GILBERTO RODRIGUES LISKA

REGRESSÃO SIMPLEX APLICADA A DELINEAMENTOS DE MISTURA E UTILIZAÇÃO DO ALGORITMO BOOSTING

LAVRAS - MG

2016

GILBERTO RODRIGUES LISKA

REGRESSÃO SIMPLEX APLICADA A DELINEAMENTOS DE MISTURA E UTILIZAÇÃO DO ALGORITMO BOOSTING

Tese apresentada à Universidade Federal de Lavras, como parte das exigências do Programa de Pós-graduação em Estatística e Experimentação Agropecuária, área de concentração em Estatística e Experimentação Agropecuária, para a obtenção do título de Doutor.

Prof. Dr. Marcelo Ângelo Cirillo
Orientador

Prof. Dr. Fortunato Silva de Menezes
Coorientador

LAVRAS - MG
2016

Ficha catalográfica elaborada pelo Sistema de Geração de Ficha Catalográfica da Biblioteca Universitária da UFLA, com dados informados pelo(a) próprio(a) autor(a).

Liska, Gilberto Rodrigues.
Regressão simplex aplicada a delineamentos de mistura e utilização do algoritmo boosting : Regressão simplex aplicada a delineamentos de mistura e utilização do algoritmo boosting / Gilberto Rodrigues Liska - Lavras: UFLA, 2016.
206 p. : il.

Tese(doutorado)-Universidade Federal de Lavras, 2016.
Orientador(a): Marcelo Ângelo Cirillo.
Bibliografia.

1. Modelo Linear Generalizado. 2. Proporção. 3. Algoritmo Boosting. 5. Modelo de Mistura. 6. Região SImplex I. Universidade Federal de Lavras. II. Título.

GILBERTO RODRIGUES LISKA

REGRESSÃO SIMPLEX APLICADA A DELINEAMENTOS DE MISTURA E UTILIZAÇÃO DO ALGORITMO BOOSTING

Tese apresentada à Universidade Federal de Lavras, como parte das exigências do Programa de Pós-graduação em Estatística e Experimentação Agropecuária, área de concentração em Estatística e Experimentação Agropecuária, para a obtenção do título de Doutor.

APROVADA em 19 de maio de 2016.

Dr. Fortunato Silva de Menezes	UFLA
Dr. Júlio Sílvio de Sousa Bueno Filho	UFLA
Dr. Luiz Alberto Beijo	UNIFAL-MG
Dra. Carla Regina Guimarães Brighenti	UFSJ

Prof. Dr. Marcelo Ângelo Cirillo
Orientador

LAVRAS - MG
2016

Aos meus pais Lúcia (*in memoriam*) e Istvan (*in memoriam*),

pelo amor, carinho e educação.

Aos meus queridos irmãos Estevan e Geraldo.

À minha esposa Grazielle.

DEDICO

AGRADECIMENTOS

A Deus. Muito obrigado por ter me dado a oportunidade de cursar um doutorado e de concluí-lo. Muito obrigado por ter me dado força, ânimo, paciência e comprometimento com os deveres dessa empreitada. Fico intrigado em me questionar se realmente existe um Deus que pode nos proporcionar coisas do tipo e logo fico emocionado em saber que sim, isso é possível, mas não sei explicar o porquê disso. Acredito que muitas pessoas que experimentam situações extremas, adversas e que exigem muito de si sabem o que quero dizer. Enfim, muito obrigado, meu Deus!

Aos meus pais, Istvan e Lúcia, que não estão mais entre nós, pelo amor infinito que criou todas as condições que me permitiram concluir mais esta etapa. Gostaria de agradecer pelas virtudes as quais admirava neles, como a sabedoria e a inteligência em meu pai e a coragem, a força e a garra em minha mãe.

Aos meus irmãos, Geraldo e Estevan. Estão seguindo suas vidas hoje, mas gostaria de destacar aqui a contribuição que eles tiveram na minha época de graduação, principalmente no momento em que ficamos sem nossos pais. Desejo a vocês que alcancem seus objetivos e que Deus os conduza da melhor maneira possível.

À minha esposa Grazielle Aparecida Cassimiro, pela amizade, companheirismo, paciência e força que me sustentaram durante todo o tempo em que estivemos e estamos juntos, mesmo que às vezes distantes um do outro.

Ao Professor Dr. Luiz Alberto Beijo, meu primeiro orientador, por ter me apresentado e conduzido na carreira de docente de ensino superior na área da Estatística.

Ao Professor Dr. Fortunato Silva de Menezes, pela orientação recebida no mestrado, pela paciência nas explicações, pela sabedoria nas horas difíceis e pelo

apoio e incentivo nos trabalhos.

Ao Professor Dr. Marcelo Ângelo Cirillo, que aceitou com satisfação o convite para me orientar no doutorado. Foram quatro anos de muito aprendizado, orientação e trabalho. Por isso, sou muito grato a ele! Além disso, por me orientar no caminho das pedras, ter me proporcionado conhecimentos que serão aperfeiçoados e que os levarei para o resto de minha vida.

Aos meus amigos de curso e república, Guido Gustavo Humada Gonzalez, Juliano Bortolini, Rossicley Rangel Paiva, Ben Dêivide de Oliveira Batista, Naje Clécio Nunes da Silva e demais amigos que estiveram comigo nessa batalha. Obrigado a todos que me apoiaram nos momentos difíceis!

À minha ex-chefe, Célia Pereira de Araújo, quando funcionário do Programa de Saúde da Família - Caensa, na função de Agente Comunitário de Saúde, pelo apoio, até então, na decisão mais difícil de minha vida: a saída desse serviço para a dedicação exclusiva aos estudos. Muito obrigado por acreditar no meu potencial.

A todos os brasileiros que pagam seus impostos honestamente e permitem que instituições como a UFLA, CAPES e CNPq mantenham cursos de alto nível e ofereçam bolsas aos alunos. Um agradecimento especial à agência de fomento CAPES por ter me concedido bolsa de estudos durante meu mestrado e doutorado.

Enfim, não posso deixar de agradecer a todos que torceram, incentivaram e que, diretamente ou indiretamente, contribuíram pelo sucesso desta empreitada. Muito obrigado!

"*Força!! ... Sangue!! ... Fibra!! ... Moral!! ... Ralação!! ... Vibração!! ...*
Ralação!! ... Vibração!! ..."

Canção cantada pelos atiradores durante os serviço militar nas corridas feitas pelas ruas da cidade de Alfenas-MG.

