *Yale-Brown Obsessive-Compulsive Scale* (Y-BOCS), que não discrimina explicitamente gravidades em dimensões mais homogêneas como o DY-BOCS faz.

Muitos estudos que buscam relacionar o TOC com determinadas características dos pacientes – como por exemplo com genética – falham e uma das razões pode se dar pela sua heterogeneidade, pois grande parte deles utilizam o Y-BOCS para medir gravidade.

Neste sentido, se for possível extrair informações das dimensões do TOC supracitadas através do Y-BOCS, com uma certa qualidade, assim como o DY-BOCS o faz, diversos conjuntos de dados existentes poderiam, a partir dessas informações, buscar relacionar determinados genes com dimensões mais específicas e homogêneas.

Assim, essa aplicação consiste em, a partir de características observadas do instrumento Y-BOCS, prever o resultado do DY-BOCS em cada uma das cinco dimensões. Em outras palavras,

consideramos certas características do Y-BOCS como covariáveis, $\boldsymbol{x}$, e a gravidade do DY-BOCS em uma dimensão específica como a variável resposta, Y, assumindo valores em $\Omega = \{1, 2, 3, 4, 5, 6\}$. Originalmente, o espaço de categorias é dado por $\{0, 1, 2, 3, ..., 15\}$, contudo para fins de aplicação, agregamos tais categorias em seis, sendo $\{0\}$, $\{1, 2, 3\}$, $\{4, 5, 6\}$, $\{7, 8, 9\}$, $\{10, 11, 12\}$ e $\{13, 14, 15\}$, rotuladas por, respectivamente, $1, 2, 3, 4, 5$ e 6.

O conjunto de dados é composto por cinco covariáveis extraídas do Y-BOCS:

x_1: indicadora de algum sintoma da respectiva dimensão do TOC foi destacada como importante pelo paciente (1 se sim e 0 se não) (*target symptom*);

x_2: proporção de sintomas presentes na respectiva dimensão do TOC (dentre todos os sintomas referentes à tal dimensão);

x_3: proporção de sintomas presentes fora da respectiva dimensão (dentre todos os sintomas referentes às outras dimensões);

x_4: escore Y-BOCS de obsessão;

x_5: escore Y-BOCS de compulsão.

As covariáveis x_4 e x_5 foram reescaladas por $x_4/20$ e $x_5/20$, para variarem entre 0 e 1.

Uma característica relevante deste problema é que todas as observações vêm de uma entrevista de um psiquiatra com o respectivo paciente, em que o paciente responde uma série de questões e o psiquiatra atribui o escore de gravidade DY-BOCS para cada uma das cinco dimensões do TOC, baseado em sua experiência profissional, porém subjetiva. Assim, tanto o paciente quanto o psiquiatra, no momento da entrevista, contribuem subjetivamente para um escore de gravidade, tido como razoável no que diz repeito a classificar o paciente de acordo com a sua gravidade no TOC. Isto levanta a questão do que está de fato sendo modelado a fim de ser previsto pelo modelo multinomial construído. Estamos almejando prever, dadas as observações de covariáveis para um novo paciente, qual seria o escore DY-BOCS em Ω que lhe seria dado por um psiquiatra qualquer.

A utilidade clínica desta aplicação é que a partir de bons classificadores, podemos aplicá-los em conjuntos de dados existentes e que contenham somente informações da Y-BOCS e informações a respeito do que está sendo pesquisado, como por exemplo genética, neuropsicologia etc., com o fim de relacionar tais características com as características dimensionais da doença, a partir de um fenótipo mais puro, em um certo sentido, do TOC (todos os detalhes relevantes na apresentação do problema clínico podem ser vistos em Shavitt *et al.* (2017) e nas suas respectivas referências bibliográficas).

Para cada uma das 5 dimensões do TOC, a amostra é composta de 1183 observações independentes de uma variável aleatória supondo distribuição de Bernoulli 6-variada, isto é, $Y|\boldsymbol{x};\boldsymbol{p} \sim BM_6(\boldsymbol{p})$. Para fins de aplicação, consideramos cada dimensão do TOC separadamente como uma aplicação distinta, ou seja, aplicamos o que foi discutido nos capítulos anteriores em cinco variáveis resposta distintas. A Tabela 4.1 mostra a quantidade observada de pacientes em cada categoria de $\Omega = \{1, 2, 3, 4, 5, 6\}$, em cada dimensão.

Tabela 4.1: *Quantidade observada de pacientes para cada dimensão do TOC.*

	1	**2**	**3**	**4**	**5**	**6**
Agressão	415	51	135	229	245	108
Sexual/Religioso	627	57	97	166	173	63
Simetria	295	70	153	232	297	136
Contaminação	414	54	113	200	258	144
Colecionismo	660	111	157	134	92	29

Para cada dimensão, dividimos a amostral total (1183 indivíduos) em amostra de treino e de teste, sendo 70% (828) e 30% (355) em cada, respectivamente, garantindo que no mínimo 3 pacientes estivessem presentes nas duas amostras para cada categoria. A amostra de treino é utilizada para estimar os parâmetros do modelo multinomial (tanto os parâmetros das regressões binomiais, quanto a bipartição total λ e também os pontos de corte nas classificações binomiais). Já a amostra de teste é utilizada para a validação do modelo construído, através da estimação das medidas de acurácia da classificação multinomial e também através do escore de Brier.

4.1 Obtenção dos modelos multinomiais

Nesta seção, exploramos a dimensão de agressão, apresentando a construção da bipartição total discutidas nas Subseções 3.3.2 e 3.3.3 e os resultados com relação aos modelos multinomiais obtidos, tanto com relação às acurácias dos classificadores, quanto com relação às estimativas dos parâmetros do modelo de regressão. Aqui, apresentamos, a partir de uma divisão específica da amostra em treino e teste, a construção de uma bipartição total com base nas minimizações de riscos de classificação binomial.

Para fins de comparação, consideramos ao todo 4 diferentes modelos de regressão e classificações multinomial, fora o logístico com categoria de referência. Sendo as 2 abordagens de busca da bipartição total:

1. *one-versus-one* passo-a-passo (Subseção 3.3.2);

2. *one-versus-rest* passo-a-passo (Subseção 3.3.3),

em que, para cada uma delas, para cada um dos 5 classificadores binomiais, γ_1, γ_2, γ_3, γ_4, e γ_5 (ver Seção 3.1), consideramos dois diferentes riscos a serem minimizados, para todo $j = 1, 2, 3, 4, 5$:

a) $R_1(\gamma_j) = |1 - AC(\gamma_j)|$;

b) $R_2(\gamma_j) = \sqrt{(AC(\gamma_j) - 1)^2 + (SEN(\gamma_j) - 1)^2 + (ESP(\gamma_j) - 1)^2 + (VPP(\gamma_j) - 1)^2 + (VPN(\gamma_j) - 1)^2}$

Por facilidade, na apresentação dos resultados vamos denominar os 4 modelos por $ovoR_1$, $ovrR_1$, $ovoR_2$ e $ovrR_2$, respectivamente, as abordagens combinando os itens acima: 1.a, 2.a, 1.b e 2.b.

4.1.1 Busca da bipartição total

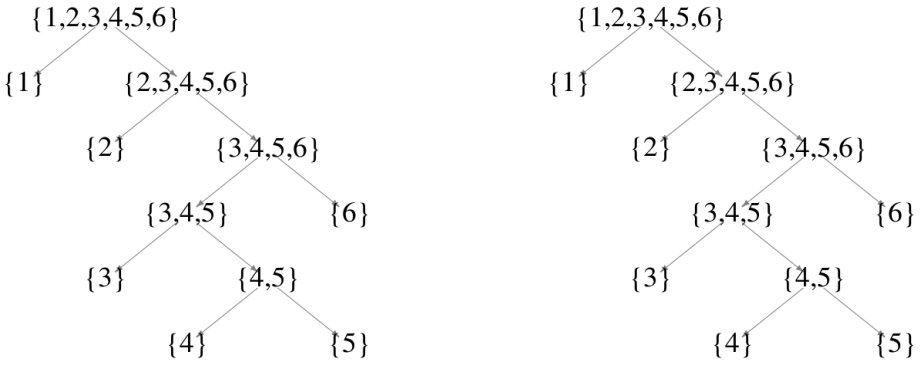
Como $\Omega = \{1, 2, 3, 4, 5, 6\}$, desejamos construir $\lambda(\Omega) = \{\Lambda_1, \Lambda_2, ..., \Lambda_5\}$, em que $\Lambda_j = \{A_{j1}, A_{j2}\}$, $j = 1, 2, 3, 4, 5$. As bipartições totais obtidas a partir das minimizações das estimativas amostrais

para os riscos supracitados, utilizando a amostra de treino, para cada uma das 4 abordagens, são apresentadas na Tabela 4.2. Vale ressaltar que estas são as bipartições totais om base em uma divisão específica da amostra em treino e teste.

Tabela 4.2: *Bipartições totais para cada uma das* 4 *abordagens.*

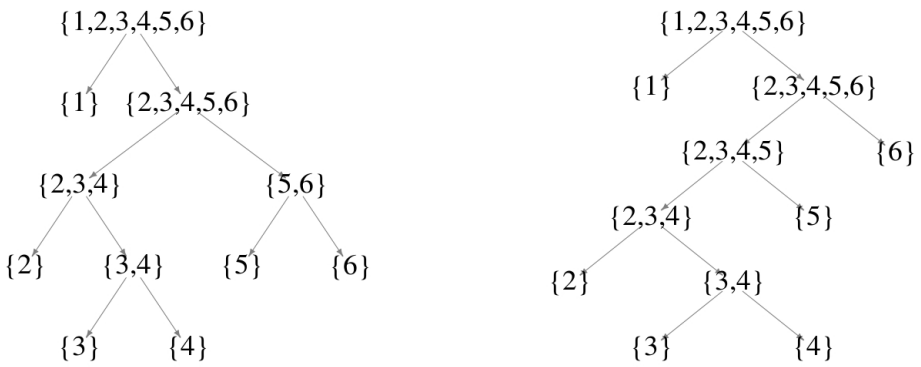
	Λ_1	Λ_2	Λ_3	Λ_4	Λ_5
$ovoR_1$	$\{\{1\},\{2,3,4,5,6\}\}$	$\{\{2\},\{3,4,5,6\}\}$	$\{\{3,4,5\},\{6\}\}$	$\{\{3,4\},\{5\}\}$	$\{\{4\},\{5\}\}$
$ovrR_1$	$\{\{1\},\{2,3,4,5,6\}\}$	$\{\{2\},\{3,4,5,6\}\}$	$\{\{3,4,5\},\{6\}\}$	$\{\{3,4\},\{5\}\}$	$\{\{4\},\{5\}\}$
$ovoR_2$	$\{\{1\},\{2,3,4,5,6\}\}$	$\{\{2,3,4\},\{5,6\}\}$	$\{\{2\},\{3,4\}\}$	$\{\{5\},\{6\}\}$	$\{\{3\},\{4\}\}$
$ovrR_2$	$\{\{1\},\{2,3,4,5,6\}\}$	$\{\{2,3,4,5\},\{6\}\}$	$\{\{2,3,4\},\{5\}\}$	$\{\{2\},\{3,4\}\}$	$\{\{3\},\{4\}\}$

Note que a bipartição total resultante em $ovoR_1$ é exatamente a mesma em $ovrR_1$. Por outro lado, utilizando o risco R_2, as bipartições totais obtidas em $ovoR_2$ e $ovrR_2$ são distintas entre si, porém compartilham de duas bipartições idênticas, sendo: a bipartição Λ_3 em $ovoR_2$ representada por Λ_4 em $ovrR_2$; bem como a bipartição representada por Λ_5 em ambos os casos. É interessante notar também que para todos os casos Λ_1 foi o mesmo, ou seja, aquele em que $W_1 = 1$ se $Y \in \{1\}$ e $W_1 = 0$ se $Y \in \{2,3,4,5,6\}$, isto é, a classificação em "doente" e "não-doente" com relação a dimensão de agressão. Essas diferenças e similaridades ficam explícitas nas representações gráficas nas Figuras 4.1 e 4.2.



(a) $ovoR_1$ (b) $ovrR_1$

Figura 4.1: *Representações das bipartições totais para* $ovoR_1$ *e* $ovrR_1$.



(a) $ovoR_2$ (b) $ovrR_2$

Figura 4.2: *Representações das bipartições totais para $ovoR_2$ e $ovrR_2$.*

4.1.2 Comparação entre as abordagens

Os riscos R_1 e R_2 são para os casos binomiais e podemos estimá-los como discutimos na Seção 3.1. Apresentamos tais estimativas ainda na amostra de treino na Tabela 4.3. Note, contudo, que R_1 e R_2 não são comparáveis e apresentamos as estimativas apenas com fins descritivos, para entendermos em quais casos – dentro de cada abordagem – temos maiores estimativas ou, em outras palavras, maiores erros de classificação. Vemos, por exemplo, que para Λ_1 os erros são pequenos em todos os sentidos que R_1 e R_2 trazem, com relação à proporção de erros e à sensibilidade, especificidade, valores preditivos positivo e negativo. Por outro lado, para Λ_5 – em todos os casos – os erros são relativamente maiores, evidenciando a maior dificuldade de se classificar em categorias mais próximas, nos casos do risco R_1 para $\{4\} \times \{5\}$, e nos casos do risco R_2 para $\{3\} \times \{4\}$.

Tabela 4.3: *Estimativas amostrais para os riscos binomiais.*

	Λ_1	Λ_2	Λ_3	Λ_4	Λ_5
$ovoR_1$	$0,027$	$0,068$	$0,126$	$0,212$	$0,321$
$ovrR_1$	$0,027$	$0,068$	$0,126$	$0,212$	$0,321$
$ovoR_2$	$0,074$	$0,661$	$0,770$	$0,756$	$0,888$
$ovrR_2$	$0,074$	$0,776$	$0,731$	$0,770$	$0,888$

Entretanto, a fim de comparar as abordagens entre si e, também com a logito usual, devemos estimar as medidas de acurácia multinomiais para comparação entre as 4 abordagens (mais a "usual"). Para tal, podemos construir as tabelas cruzadas de classificação, na amostra de teste, comparando os observados e os preditos para todos os casos, como apresentado na Tabela 4.4.

Podemos ver que as abordagens com risco R_1, bem como a usual, negligenciam as categorias menos frequentes, $\{2\}, \{3\}$ e $\{6\}$, classificando a maior parte dos sujeitos em $\{1\}, \{4\}$ e $\{5\}$. Por outro lado, as abordagens com risco R_2 são mais equilibradas neste sentido, classificando os indivíduos em todas as categorias de modo relativamente proporcional ao observado.

Estas similaridades e diferenças ficam mais evidentes ao apresentarmos as estimativas das medidas de acurácia. Na Tabela 4.5 apresentamos a proporção de acertos, sensibilidades, valores preditivos positivos e também o escore de Brier, denotados, respectivamente, por $\widehat{AC}$, $\widehat{SEN}_k$, $\widehat{VPP}_k$, $k = 1, 2, 3, 4, 5, 6$, e BS. Omitimos as especificidades e valores preditivos negativos pois são todos

Tabela 4.4: *Tabelas cruzadas de classificação para $ovoR_1$, $ovrR_1$, $ovoR_2$, $ovrR_2$ e usual.*

$ovoR_1$

Preditos \ Observados	1	2	3	4	5	6	Total
1	104	0	0	0	0	0	104
2	0	0	0	0	0	0	0
3	2	3	1	0	0	0	6
4	6	9	29	47	16	7	114
5	3	3	11	33	48	25	123
6	0	0	1	1	2	4	8
Total	115	15	42	81	66	36	355

$ovrR_1$

Preditos \ Observados	1	2	3	4	5	6	Total
1	104	0	0	0	0	0	104
2	0	0	0	0	0	0	0
3	2	3	1	0	0	0	6
4	6	9	29	47	16	7	114
5	3	3	11	33	48	25	123
6	0	0	1	1	2	4	8
Total	115	15	42	81	66	36	355