RESUMO GERAL

Na composição deste trabalho estão presentes duas partes. A primeira parte contém a fundamentação teórica do presente estudo. A segunda parte contém dois artigos científicos. No primeiro artigo são abordados dois modelos da classe dos modelos lineares generalizados para analisar um experimento de mistura que consistiu em estudar o efeito de diferentes dietas compostas por gordura, carboidrato e fibra sobre a expressão de tumor nas glândulas mamárias em ratos fêmeas, dada pela proporção de ratos que tiveram a expressão do tumor numa determinada dieta. Experimentos de mistura são caracterizados por apresentarem o efeito da colinearidade e tamanho amostral reduzido. Nesse sentido, assumir normalidade para a resposta a ser maximizada ou minimizada pode ser inadequado. Diante desse fato, são abordadas as principais características dos modelos de regressão logística e simplex. Os modelos foram comparados mediante os critérios de seleção de modelos AIC, BIC e ICOMP, gráficos de envelope simulado para os resíduos dos modelos ajustados, gráficos das razões de chances e seus respectivos intervalos de confiança para cada componente de mistura. Concluiu-se nesse primeiro artigo que o modelo de regressão simplex apresentou melhor qualidade de ajuste e produziu intervalos de confiança para a razão de chances mais precisos. O segundo artigo apresenta o modelo *Boosted Simplex Regression*, a versão boosting do modelo de regressão simplex, como uma alternativa de aumentar a precisão dos intervalos de confiança para a razão de chances em cada componente de mistura. Para tal, foi utilizado o método de Monte Carlo para a construção dos respectivos intervalos de confiança. Além disso, é apresentado de maneira inovadora o gráfico de envelope simulado para os resíduos do modelo ajustado via algoritmo boosting. Foi possível concluir que o modelo *Boosted Simplex Regression* se ajustou satisfatoriamente e produziu intervalos de confiança para a razão de chances acurados e ligeiramente mais precisos do que sua versão ajustada pelo método da máxima verossimilhança.

Palavras-chave: Modelo Linear Generalizado. Proporção. Algoritmo Boosting. Modelo de Mistura. Região Simplex.

GENERAL ABSTRACT

In the composition of this work are present two parts. The first part contains the theory used. The second part contains the two articles. The first article examines two models of the class of generalized linear models for analyzing a mixture experiment, which studied the effect of different diets consist of fat, carbohydrate, and fiber on tumor expression in mammary glands of female rats, given by the ratio mice that had tumor expression in a particular diet. Mixture experiments are characterized by having the effect of collinearity and smaller sample size. In this sense, assuming normality for the answer to be maximized or minimized may be inadequate. Given this fact, the main characteristics of logistic regression and simplex models are addressed. The models were compared by the criteria of selection of models AIC, BIC and ICOMP, simulated envelope charts for residuals of adjusted models, odds ratios graphics and their respective confidence intervals for each mixture component. It was concluded that first article that the simplex regression model showed better quality of fit and narrowest confidence intervals for odds ratio. The second article presents the model Boosted Simplex Regression, the boosting version of the simplex regression model, as an alternative to increase the precision of confidence intervals for the odds ratio for each mixture component. For this, we used the Monte Carlo method for the construction of confidence intervals. Moreover, it is presented in an innovative way the envelope simulated chart for residuals of the adjusted model via boosting algorithm. It was concluded that the Boosted Simplex Regression model was adjusted successfully and confidence intervals for the odds ratio were accurate and lightly more precise than the its maximum likelihood version.

Keywords: Generalized Linear Model. Proportion. Boosting Algorithm. Mixture Model. Simplex Space.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE SIGLAS

FGD	Functional Gradient Descent (Gradiente de Descida Funcional)
GBF	Gradiente Boosting de Friedman
MLE	Maximum Likelihood Estimator (Estimador de Máxima Verossimilhança)
MLG	Modelo Linear Generalizado
BFGS	Método de Quase-Newton desenvolvido por Broyden, Fletcher, Goldfarb e Shanno
AIC	Akaike Information Criterion (Critério de Informação de Akaike)
BIC	Bayesian Information Criterion (Critério de Informação Bayesiano)
ICOMP	Índice de complexidade da informação de Bozdogan
DMBA	Dimetil-benzathracene (Dimetil-benzantraceno)

LISTA DE SÍMBOLOS

q	Número de componentes de mistura
p	Número de parâmetros
n	Número de observações
x_i	Proporção do i-ésimo componente de mistura $(i = 1, ..., q)$
L_i	Limite inferior do i-ésimo componente de mistura $(i = 1, ..., q)$
U_i	Limite superior do i-ésimo componente de mistura $(i = 1, ..., q)$
x_i'	Proporção do i-ésimo pseudo-componente de mistura (i = 1, ..., q)
β_i^*	i-ésimo parâmetro confundido com β_0 associado ao i-ésimo componente de mistura
η	Função das variáveis de mistura x_i. É também a componente sistemática em um MLG
ϵ	Erro experimental
β_{-i}	Parâmetro do i-ésimo termo inverso x $(i = 1, ..., q)$
w_i	i-ésima variável razão $(i = 1, ..., q - 1)$
M	Número de iterações do algoritmo Boosting $(m = 1, ..., M)$
$f_m(\mathbf{x})$	Procedimento base do algoritmo AdaBoost
$F(\mathbf{x})$	Classificador final do algoritmo AdaBoost
W_i^m	Peso da i-ésima observação $(i = 1, ..., n)$ na m-ésima iteração do algoritmo AdaBoost
Z^m	Constante de normalização no algoritmo AdaBoost
$f*(\cdot)$	Função de predição ótima, ou minimizador populacional, no algoritmo GBF
$\rho(y, g)$	Função perda do algoritmo GBF
$g(\cdot)$	Procedimento base no algoritmo GBF. É também a função de ligação em um MLG.
$\hat{g}^{(m)}(\cdot)$	Procedimento base ajustado na iteração m do algoritmo GBF (m=1, ..., M)
$\hat{f}^{(m)}(\cdot)$	Modelo ajustado na iteração m do algoritmo GBF
$R[\cdot, \cdot]$	Risco empírico
$\mathbf{u}^{(m)}$	Vetor gradiente negativo no algoritmo GBF
v	Comprimento do passo no algoritmo GBF
$\tilde{Y}, \tilde{y}$	Valor marginal
$\pi(\mathbf{x})$	Probabilidade de ocorrer $Y = 1$ dado $\mathbf{X} = \mathbf{x}$ em um ensaio Bernoulli
$ln(\cdot)$	Logaritmo neperiano de um dado argumento
$\hat{\beta}^{(j)}$	Parâmetro estimado na variável j $(j = 1, ..., p)$ no algoritmo GBF