$ovoR_2$

Preditos \ Observados	1	2	3	4	5	6	Total
1	105	2	1	0	0	0	108
2	3	4	6	7	2	1	23
3	3	5	10	10	3	2	33
4	1	1	13	29	11	2	57
5	1	1	9	32	36	14	93
6	2	2	3	3	14	17	41
Total	115	15	42	81	66	36	355

$ovrR_2$

Preditos \ Observados	1	2	3	4	5	6	Total
1	105	2	1	0	0	0	108
2	2	4	6	8	2	1	23
3	3	5	10	10	4	2	34
4	1	1	15	32	15	5	69
5	1	1	7	28	32	13	82
6	3	2	3	3	13	15	39
Total	115	15	42	81	66	36	355

usual

Preditos \ Observados	1	2	3	4	5	6	Total
1	109	2	3	5	1	0	120
2	0	0	0	0	0	0	0
3	1	2	0	2	0	0	5
4	3	9	21	26	10	3	72
5	2	1	17	46	52	27	145
6	0	1	1	2	3	6	13
Total	115	15	42	81	66	36	355

relativamente altos e similares entre si. Note que não há uma estimativa para o valor preditivo positivo para a categoria 2 nos casos $ovoR_1$ e $ovrR_1$, uma vez que nenhum indivíduo foi classificado em tal categoria.

Tabela 4.5: *Proporção de acertos, sensibilidades, valores preditivos positivos e escore de Brier.*

	$ovoR_1$	$ovrR_1$	$ovoR_2$	$ovrR_2$	*usual*
proporção de acertos	0,575	0,575	0,566	0,558	0,544
sensibilidades					
1	0,904	0,904	0,913	0,913	0,948
2	0	0	0,267	0,267	0
3	0,024	0,024	0,238	0,238	0
4	0,58	0,58	0,358	0,395	0,321
5	0,727	0,727	0,545	0,485	0,788
6	0,111	0,111	0,472	0,417	0,167
valores preditivos positivos					
1	1	1	0,972	0,972	0,908
2	–	–	0,174	0,174	–
3	0,167	0,167	0,303	0,294	0
4	0,412	0,412	0,509	0,464	0,361
5	0,39	0,39	0,387	0,39	0,359
6	0,5	0,5	0,415	0,385	0,462
escore de Brier	0,507	0,507	0,508	0,507	0,546

Desse modo, vemos que as sensibilidades para as categorias menos frequentes $\{2\}$, $\{3\}$ e $\{6\}$ são maiores nas abordagens com risco R_2 do que no restante. Com relação aos valores preditivos positivos, essas diferenças são menos discrepantes. Com relação ao escore de Brier, valores próximos entre si são observados para as 4 abordagens propostas, sendo que a usual apresentou um valor mais elevado, evidenciando, para este caso, que as estimativas do vetor de parâmetros $\boldsymbol{p}$ para cada indivíduo, em média, foram mais distantes às observações y da variável resposta.

4.1.3 Estimativas dos parâmetros da regressão

Com relação ao modelo de regressão, apresentamos os resultados das estimativas pontuais (por máxima verossimilhança), juntamente com as estimativas do respectivo erro padrão e intervalos de confiança de 95%, analogamente ao caso usual, nas Tabelas 4.6, 4.7, 4.8 e 4.9, para, respectivamente, $ovoR_1$ (idêntico para $ovrR_1$), $ovoR_2$, $ovrR_2$ e para o usual tomando como categoria de referência a categoria $\{6\}$.

O modelo binomial em Λ_1, para qualquer uma das abordagens propostas, é com relação a $W_1 = 1$ se $Y = 1$ e W_0 se $Y \neq 1$. Para este caso, vemos que x_2 apresenta uma estimativa alta, ou seja, quanto maior a proporção de sintomas de agressão o indivíduo tem, maior a probabilidade de $W_1 = 0$, isto é, $Y \neq 1$ (ter gravidade do TOC positiva para agressão através do DY-BOCS).

Com relação à bipartição Λ_2, para os casos $ovoR_1$ e $ovrR_1$, x_2 apresenta estimativa bem menor do que em Λ_1, mas ainda parece ser importante para diferenciar um sujeito em $\{2\}$ *versus* $\{3, 4, 5, 6\}$, juntamente com x_5, ou seja, quanto maior o escore Y-BOCS de compulsão, maior a probabilidade de tal paciente ter escore DY-BOCS de agressão em $\{3, 4, 5, 6\}$. Também é interessante notar as estimativas com relação à bipartição Λ_3, ainda para os casos $ovoR_1$ e $ovrR_1$, em que a variável mais importante nessa relação parece ser x_4, ou seja, quanto maior o escore Y-BOCS de obsessão,

maior a probabilidade de o indivíduo ter uma gravidade DY-BOCS de agressão em $\{6\}$ do que em $\{3, 4, 5\}$.

As interpretações com relação às Tabelas 4.7, 4.8 são análogas.

Com relação à Tabela 4.9, ela apresenta os resultados como se faz usualmente e cada estimativa, por mais que seja obtida através da verossimilhança do modelo multinomial, diz respeito somente ao par de categorias $\{j\}$ e $\{D\}$, fixando-se, por exemplo, a categoria $\{D\}$ como a de referência.

Durante toda esta seção apresentou resultados a partir de uma divisão específica da amostra em treino e teste, de forma aleatória. Para fazer um estudo de performance entre os métodos, de fato, devemos fazer esta divisão um número grande de vezes, como exploramos na seção seguinte.

Tabela 4.6: *Estimativas para os parâmetros da regressão – $ovoR_1$ e $ovrR_1$.*

Bipartição/Variável	estimativa pontual	erro padrão	intervalo de confiança de 95%
$\Lambda_1 : \{1\} \times \{2, 3, 4, 5, 6\}$			
Intercepto	3,856	0,631	[2,672 ; 5,155]
x_1	-1,554	0,434	[-2,431 ; -0,724]
x_2	-35,613	3,187	[-42,324 ; -29,799]
x_3	1,169	1,091	[-0,963 ; 3,328]
x_4	-1,15	1,208	[-3,573 ; 1,126]
x_5	-0,875	1,175	[-3,163 ; 1,411]
$\Lambda_2 : \{2\} \times \{3, 4, 5, 6\}$			
Intercepto	0,582	0,695	[-0,788 ; 1,95]
x_1	-0,495	0,395	[-1,299 ; 0,26]
x_2	-2,625	1,139	[-4,976 ; -0,488]
x_3	3,464	1,179	[1,197 ; 5,837]
x_4	-1,103	1,428	[-3,924 ; 1,686]
x_5	-4,465	1,254	[-6,947 ; -2,004]
$\Lambda_3 : \{3, 4, 5\} \times \{6\}$			
Intercepto	8,26	1,004	[6,402 ; 10,347]
x_1	0,179	0,287	[-0,381 ; 0,748]
x_2	-0,865	0,701	[-2,239 ; 0,521]
x_3	1,577	0,791	[0,041 ; 3,149]
x_4	-9,839	1,644	[-13,211 ; -6,73]
x_5	1,117	1,21	[-1,281 ; 3,507]
$\Lambda_4 : \{3, 4\} \times \{5\}$			
Intercepto	1,413	0,557	[0,332 ; 2,524]
x_1	-0,557	0,256	[-1,068 ; -0,06]
x_2	-2,311	0,756	[-3,847 ; -0,874]
x_3	1,325	0,74	[-0,117 ; 2,79]
x_4	-1,114	1,091	[-3,25 ; 1,043]
x_5	-2,186	0,979	[-4,123 ; -0,271]
$\Lambda_5 : \{3\} \times \{4\}$			
Intercepto	2,899	0,605	[1,745 ; 4,123]
x_1	-0,49	0,239	[-0,963 ; -0,023]
x_2	-1,844	0,661	[-3,17 ; -0,573]
x_3	0,993	0,679	[-0,33 ; 2,34]
x_4	-3,332	1,161	[-5,664 ; -1,093]
x_5	-0,334	1,048	[-2,391 ; 1,734]