LISTA DE SÍMBOLOS

$\hat{\lambda}$	Índice da variável que minimiza a soma de quadrados dos desvios do vetor gradiente com o procedimento base ajustado para a variável j
$\hat{\beta}^{(\hat{\lambda}_m)}$	Estimativa do parâmetro da variável selecionada na m-ésima iteração do algoritmo GBF
$x^{(\hat{\lambda}_m)}$	Variável selecionada na m-ésima iteração do algoritmo GBF
$l(\cdot)$	Logaritmo da função de verossimilhança
$U_\beta(\cdot)$	Vetor Escore ou vetor gradiente
I_β	Matriz hessiana
$I_E(\cdot)$	Matriz de Informação de Fisher esperada
$\mathbf{X}$	Matriz de delineamento de dimensão $n \times p$
$\mathbf{Y}$	Vetor dos valores y_i de dimensão $n \times 1$
$\mathbf{V}$	Vetor das probabilidades ajustadas de dimensão $n \times 1$ do modelo de regressão logística
$\mathbf{x}_i$	Vetor correspondente à i-ésima linha da matriz $\mathbf{X}$ de dimensão $p \times 1$
$\mathbf{W}$	Matriz diagonal de dimensão $n \times n$
δ	Nível de tolerância de um algoritmo de otimização
t_{D_i}	Resíduo padronizado d modelo de regressão logística
$\hat{h}_{ii}$	medida de influência (leverage) estimada
K	Número de iterações do algoritmo para construção do envelope simulado $(k = 1, ..., K)$
$\mathbf{H}$	Matriz chapéu (hat) de dimensão $n \times n$ que contém os $\hat{h}_{ii}$
r_i^{pp}	Resíduo ponderado padronizado do modelo de regressão simplex
$df\,(m)$	Graus de liberdade na m-ésima iteração do algoritmo GBF
c_i	Proporção do i-ésimo componente da mistura de referência $(i = 1, ..., q)$
Δ_i	Incremento no componente de mistura x_i na direção Cox
ρ_{x_i}	Razão do componentes exceto x_i no ponto de referência
$OR(x_i)$	Razão de chances de um grupo controle (ou mistura de referência) em relação à incrementos no componente de mistura x_i
z_α	Quantil da distribuição normal padrão com nível α de significância

SUMÁRIO

PRIMEIRA PARTE

1 INTRODUÇÃO GERAL

Experimentos de mistura são frequentemente utilizados em muitos campos da ciência, como em química, farmácia e indústria de produtos para consumidores, como por exemplo na formulação dos ingredientes que minimizam o teor de açucar de uma dieta. Basicamente, um experimento de mistura consiste em otimizar uma variável resposta considerando-se algumas variáveis de entrada, ou variáveis independentes, em que a quantidade de cada ingrediente é dada como uma proporção dos ingredientes e a soma das proporções dos ingredientes deve somar um. Esse conceito pode ser estendido para otimização de mais de uma resposta, constituindo o contexto de otimização de respostas simultâneas. Devido à sua ampla aplicação, muitos estudos têm sido feitos na construção de modelos, como os modelos de Scheffé (SCHEFFÉ, 1958) e de termos inversos (DRAPER, ST. JOHN, 1977). Novos modelos têm sido propostos quando a resposta de interesse não é proveniente de uma distribuição normal, como por exemplo quando a variável a ser analisada é resultante da proporção de um dado evento de interesse (CHEN; JACKSON, 1996).

restritas ao intervalo unitário, logo, assumir normalidade para os erros pode levar a predições fora do intervalo unitário, ou seja, pode fornecer predições negativas ou maiores do que um. Além disso, nessa abordagem, a estimação intervalar e testes de hipóteses assume que a resposta é distribuída simetricamente, o que pode levar a conclusões espúrias.

Dada a natureza da variável resposta do presente estudo, foram consideradas as distribuições Binomial e Simplex na abordagem de modelos lineares generalizados (MLG). Os parâmetros dos modelos de regressão em MLG geralmente são estimados por máxima verossimilhança ou mínimos quadrados reponderados.

Uma observação a ser feita é a de que os delineamentos usuais de mistura, como o Centróide-Simplex ou Vértice Extremo, são caracterizados por apresentarem tamanho amostral reduzido. Esse fato pode influenciar de certa forma na variância do estimador de máxima verossimilhança, inflacionando-o. Para contornar esse problema, foi proposta a versão boosting do modelo de regressão simplex.

Outra característica dos experimentos de mistura é o fato de que os modelos de mistura são altamente afetados pela colinearidade. Essa colinearidade é devida à restrição unitária dos componentes de mistura, o que faz com que as colunas associadas aos componentes de mistura da matriz de delineamento sejam linearmente dependentes. Esse fato influencia ainda na covariância dos parâmetros do modelo, fazendo com que as mesmas sejam maiores quanto maior for o grau de colinearidade. Várias alternativas têm sido propostas na literatura para contornar esse problema e trabalhos mostram que um meio é considerar no preditor linear de um MLG as variáveis razão, o que diminui a covariância entre os parâmetros do preditor linear do modelo.

Especialmente, quando a variável resposta assume uma natureza binária ou proporção, uma medida de grande interesse prático é a razão de chances e, nesse caso, os métodos convencionais de análise e interpretação dos parâmetros de um modelo de mistura não são adequados, uma vez que quando um componente varia, os outros componentes devem variar conjuntamente. Nesse caso, a direção Cox é utilizada para quantificar o efeito dos componentes sobre a resposta predita.

Diante do exposto, o trabalho tem por objetivo comparar os modelos de regressão, considerando a distribuição simplex e binomial, na análise de delineamentos de mistura.

Dentre os objetivos específicos destacam-se:

a) Comparar o emprego do modelo de Scheffé, comumente usado em experimen-

tos de mistura, em relação ao modelo com variáveis razão. Ambos modelos foram avaliados considerando-se a suposição de resposta binomial e simplex.

b) Comparar os gráficos de razões de chances e seus respectivos intervalos de confiança, construídos utilizando-se a direção Cox, com relação à amplitude dos mesmos.

c) Propor o modelo *Boosted Simplex Regression* como um método capaz de estimar razões de chances acuradas e precisas em experimento de mistura. Fornecer gráficos de disgnóstico baseados em bandas de confiança obtidas por simulação.

d) Desenvolver e disponibilizar um pacote em R para permitir a reprodutibilidade da metodologia proposta. Tal pacote será chamado de *SimplexMixmodel*.

O trabalho está organizado em formato de artigo, sendo constituída por partes que se compõem da seguinte forma:

- A primeira parte é composta de uma introdução geral, dos objetivos e na sequência é exposto o referencial teórico para fundamentação do que é proposto nos artigos.
- A segunda parte é constituída por dois artigos:
  - **i.** O artigo 1 consistiu em se ajustarem os modelos de Scheffé e de variáveis razão utilizando-se os modelos de regressão logística e simplex. A partir dos resultados obtidos foram feitas considerações sobre os modelos de regressão logística ajustados em relação aos modelos de regressão simplex ajustados.

ii. O artigo 2 consistiu em apresentar o modelo *Boosted Simplex Regression*, a versão boosting da regressão simplex. O modelo proposto foi comparado com sua versão ajustada pelo método da máxima verossimilhança e a regressão logística. Os principais aspectos estatísticos foram discutidos.

- As considerações finais são apresentadas no fim da segunda parte da tese, bem como aspectos sobre o tema do trabalho desenvolvido e algumas perspectivas de estudos futuros.

- Nos apêndices se encontram uma ilustração didática do algoritmo boosting e uma rotina no software R da versão preliminar do pacote *SimplexMixmodel*.

2 REFERENCIAL TEÓRICO

Serão apresentadas inicialmente nesta seção as principais características sobre a região experimental avaliada no presente trabalho, a região simplex. Será feita uma introdução do método de boosting e o algorítmo utilizado para ajustar modelos lineares generalizados. Serão apresentados também os modelos de regressão logística e simplex. Em seguida, serão discutidos os critérios de adequabilidade de ajuste utilizados para se efetuarem as comparações dos modelos ajustados via boosting e os mesmos estimados via máxima verossimilhança.

2.1 Experimentos de Mistura

Em experimentos de mistura, assume-se que a resposta mensurada é dependente apenas das proporções dos ingredientes presentes em uma mistura. Assumindo uma mistura com q componentes, sendo que x_i ($0 \leq x_i \leq 1$) representa a proporção do i-ésimo componente ($i = 1, 2, \ldots, q$). Segundo Box e Draper (2007), a restrição que caracteriza simplex é dada por.

$$\sum_{i=1}^{q} x_i = 1. \tag{2.1}$$

O espaço que compõe os q componentes assume a forma de um simplex regular de dimensão $(q-1)$. Consideremos uma situação simples, definindo $q = 2$, portanto $x_1 + x_2 = 1$. Nesse caso, x_1 e x_2 devem estar no intervalo $0 \leq x_i \leq 1$. Caso um componente tenha proporção igual a um, ou seja $x_i = 1$, os demais componentes terão proporção iguais a zero e a referida mistura recebe o nome de *mistura pura* para um determinado componente.

Os extremos do espaço experimental ocorrem em $x_1 = 1$, $x_2 = 0$ e $x_1 =$

0, $x_2 = 1$. Geometricamente, todas as possíveis misturas estão contidas na linha $x_1 + x_2 = 1$ no espaço (x_1, x_2), que é representado por um plano. Nesse caso, o espaço da mistura é compreendido por todos os pontos que estão entre (0,1) e (1,0), inclusive eles. Essa situação é ilustrada na Figura 1.

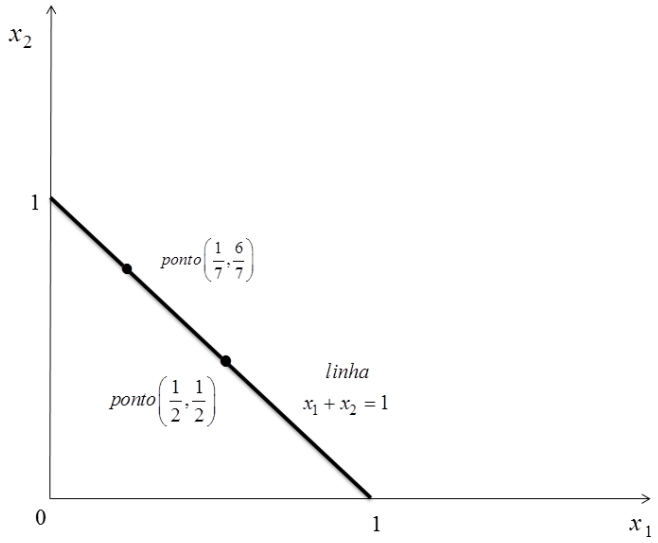


Figura 1 Região experimental de uma mistura com 2 componente, $x_1 + x_2 = 1$.

Em uma situação em que $q = 3$, a região simplex é um espaço de mistura triangular, com $x_1 + x_2 + x_3 = 1$. As três misturas $(1, 0, 0)$, $(0, 1, 0)$ e $(0, , 0, 1)$ definem os vértices, também chamados de vertex, cujos pares formam uma linha.

A junção das três linhas formadas por esses vértices definem um triângulo, cuja região interna é o espaço de mistura reduzido para os três componentes, ou seja, dentre todos os possíveis espaços simplex que satisfazem a restrição em (2.1), somente o construído com os vértices acima contém valores positivos para

todos os componentes individuais. Observe que fora dos limites compostos pelos vértices, pelo menos um x_i é negativo, o que não é possível em misturas reais. Por exemplo, o ponto $(-0.5, 0.5, 1)$ satisfaz a restrição em (2.1), porém não faz sentido considerar uma proporção negativa para o componente x_1.

Cada ponto interior ao triângulo possui três coordenadas, mas são necessárias apenas duas coordenadas para especificar um ponto qualquer ao triângulo, uma vez que a terceira coordenada assume o valor 1 - (soma dos outros dois componentes). Note-se que cada face do triângulo (Figura 2) representa o espaço de mistura de dois componentes, conforme ilustrado na Figura 1.

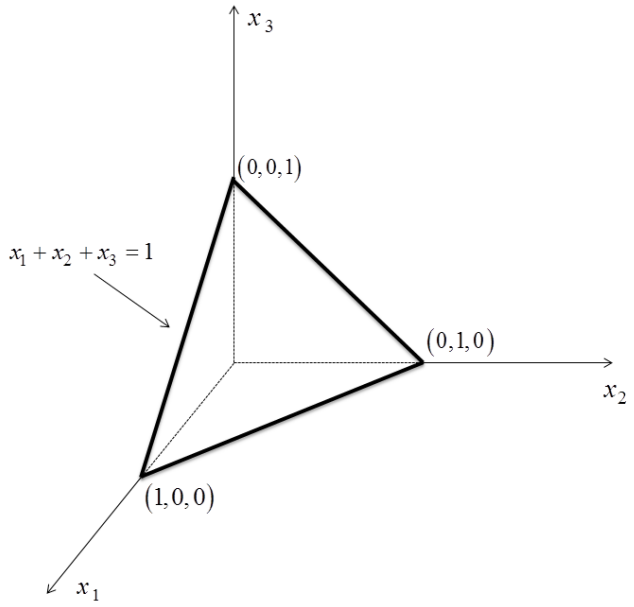


Figura 2 Região experimental de uma mistura com $q = 3$ componentes, $x_1 + x_2 + x_3 = 1$.

A identificação de cada componente dentro do simplex é feita por “linhas

imaginárias", paralelas ao vértice correspondente. Visando melhores esclarecimentos, consideremos inicialmente $x_1 = 0.8$ (Figura 3). Aumentando x_1 no sentido $C \to A$, segmento $\overrightarrow{CA}$, sendo $x_1 = 0$ no segmento $\overrightarrow{BC}$, "linhas imaginárias" são traçadas paralelamente ao segmento $\overrightarrow{BC}$. Da mesma forma, procede-se para qualquer vértice do simplex.