Tabela 4.7: *Estimativas para os parâmetros da regressão – ovoR_2.*

Variável	estimativa pontual	erro padrão	intervalo de confiança de 95%
$\Lambda_1 : \{1\} \times \{2,3,4,5,6\}$	$= \Lambda_1$ na Tabela 4.6		
$\Lambda_2 : \{2,3,4\} \times \{5,6\}$			
Intercepto	4,539	0,513	[3,566 ; 5,578]
x_1	-0,485	0,199	[-0,876 ; -0,097]
x_2	-2,368	0,539	[-3,445 ; -1,33]
x_3	1,547	0,563	[0,454 ; 2,664]
x_4	-4,521	0,933	[-6,398 ; -2,729]
x_5	-1,121	0,811	[-2,72 ; 0,469]
$\Lambda_3 : \{2\} \times \{3,4\}$			
Intercepto	-0,024	0,764	[-1,538 ; 1,473]
x_1	-0,329	0,404	[-1,151 ; 0,447]
x_2	-1,876	1,189	[-4,324 ; 0,366]
x_3	3,036	1,245	[0,63 ; 5,538]
x_4	0,431	1,639	[-2,741 ; 3,706]
x_5	-4,368	1,39	[-7,19 ; -1,703]
$\Lambda_4 : \{5\} \times \{6\}$			
Intercepto	5,836	1,101	[3,791 ; 8,118]
x_1	0,368	0,307	[-0,233 ; 0,976]
x_2	0,033	0,743	[-1,419 ; 1,507]
x_3	0,981	0,804	[-0,587 ; 2,578]
x_4	-8,102	1,792	[-11,812 ; -4,748]
x_5	1,318	1,347	[-1,303 ; 4,021]
$\Lambda_5 : \{3\} \times \{4\}$			
Intercepto	0,817	0,62	[-0,387 ; 2,054]
x_1	-0,338	0,282	[-0,897 ; 0,212]
x_2	-1,441	0,851	[-3,158 ; 0,191]
x_3	1,004	0,859	[-0,674 ; 2,706]
x_4	0,161	1,208	[-2,198 ; 2,565]
x_5	-1,831	1,054	[-3,942 ; 0,211]

Tabela 4.8: *Estimativas para os parâmetros da regressão – ovrR_2.*

Variável	estimativa pontual	erro padrão	intervalo de confiança de 95%
$\Lambda_1 : \{1\} \times \{2,3,4,5,6\}$	$= \Lambda_1$ na Tabela 4.6		
$\Lambda_2 : \{2,3,4,5\} \times \{6\}$			
Intercepto	8,458	1,001	[6,609 ; 10,54]
x_1	0,162	0,286	[-0,396 ; 0,729]
x_2	-0,966	0,7	[-2,336 ; 0,416]
x_3	1,652	0,789	[0,119 ; 3,221]
x_4	-9,738	1,62	[-13,05 ; -6,663]
x_5	0,843	1,194	[-1,535 ; 3,189]
$\Lambda_3 : \{2,3,4\} \times \{5\}$			
Intercepto	4,06	0,531	[3,052 ; 5,139]
x_1	-0,585	0,211	[-1,001 ; -0,172]
x_2	-2,354	0,569	[-3,491 ; -1,257]
x_3	1,395	0,613	[0,205 ; 2,611]
x_4	-2,873	1,015	[-4,902 ; -0,911]
x_5	-1,578	0,899	[-3,364 ; 0,171]
$\Lambda_4 : \{2\} \times \{3,4\}$	$= \Lambda_3$ na Tabela 4.7		
$\Lambda_5 : \{3\} \times \{4\}$	$= \Lambda_5$ na Tabela 4.7		

Tabela 4.9: *Estimativas para os parâmetros da regressão – usual (categoria de referência 6).*

Variável	estimativa pontual	erro padrão	intervalo de confiança de 95%
$\{1\} \times \{6\}$			
Intercepto	9,406	1,853	[6,216 ; 13,571]
x_1	-1,517	0,765	[-3,03 ; 0,005]
x_2	-19,363	3,009	[-26,037 ; -14,128]
x_3	1,059	1,851	[-2,44 ; 4,908]
x_4	-8,031	2,628	[-13,341 ; -2,891]
x_5	0,267	1,877	[-3,948 ; 3,599]
$\{2\} \times \{6\}$			
Intercepto	6,815	1,606	[4,013 ; 10,385]
x_1	-0,163	0,645	[-1,46 ; 1,113]
x_2	-2,94	1,581	[-6,307 ; -0,005]
x_3	3,251	1,693	[0,101 ; 6,828]
x_4	-7,798	2,635	[-13,647 ; -3,064]
x_5	-2,882	2,196	[-7,137 ; 1,69]
$\{3\} \times \{6\}$			
Intercepto	9,009	1,443	[6,428 ; 12,125]
x_1	-0,001	0,438	[-0,861 ; 0,868]
x_2	-3,306	1,106	[-5,58 ; -1,215]
x_3	1,89	1,228	[-0,467 ; 4,382]
x_4	-10,755	2,288	[-15,679 ; -6,64]
x_5	-0,345	1,682	[-3,585 ; 3,049]
$\{4\} \times \{6\}$			
Intercepto	8,463	1,241	[6,214 ; 11,106]
x_1	-0,061	0,363	[-0,773 ; 0,657]
x_2	-1,754	0,99	[-3,744 ; 0,156]
x_3	1,627	0,999	[-0,302 ; 3,633]
x_4	-10,81	1,952	[-14,926 ; -7,231]
x_5	0,755	1,48	[-2,076 ; 3,762]
$\{5\} \times \{6\}$			
Intercepto	5,836	1,101	[3,791 ; 8,118]
x_1	0,368	0,307	[-0,233 ; 0,976]
x_2	0,033	0,743	[-1,419 ; 1,507]
x_3	0,981	0,804	[-0,587 ; 2,578]
x_4	-8,102	1,792	[-11,812 ; -4,748]
x_5	1,318	1,347	[-1,303 ; 4,021]

4.2 Performance dos classificadores multinomiais

Na seção anterior apresentamos os resultados para uma divisão da amostra em treino e teste, obtendo portanto uma única proposta de classificador multinomial, para cada uma das abordagens, bem como um único modelo de regressão multinomial.

A fim de estudar a performance de tais classificadores e regressões, sem ser possivelmente favorecido por uma amostra de treino e teste específicas, nesta seção repetimos o processo da divisão da amostra $B = 250$ vezes e assim obtemos diferentes estimativas das medidas de acurácia e também do escore de Brier, nos possibilitando apresentar como os resultados variam de acordo com as abordagens.

Apresentamos as estimativas para a acurácia, sensibilidades, valores preditivos positivos e os escores de Brier (denotadas, respectivamente, por $\widehat{AC}$, $\widehat{SEN}_k$ e $\widehat{VPP}_k$, $k = 1, 2, 3, 4, 5, 6$, e BS) através de gráficos de boxplot, a fim de comparar as 4 abordagens propostas e também o logito usual. Desse modo, nesta seção, apresentamos o resultado para todas as dimensões do TOC, consideradas portanto 5 diferentes aplicações.

Apesar de não termos abordado especificamente o caso ordinal durante este trabalho, esta aplicação traz este caráter naturalmente e, na seção anterior, podemos ver como as propostas, sem uma imposição com relação à ordem na variável resposta, estruturam as bipartições totais de modo a agrupar as categorias mais similares de acordo com as covariáveis. Assim sendo, nesta seção, exploramos uma medida de acurácia que atribui pesos maiores para erros considerados mais graves, isto é, para classificações mais distantes das respectivas observações. Definimos, e denotamos por $\widehat{EP}$, o erro médio ponderado, dado por

$$\widehat{EP} = \frac{1}{15} \sum_{m=0}^{5} \#\{|y_i - \delta(\boldsymbol{x}_i)| = m\} \times m,$$

isto é, atribuímos um peso igual a m se a distância absoluta entre o predito e o observado é igual a m, em que $m = 0, 1, 2, 3, 4, 5$.