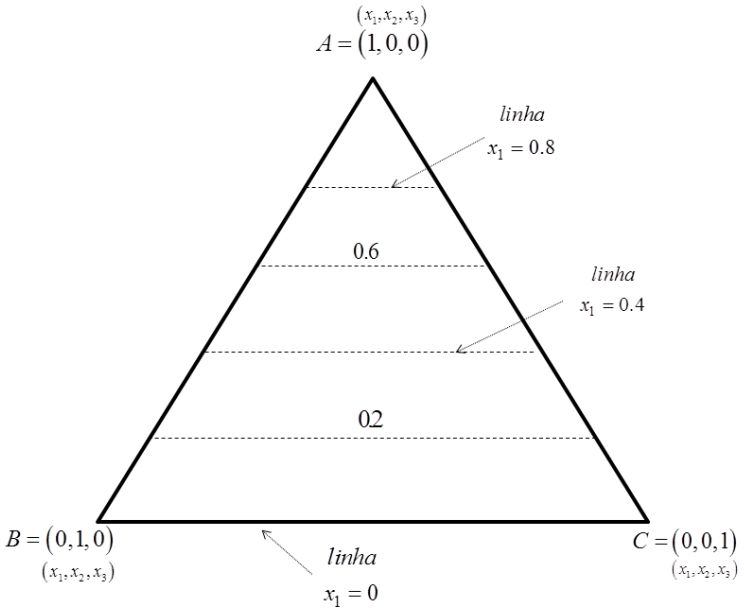


Figura 3 Coordenadas das linhas paralelas em que $x_1 = 0; 0.2; 0.4; 0.6; 0.8$ no espaço de mistura $x_1 + x_2 + x_3 = 1$.

A Figura 4 mostra o único ponto definido pela interseção das linhas $x_1 = 0.35$ e $x_2 = 0.45$. Naturalmente, $x_3 = 1 - x1 - x2 = 0.20$, então o ponto $(x_1, x_2, x_3) = (0.35, 0.45, 0.20)$.

Não é possível projetar uma espaço de quatro dimensões em (x_1, x_2, x_3, x_4), como na figura 2, mas pode-se estender o conceito de mistura fazendo uma analo-

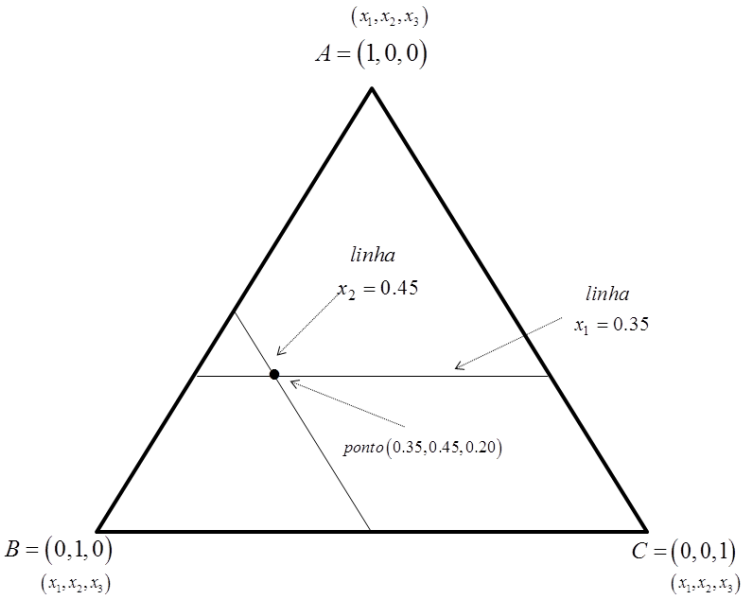


Figura 4 O ponto $(0.35, 0.45, 0.20)$ é definido como a interseção das duas linhas de coordenadas $x_1 = 0.35$ e $x_2 = 0.45$. Note que $x_3 = 1 - x_1 - x_2 = 0.20$.

gia do espaço de mistura na figura 3 como sendo uma caso particular de um espaço de mistura de três dimensões, em que $x_4 = 0$. Nesse caso, o espaço da mistura é um tetraedro regular e ao invés de retas paralelas definindo as misturas, tem-se uma série de triângulos paralelos a uma base qualquer selecionada tetraedro.

Em síntese, o conceito de espaço de mistura pode ser estendido para qualquer espaço de dimensão $q - 1$, que é um simplex regular formado por q pontos igualmente espaçados. Quando $q \geq 6$, qualquer visualização da região simplex se torna uma tarefa extremamente difícil (BOX; DRAPER, 2007).

Por questões econômicas e/ou físicas, às vezes são impostas restrições adicionais sobre os componentes individuais

$$0 \leq L_i \leq x_i \leq U_i \leq 1;\ i = 1,\ 2,\ \ldots,\ q, \tag{2.2}$$

sendo L_i e U_i, respectivamente, os limites inferiores e superiores. A restrição (2.2) reduz a região restrita dada na equação (2.1).

Devido à restrição (2.2) feita nos componentes, sub-regiões dentro do simplex são criadas, de forma que, todas as possíveis combinações das proporções dos componentes envolvidos na mistura sejam contemplados. A título de ilustração, observe-se a sub-região definida no simplex da Figura 5. Os componentes de mistura dessas sub-regiões podem ser reescritos em termos de suas restrições e componentes originais, como pode ser visto na equação (2.3),

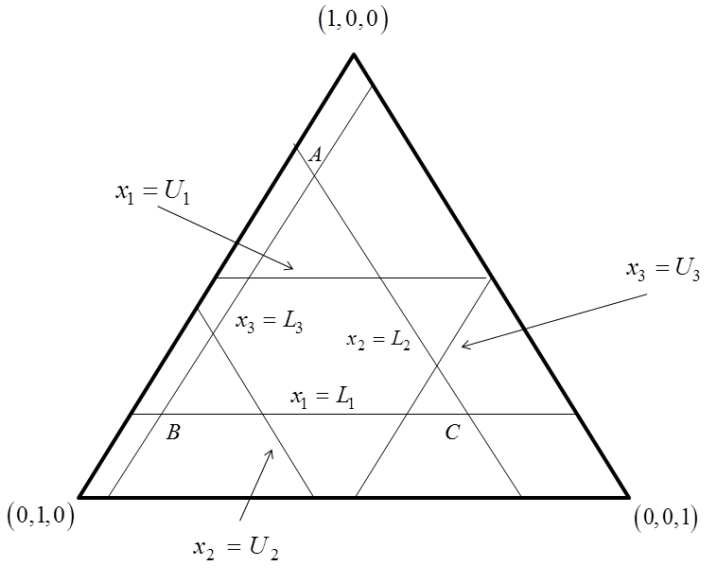


Figura 5 Região simplex restrita pelos limites inferiores L_i e superiores U_i para cada componente de mistura x_i, em que $q = 3$.

$$x_i' = \frac{x_i - L_i}{1 - L}, \tag{2.3}$$

em que $L = \sum_{i=1}^{q} L_i$, cujo valore é inferior à restrição em (2.1). Caso a sub-região do simplex seja também um simplex, o uso dos L-pseudo-componentes simplifica a construção do experimento por permitir a aplicabilidade dos delineamentos Látice-Simplex e Centróide-Simplex.

Esses novos componentes são chamados de pseudo-componentes e foram definidos por Piepel (1982), com o propósito de reduzir o efeito da multicolinearidade presente em experimentos do tipo. O termo "efeito" quando relacionado à presença de multicolinearidade é no sentido de que trata-se de um efeito que não é estimado ou faz parte do modelo, e sim de uma característica que ocorre naturalmente em modelos cujos termos podem ser reescritos como combinações lineares de outros termos. Na seção 2.1.2 são apresentadas algumas alternativas para reduzir o efeito da multicolinearidade além da inclusão de pseudo-componentes.

Para mais esclarecimentos, retornamos ao exemplo com $q = 3$ componentes, porém os limites inferiores L_1, L_2 e L_3 associados a cada componente, de modo que $x_1 \geq L_1$, $x_2 \geq L_2$ e $x_3 \geq L_3$, logo $L = L_1 + L_2 + L_3$ e, naturalmente, $L \leq 1$. Se $L = 1$, a região restrita consiste de um ponto apenas, ou seja, (L_1, L_2, L_3), então é coerente especificar $L < 1$.

Uma ilustração dessa região restrita é apresentada na Figura 6. Note-se que a nova região simplex assume exatamente a mesma forma que a região original e está totalmente contida nela. O vértice A^*, que está posicionado na interseção das linhas $x_2 = L_2$ e $x_3 = L_3$, tem coordenadas $(1 - L_2 - L_3, L_2, L_3)$. Analogamente, o ponto B^* em $(L_1, 1 - L_1 - L_3, L_3)$ e C^* em $(L_1, L_2, 1 - L_1 - L_2)$.

Retornando à definição de pseudo-componentes (2.3), porém substituindo $x_i = L_i$, pode-se observar que $x_i' = 0$ e que

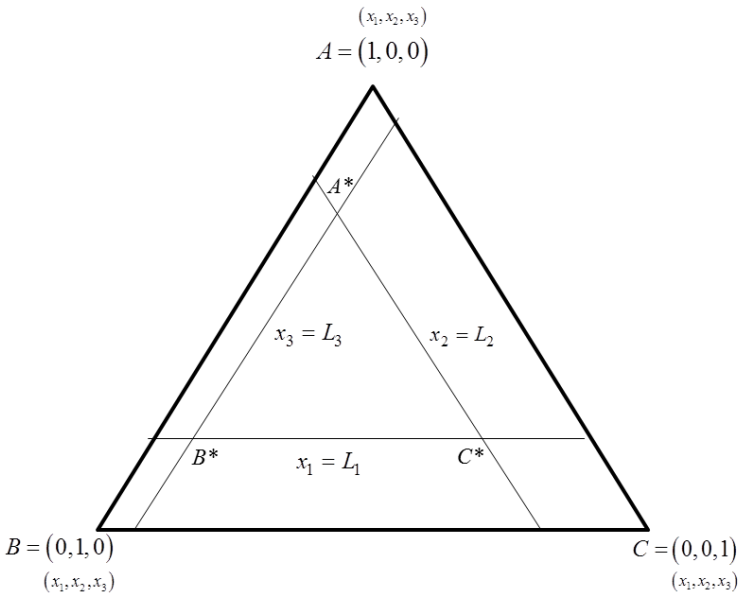


Figura 6 Região restrita para $q = 3$ ingredientes definido pelos três limites inferiores L_1, L_2 e L_3.

$$x'_1 + x'_2 + x'_3 = \frac{x_1 - L_1}{1 - L} + \frac{x_2 - L_2}{1 - L} + \frac{x_3 - L_3}{1 - L} = \frac{(x_1 + x_2 + x_3 - L_1 - L_2 - L_3)}{1 - L} = 1.$$

Portanto, pode-se concluir que a soma das proporções dos pseudo-componentes é 1. A Figura 7 ilustra a sub-região formada pelos pseudo-componentes, em que o ponto A se torna $(x'_1, x'_2, x'_3) = (1, 0, 0)$, B é $(0, 1, 0)$ e C é $(0, 0, 1)$. A extensão de pseudo-componentes para misturas com outros valores de q é feita de maneira similar ao caso apresentado.

Em relação à interpretação dos resultados, deve-se tomar o cuidado ao utilizar os pseudo-componentes, uma vez que essa transformação não proporciona

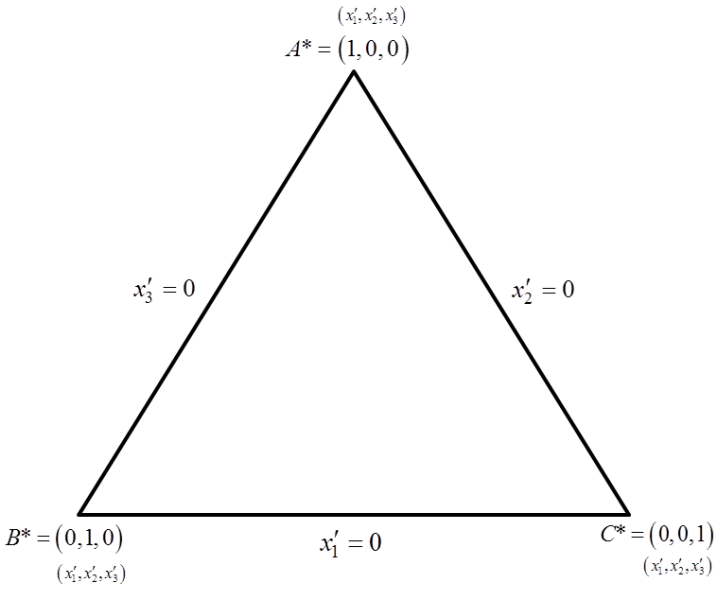


Figura 7 Região simplex restrita em pseudo-componentes para $q = 3$.

uma inferência direta sobre os componentes originais, pois o ajuste do modelo foi feito com os pseudo-componentes. Cornell (2002) recomenda que deve-se fazer a transformação inversa dos pseudo-componentes para a escala original, isolando x_i na equação (2.3). Assim, a relação linear entre a escala original e pseudo-componentes é explicada em (2.4).

$$x_i = L_i + (1 - L)\, x_i'. \tag{2.4}$$

2.1.1 Descrição do experimento de mistura utilizado

O presente estudo fará uso dos dados disponibilizados por Akay e Tez (2011). Os dados de mistura são referentes a ocorrência de tumor em ratos. O experimento de mixtura foi conduzido para estudar os efeitos de calorias da dieta composta por gordura, carboidrato e fibra sobre a expressão (promoção) de tumor nas glândulas mamárias induzido por *Dimetil-benzathracene* (DMBA) em ratos fêmeas. Nesse experimento, ratos da raça Sprague-Dwaley de 38 a 42 dias de idade foram aleatoriamente atribuídos a nove grupos, sendo 30 animais por grupo. Ao dia 52, os animais foram administrados com 7,5 mg de DMBA em óleo de milho por um tubo ao estômago. Uma semana depois do tratamento com DMBA, os animais em cada grupo foram alimentados com suas correspondentes dietas em igual quantidade de calorias totais, mas com diferentes níveis de gordura, carboidrato e fibra. As dietas foram administradas durante 26 semanas, tempo total de duração do experimento.