As Figuras 4.3, 4.4, 4.5, 4.6 e 4.7 apresentam os boxplots, respectivamente, para as dimensões do TOC de agressão, sexual/religioso, simetria, contaminação e colecionismo. Cada uma das figuras apresenta os boxplots, respectivamente, para $\widehat{AC}$, $\widehat{EP}$, BS, $\widehat{SEN}_k$ e $\widehat{VPP}_k$, $k = 1, 2, 3, 4, 5, 6$.

Com relação às proporções de acerto, as abordagens baseadas no risco R_2 são, em geral, menores do que as outras, uma vez que buscamos maximizá-las, mas levando em conta também as outras medidas de acurácia. A logito usual e as abordagens baseadas no risco R_1 são parecidas neste aspecto, a não ser em agressão que o logito usual ficou próximo à $ovoR_2$ e $ovrR_2$.

Já para as sensibilidades vemos uma grande diferença – assim como tínhamos notado na seção anterior – para as categorias menos frequentes, {2} {3} e {6}. Os métodos baseados no risco R_2, no geral, são mais equilibrados e apresentam maiores sensibilidades para tais categorias, por mais que apresentem menores valores para as categorias {4} e {5} comparando com as outras 3 abordagens. Para a categoria {1}, em todas as dimensões, as sensibilidades para todas as abordagens foram relativamente altas. Para os valores preditivos positivos, todas as abordagens foram similares.

Para o erro ponderado, as abordagens baseadas no risco R_1, no geral, apresentaram melhor desempenho. Já com relação ao escore de Brier, o logito usual apresentou pior performance, a não ser nas dimensões sexual/religioso e colecionismo em que foi similar às outras 4 abordagens.