Os nutrientes gordura, carboidrato e fibra foram administrados sobre todo o alcance das dietas que poderiam ser consideradas como sendo fisiológicas. Por exemplo, uma dieta contendo zero ou cem por cento de qualquer um dos nutrientes poderia ser considerada como sendo não-fisiológica (não-nutritiva) e produziria resultados sem relevância quando comparados com dietas usuais. Essas dietas tiveram uma combinação de baixo, médio e alto níveis de cada um dos três nutrientes, num total de nove grupos de teste.

A proporção das três dietas foram restritas pelos seguintes limites inferiores e superiores

$$0,133 \leq Gordura \leq 0,730$$

$$0,267 \leq Carboidrato \leq 0,864$$

$$0,003 \leq Fibra \leq 0,600.$$

A Tabela 1 contém as respostas das proporções de tumor observadas de nove grupos de dieta com diferentes proporções calóricas de gordura (x_1), carboidrato (x_2) e fibra (x_3). Na Tabela 1, as dietas 1 a 3 são constituídas de baixa gordura e alto carboidrato, as dietas 4 a 6 são constituídas de valores médios de gordura e carboidrato, as dietas 7 a 9 são possuem alta gordura e baixo carboidrato. As dietas 1, 4 e 7 são altas em fibras, as dietas 2, 5 e 8 são médias em fibras e as dietas 3, 6 e 9 são baixas em fibras. Em todas as dietas, a gordura e o carboidrato são as duas maiores origens de calorias.

Tabela 1 O número observado de tumores induzidos DMBA em glândulas mamárias de ratos tratados com diferentes proporções calóricas de fibra, gordura e carboidrato (componentes originais). Cada grupo é composto por 30 ratos e tem igual quantidade de calorias totais.

Grupo	Componentes Originais			Ratos c/ Tumor	Taxa de Tumor
	Gordura (x_1)	Carboidrato (x_2)	Fibra (x_3)		
1	0,175	0,775	0,050	17	0,567
2	0,153	0,820	0,027	15	0,500
3	0,133	0,863	0,004	17	0,567
4	0,491	0,470	0,039	24	0,800
5	0,440	0,538	0,022	21	0,700
6	0,390	0,607	0,003	23	0,767
7	0,701	0,267	0,032	18	0,600
8	0,638	0,343	0,019	23	0,767
9	0,576	0,421	0,003	26	0,867

A Figura 8 exibe os pontos experimentais na região simplex restrita. A região simplex em termos dos pseudo-componentes pode ser facilmente visualizada a partir da Figura 8.

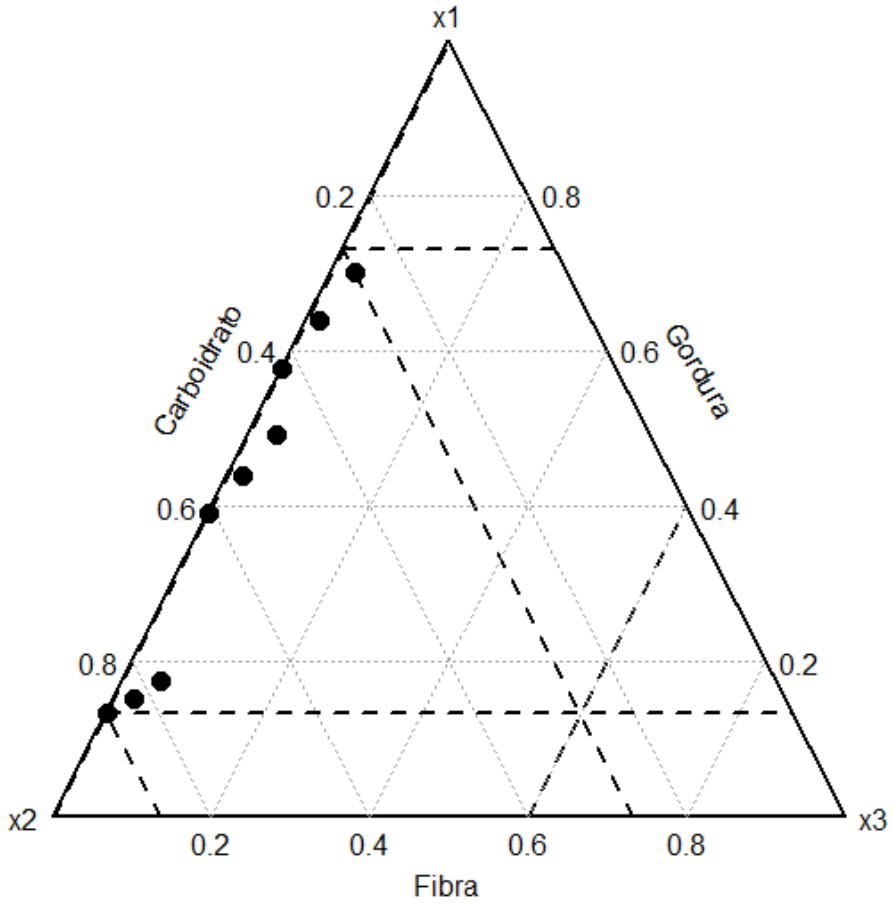


Figura 8 Região simplex dos grupos de ratos tratados com diferentes proporções calóricas de fibra, gordura e carboidrato (componentes originais).

A Tabela 2 reúne os dados das nove dietas apresentadas na Tabela 1 em pseudo componentes.

Tabela 2 O número observado de tumores induzidos DMBA em glândulas mamárias de ratos tratados com diferentes proporções calóricas de fibra, gordura e carboidrato (pseudo componentes). Cada grupo tem igual quantidade de calorias totais.