4.2.1 Agressão

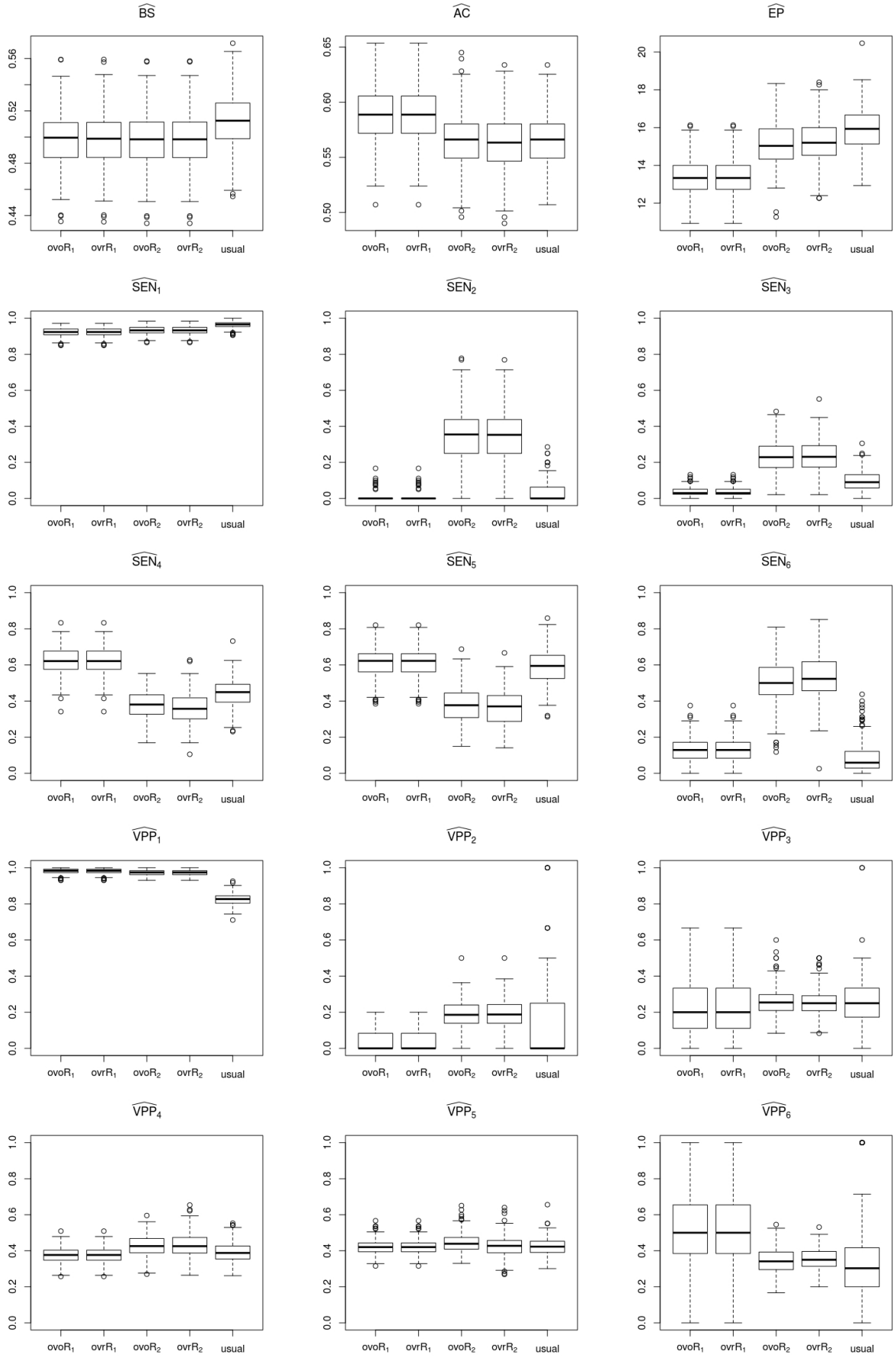


Figura 4.3: *Medidas de acurácia para a dimensão de agressão.*

4.2.2 Sexual/Religioso

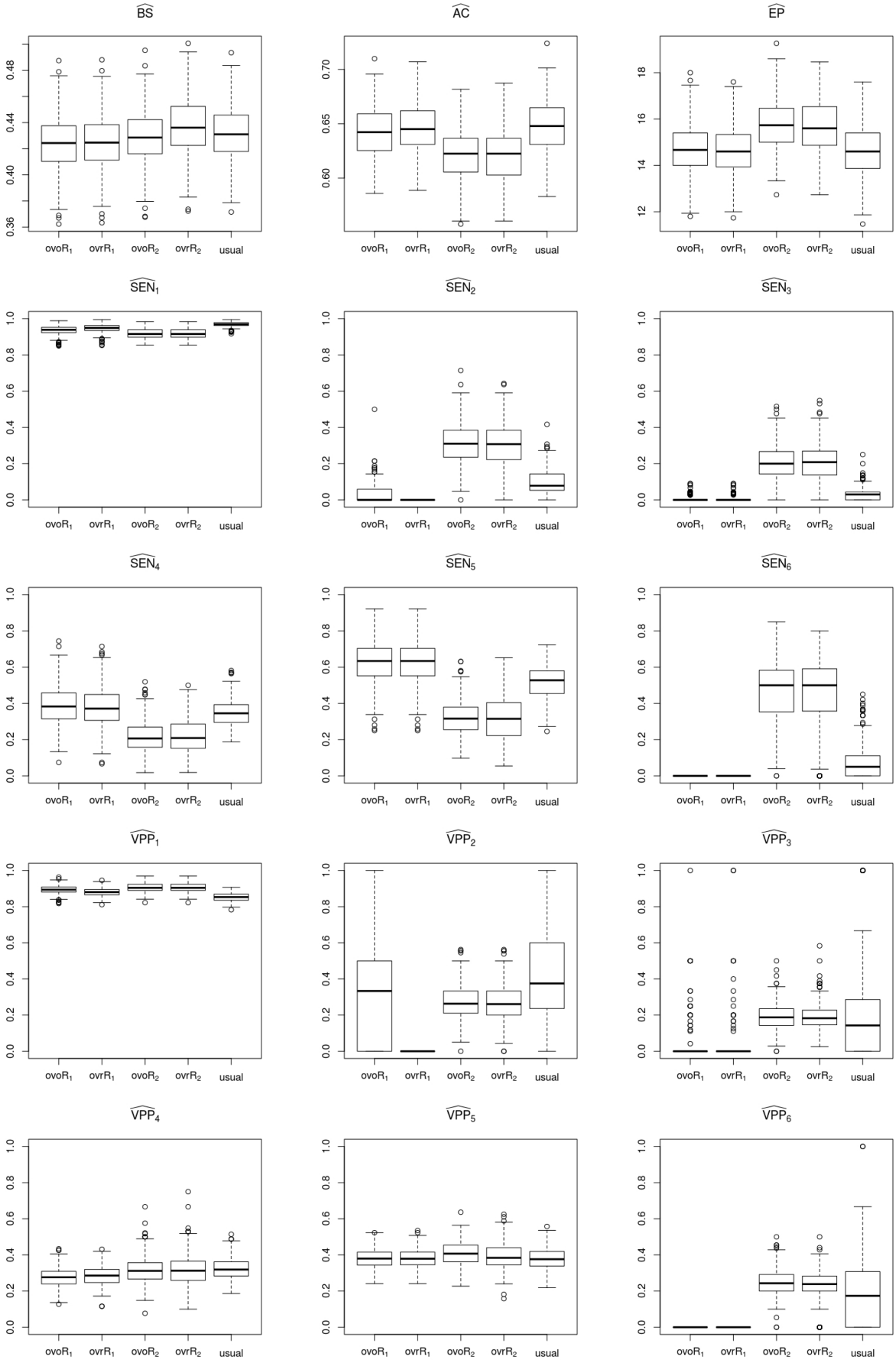


Figura 4.4: *Medidas de acurácia para a dimensão sexual/religioso.*

4.2.3 Simetria

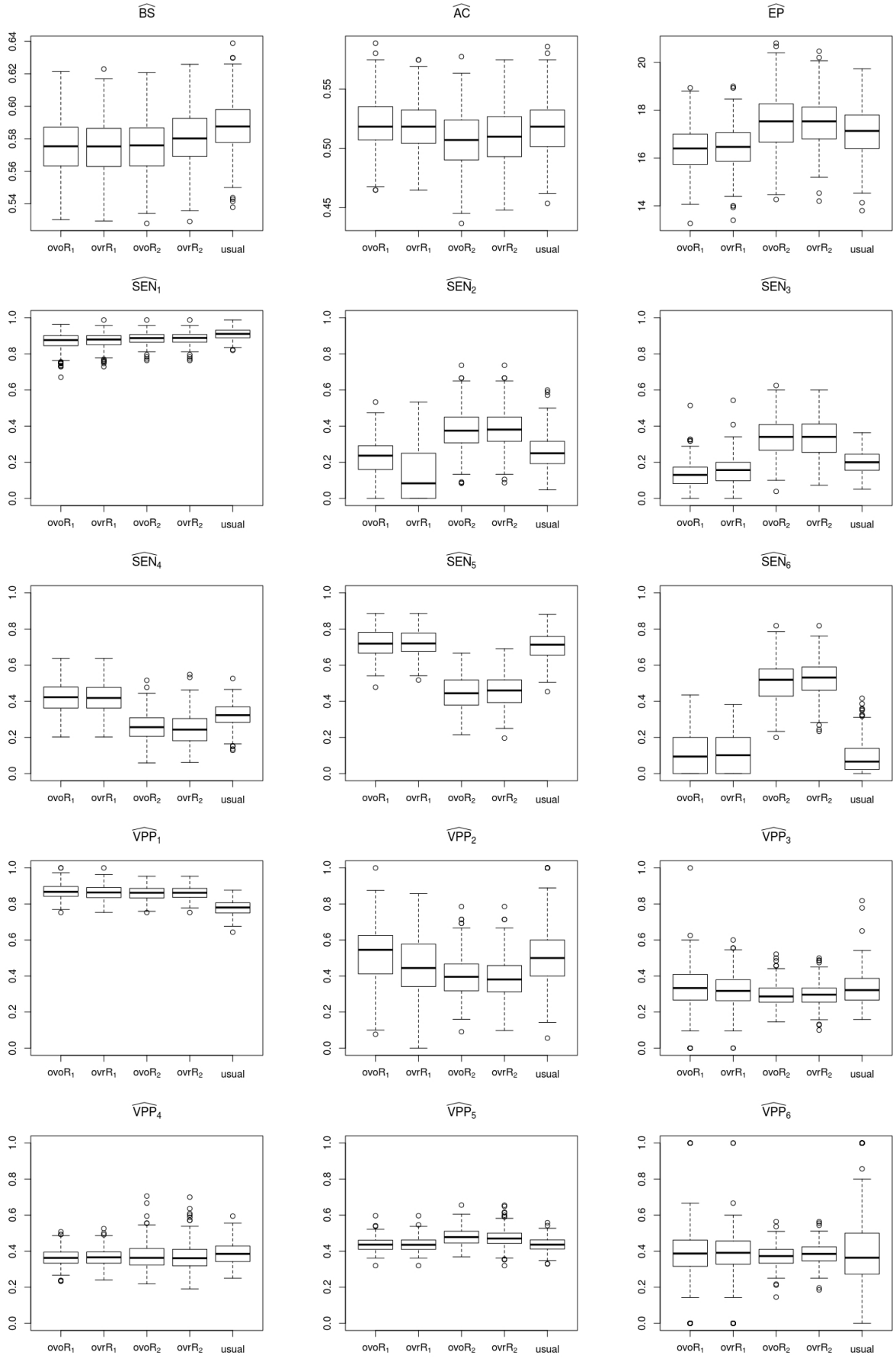


Figura 4.5: *Medidas de acurácia para a dimensão de simetria.*

4.2.4 Contaminação

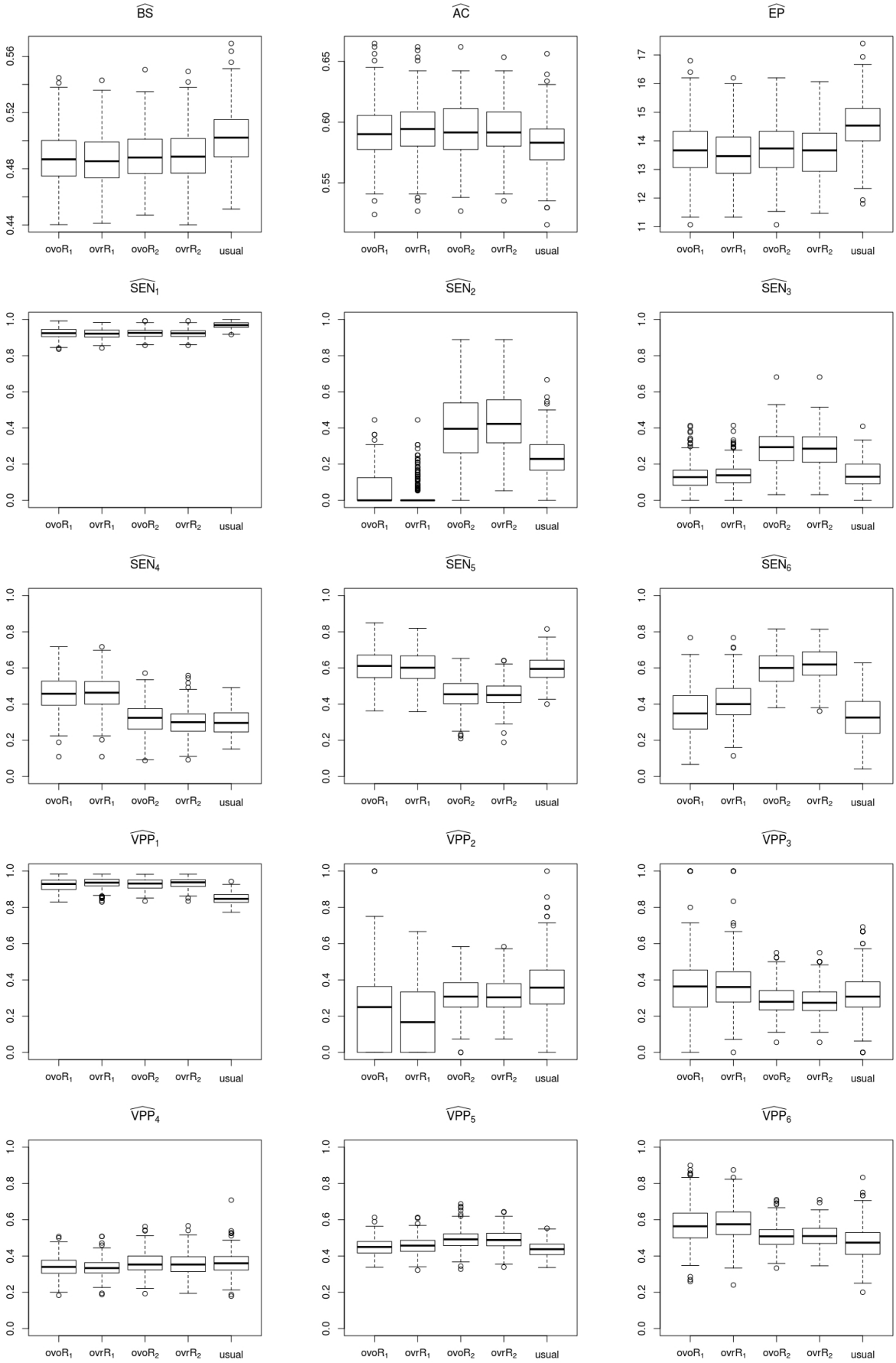


Figura 4.6: *Medidas de acurácia para a dimensão de contaminação.*

4.2.5 Colecionismo

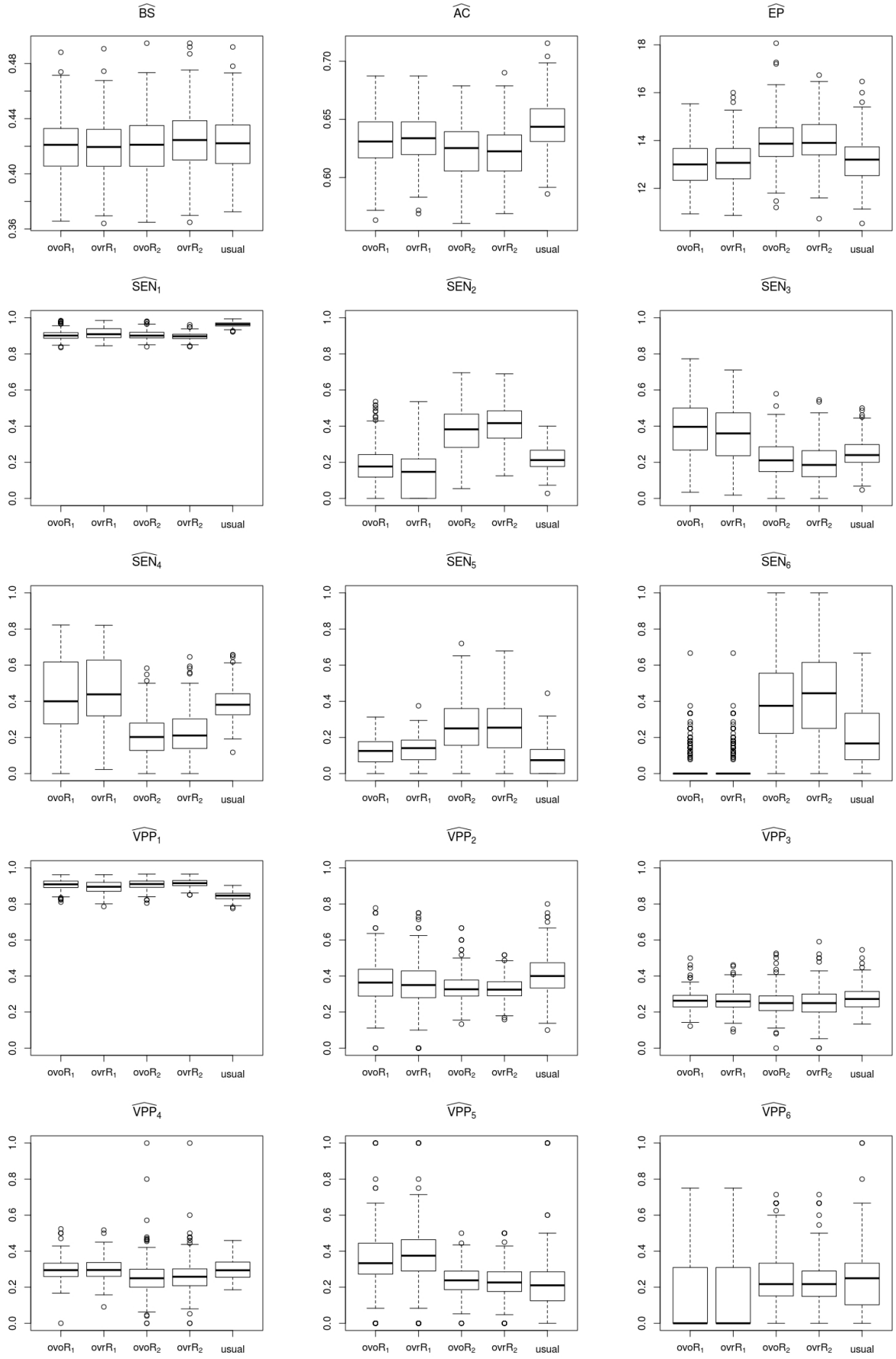


Figura 4.7: *Medidas de acurácia para a dimensão de colecionismo.*

Algumas considerações desta aplicação

Com base no estudo de performance na seção anterior, na prática, a aplicação dos métodos em TOC é bastante útil, uma vez que permite com que determinados conjuntos de dados com informações de genética e de TOC por meio do YBOCS, por exemplo, sejam utilizados, controlando-se determinados erros de classificação binomial. Na área de médica, é relevante a aplicação de modelos induzidos a partir da construção de uma bipartição total em TOC, como publicado em Shavitt *et al.* (2017).

Capítulo 5

Considerações finais

Neste trabalho, exploramos como as fatorações da distribuição de Bernoulli multivariada em Bernoulli's levam a construções de modelos de regressão alternativos ao logístico com categoria de referência, usual e amplamente utilizado nas mais diversas áreas do conhecimento, a partir de modelos de regressão binomiais. Também, de forma independente dos modelos de regressão, exploramos como tais fatorações levam a construções de classificadores politômicos com base na construção de dicotômicos.

Apesar de termos como ponto de partida motivador, de um ponto de vista estatístico, os modelos de regressão multinomial, demos um passo no sentido oposto e exploramos no Capítulo 2 como um conjunto finito qualquer pode ser representado equivalentemente por uma classe de conjuntos com dois elementos, a qual denominamos por bipartição total. Naturalmente, pela natureza finita do espaço amostral para a distribuição de Bernoulli multivariada, vimos, ainda no Capítulo 2, transformações de uma variável aleatória com distribuição Bernoulli D-variada que são induzidas pela bipartição total e também como a reparametrização é estruturada em um mesmo sentido através da bipartição total.

Com o problema politômico transformado – com relação ao espaço amostral e paramétrico – em problemas dicotômicos, a aplicação dos conceitos vistos no Capítulo 2 em modelos de regressão e classificação multinomial foi explorada no Capítulo 3. Uma questão computacional surge neste capítulo pela grande quantidade de possíveis bipartições totais e, consequentemente, de modelos de regressão e classificação multinomial. Disso, propomos duas diferentes abordagens para buscar uma bipartição total e, consequentemente, uma estrutura para a regressão e classificação.

Para cada uma dessas propostas – e de uma forma geral para qualquer bipartição total – vimos também no Capítulo 3 que podemos controlar medidas de acurácia no problema multinomial – como sensibilidades e especificidades, bem como valores preditivos positivos e negativos, dentre diversos outros riscos que podem ser definidos – através da minimização desses riscos nos problemas binomiais.

Por fim, no Capítulo 4, estudamos as performances das propostas e o comportamento e apresentação de resultados para o modelo de regressão e classificação multinomial através das discussões feitas nos capítulos anteriores. Evidentemente, essa construção é aplicável a qualquer problema de regressão e classificação multinomial.

Este trabalho abre caminho para diversos trabalhos futuros: i) Como vimos no Capítulo 3 ao estudarmos o modelo de regressão multinomial logístico usual, que esta abordagem também é proposta como fruto da decomposição em problemas binomiais, uma classe bastante geral de regressões

multinomiais pode ser proposta combinando esta e a decomposição explorada no decorrer deste trabalho. Podemos, por exemplo, construir R-partições totais e, para cada problema R-variado, utilizar um modelo multinomial como usualmente feito; ii) Toda a flexibilização e generalização com relação às funções de ligação que existem na literatura sobre modelos de regressão binomial, que em geral não são discutidos no caso multinomial, pode ser incorporada diretamente na nossa construção, bem como na construção usual na qual mostramos o papel da função de distribuição logística padrão. Pode-se, inclusive, utilizar funções de ligação distintas para cada modelo binomial e também covariáveis distintas. É de se esperar que através de tais flexibilizações obtenhamos resultados mais acurados. iii) Um algoritmo computacional – mais geral do que os apresentados nesta tese – de busca por uma, ou uma mistura de, bipartições totais, pode ser construído. Propomos duas abordagens empíricas passo-a-passo nessa construção, mas elas são aditivas em um certo sentido: uma vez que selecionamos uma categoria para um ou outro conjunto de uma bipartição, a deixamos fixa e um algoritmo computacional pode generalizar essa ideia, buscando uma bipartição que leve a resultados mais precisos na classificação. Diversos outros métodos de classificação dicotômica podem ser aplicados diretamente nessa construção. iv) O tamanho amostral em cada categoria do problema pode ser um elemento importante na busca pelas bipartições que minimizam os riscos binomiais, na construção da bipartição total. Medidas como AIC e BIC não foram explorados neste trabalho justamente por estarmos comparando modelos binomiais em espaços amostrais diferentes, tendo consequentemente tamanhos amostrais distintos. Logo, os tamanhos amostrais em cada categoria podem ser levados em consideração na busca pela bipartição total. v) A estimação dos parâmetros não foi nosso foco durante este trabalho e utilizamos estimativas amostrais e de máxima verossimilhança. Contudo, uma abordagem Bayesiana é aplicável tanto na estimação dos parâmetros da regressão e outros parâmetros envolvidos, quanto na estimação dos riscos binomiais e, consequentemente, na busca por uma bipartição total. vi) Como explorado em Basu e Pereira (1982), a distribuição Dirichlet também pode ser fatorada em betas e toda a formulação apresentada neste trabalho se estende neste contexto.

Referências Bibliográficas

Agresti(2002) Alan Agresti. *Categorical data analysis*. John Wiley & Sons, 2ª edição.

Aitchison(1986) John Aitchison. *The statistical analysis of compositional data*. Chapman and Hall.

Aitchison *et al.***(2004)** John Aitchison, Jim W. Kay e Ian J. Lauder. *Statistical concepts and applications in clinical medicine*. CRC Press.

Andersen(1970) Erling B. Andersen. Sufficiency and exponential families for discrete sample spaces. *Journal of the American Statistical Association*, 65(331):1248–1255.

Aranda-Ordaz(1981) Francisco J. Aranda-Ordaz. On two families of transformations to additivity for binary response data. *Biometrika*, 68(2):357–363.

Basu e Pereira(1982) Debrabata Basu e Carlos Alberto de B. Pereira. On the bayesian analysis of categorical data: the problem of nonresponse. *Journal of Statistical Planning and Inference*, 6 (4):345–362.

Berkson(1944) Joseph Berkson. Application of the logistic function to bio-assay. *Journal of the American Statistical Association*, 39(227):357–365.

Bishop(2011) Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Springer.

Bliss(1935) Chester I. Bliss. The calculation of the dosage-mortality curve. *Annals of Applied Biology*, 22(1):134–167.

Brier(1950) Glenn W. Brier. Verification of forecasts expressed in terms of probability. *Monthey Weather Review*, 78(1):1–3.

Caron e Polpo(2009) Renaut Caron e Adriano Polpo. Binary data regression: Weibull distribution. Em *Bayesian Inference and Maximum Entropy methods in Science and Engineering: The 29th International Workshop on Bayesian Inference and Maximum Entropy Methods in Science and Engineering*, volume 1193, páginas 187–193. AIP Publishing.

Catalan(1838) Eugene Catalan. Note sur une équation aux différences finies. *Journal de mathématiques pures et appliquées*, 3:508–516.

Comtet(1974) Louis Comtet. *Advanced Combinatorics: The Art of Finite and Infinite Expansions*. D. Reidel Publishing Company.

Cox(1958) David R. Cox. The regression analysis of binary sequences. *Journal of the Royal Statistical Society. Series B (Methodological)*, 20(2):215–242.

Diniz(2015) Márcio A. Diniz. *Modelos Bayesianos semi-paramétricos para dados binários*. Tese de Doutorado, Instituto de Matemática e Estatística, Universidade de São Paulo.

Eugenio(2016) Nicholas W. Eugenio. Modelo de regressão para dados binários com mistura de funções de ligação. Dissertação de Mestrado, Instituto de Matemática e Estatística, Universidade de São Paulo.

Feller(1968) William Feller. *An introduction to probability theory and its applications*, volume 1. Wiley, 3ª edição.

Fisher(1922) Ronald A. Fisher. On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, 222(594-604):309–368.

Friedman(1996) Jerome Friedman. Another approach to polychotomous classification. Relatório técnico, Department of Statistics, Stanford University.

Hastie e Tibshirani(1998) Trevor Hastie e Robert Tibshirani. Classification by pairwise coupling. Em *Advances in neural information processing systems*, páginas 507–513.

Hastie *et al.***(2016)** Trevor Hastie, Robert Tibshirani e Jerome Friedman. *The elements of statistical learning*. Springer, 2ª edição.

Hosmer *et al.***(2013)** David W. Hosmer, Stanley Lemeshow e Rodney X. Sturdivant. *Applied logistic regression*. John Wiley & Sons, 3ª edição.

Izbicki(2018) Rafael Izbicki. *Machine Learning sob a ótica estatística: Uma abordagem preditivista para a estatística com exemplos em R*. URL https://rizbicki.files.wordpress.com/2016/09/main2.pdf. Acesso em 10 de junho de 2018.

Moon(1970) John W. Moon. *Counting labelled trees*. William Clowes and Sons.

Murtagh(1984) Fionn Murtagh. Counting dendrograms: a survey. *Discrete Applied Mathematics*, 7(2):191–199.

Nelder e Wedderburn(1972) John A. Nelder e Robert W. M. Wedderburn. Generalized linear models. *Statist. Soc A*, 1972.

Pearl e Reed(1920) Raymond Pearl e Lowell J. Reed. On the rate of growth of the population of the united states since 1790 and its mathematical representation. *Proceedings of the national academy of sciences*, 6(6):275–288.

Pereira(1980) Carlos Alberto de B. Pereira. *Bayesian solutions to some classical problems of statistics*. Tese de Doutorado, The Florida State University.

Pereira e Stern(2008) Carlos Alberto de B. Pereira e Julio M. Stern. Special characterizations of standard discrete models. *REVSTAT–Statistical Journal*, 6(3):199–230.

Prentice(1976) Ross L. Prentice. A generalization of the probit and logit methods for dose response curves. *Biometrics*, páginas 761–768.

Reed e Berkson(1929) Lowell J. Reed e Joseph Berkson. The application of the logistic function to experimental data. *The Journal of Physical Chemistry*, 33(5):760–779.

Shavitt *et al.***(2017)** Roseli G. Shavitt, Guaraci Requena, Pino Alonso, Gwyneth Zai, Daniel LC Costa, Carlos Alberto de B Pereira *et al.* Quantifying dimensional severity of obsessive-compulsive disorder for neurobiological research. *Progress in Neuro-Psychopharmacology and Biological Psychiatry*, 79:206–212.

Stanley(1999) Richard P. Stanley. *Enumerative Combinatorics*, volume 1. Cambridge University Press.

Stukel(1988) Therese A. Stukel. Generalized logistic models. *Journal of the American Statistical Association*, 83(402):426–431.

Vapnik(1998) Vladimir Vapnik. *Statistical learning theory*. Wiley, New York.