Grupo	Pseudo componentes			Ratos c/ Tumor	Taxa de Tumor
	x'_1	x'_2	x'_3		
1	0,070	0,851	0,079	17	0,567
2	0,034	0,926	0,040	15	0,500
3	0,000	1,000	0,000	17	0,567
4	0,600	0,340	0,060	24	0,800
5	0,514	0,454	0,032	21	0,700
6	0,430	0,570	0,000	23	0,767
7	0,951	0,000	0,049	18	0,600
8	0,846	0,127	0,027	23	0,767
9	0,742	0,258	0,000	26	0,867

2.1.2 Modelos de regressão aplicados em experimentos de mistura

Além das pressuposições estatísticas relacionadas à distribuição dos resíduos, o modelo de regressão linear considera que as covariáveis x_i $(i = 1, ..., q)$ sejam independentes. Tal suposição não pode ser estendida dos modelos que envolvem misturas, uma vez que $\sum_{i=1}^{q} x_i = 1$, o que implica que, por exemplo para o componente x_1, $x_1 = 1 - x_2 - x_3 - \ldots - x_q$.

Devido a esse resultado, surge uma nova classe de modelos definidos por regressão de mistura, cuja essência consiste em confundir a restrição $\sum_{i=1}^{q} x_i = 1$ aos parâmetros do modelo dado por uma função polinomial.

Visando a melhores esclarecimentos, será exemplificada uma mistura formada por três componentes dada por

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3,$$

restrito à $\sum_{i=1}^{3} x_i = 1$. O confundimento da restrição $\sum_{i=1}^{3} x_i = 1$ aos parâmetros é feito em relação ao intercepto, com as operações descritas a seguir,

$$\begin{aligned} y &= 1\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 \\ &= (x_1 + x_2 + x_3)\,\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 \\ &= \beta_0 x_1 + \beta_0 x_2 + \beta_0 x_3 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 \\ &= (\beta_0 + \beta_1)\,x_1 + (\beta_0 + \beta_2)\,x_2 + (\beta_0 + \beta_3)\,x_3 \\ y &= \beta_1^* x_1 + \beta_2^* x_2 + \beta_3^* x_3. \end{aligned}$$

Logo, o modelo de mistura é reescrito por

$$y = \beta_1^* x_1 + \beta_2^* x_2 + \beta_3^* x_3 \Rightarrow y = \sum_{i=1}^{3} \beta_i^* x_i.$$

Observa-se, portanto, que num modelo linear o parâmetro $\beta_i^* = \beta_0 + \beta_i$, ou seja, o intercepto está confundido com o parâmetro do componente de mistura i. Por questão de padronização, β_i^* será denotado por β_i apenas, uma vez que todos os modelos abordados envolverão componentes de mistura.

Seguindo o mesmo procedimento, porém expandindo as operações algébricas com as devidas simplificações, os modelos de mistura com termos quadráticos e cúbicos podem ser propostos. Contudo, convém ressaltar que os modelos quadrático e cúbico não são caracterizados pelas potências x_2 e x_3, mas pelo número de componentes envolvidos. Por exemplo, ainda considerando $q = 3$, para o modelo de mistura linear existem três parâmetros a serem estimados e no modelo de mistura quadrático existem 6 parâmetros a serem estimados (Tabela 3). A de-

dução do modelo de mistura quadrático para esse caso pode ser feita de maneira similar ao modelo de mistura linear, ou seja,

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_1^2 + \beta_5 x_2^2 + \beta_6 x_3^2.$$

Aplicando a restrição e observando que $x_1^2 = x_1 x_1 = x_1 (1 - x_2 - x_3) = x_1 - x_1 x_2 - x_1 x_3$, implicando $x_1 = x_1^2 + x_1 x_2 + x_1 x_3$, temos o modelo de mistura quadrático,

$$y = \beta_1^* x_1 + \beta_2^* x_2 + \beta_3^* x_3 + \beta_{12}^* x_1 x_2 + \beta_{13}^* x_1 x_3 + \beta_{23}^* x_2 x_3.$$

Portanto, o estudo do efeito de interação entre os componentes só é feito a partir do modelo de mistura quadrático, uma vez que os termos quadráticos são substituídos pelos termos cruzados. A Tabela 3 apresenta de maneira geral o número de parâmetros a serem estimados nos principais modelos de mistura.

Tabela 3 Número de parâmetros (p) no modelo em função do nº de componentes de mistura (q).

Nº componentes	**Linear**	**Quadrático**	**Cúbico especial**	**Cúbico completo**
q	q	$\frac{q(q+1)}{2}$	$\frac{q\left(q^2+5\right)}{6}$	$\frac{q(q+1)(q+2)}{6}$

A proposta de experimentos de mistura é construir um modelo apropriado que relacione a resposta aos componentes $x_1, x_2, \ldots, x_q$. Assume-se que a resposta de interesse η seja uma função das variáveis de mistura x_i, ou seja,

$$\eta = f\left(x_1, x_2, \ldots, x_q\right). \tag{2.5}$$

Quando um experimento é realizado, é natural assumir que as respostas observadas, que denotaremos por y_i para o i-ésimo valor $(i = 1, 2, \ldots, n)$ variam sobre a média de η_i com uma variância constante σ^2 para todo $i = 1, 2, \ldots, n$. A resposta observada contém o erro experimental aditivo ε_i, ou seja,

$$y_i = \eta_i + \varepsilon_i. \tag{2.6}$$

Assume-se que o erro experimental na equação 2.6 tem distribuição normal e identicamente distribuído com média zero e variância σ^2 constante, ou, abreviadamente $\varepsilon_i \overset{iid}{\sim} N\left(0; \sigma^2\right)$. A forma funcional da resposta $E[y] = f(x_1, \ldots, x_q)$ geralmente não é conhecida. Em várias situações, modelos de aproximação polinomial de primeiro e segundo graus podem ser utilizados.

As formas de modelos de mistura mais utilizadas são os polinômios canônicos de segundo grau introduzidos por Scheffé (1958) da forma

$$E[Y] = \eta = \sum_{i=1}^{q} \hat{\beta}_i x_i + \sum_{i=1}^{q} \sum_{i<j}^{q} \hat{\beta}_{ij} x_i x_j. \tag{2.7}$$

Contudo, existem situações em que os polinômios canônicos de Scheffé não são os melhores modelos. Por exemplo, modelos com termos inversos são utilizados com o objetivo de modelar mudanças extremas no comportamento da resposta, conforme o valor de um ou mais componentes tendem ao limite da região simplex. O seguinte modelo de primeiro grau que inclui um termo inverso foi proposto por Draper e St John (1977),

$$E[Y] = \sum_{i=1}^{q} \hat{\beta}_i x_i + \sum_{i=1}^{q} \hat{\beta}_{-i} x_i^{-1}, \tag{2.8}$$

em que β_{-i} representa o parâmetro do i-ésimo termo inverso x.

Em muitas situações práticas, os modelos polinomiais de Scheffé e outros

modelos especiais de mistura fornecem estimativas ruins dos coeficientes, devido ao fato de que as restrições dadas pelas equações (2.1) e (2.2) sobre a região da mistura provocam colinearidade. Existem algumas alternativas para reduzir os feitos decorridos da colinearidade, como por exemplo, remover as restrições em (2.1) entre os componentes de mistura.

Uma maneira de remover a soma constante entre os termos lineares da mistura em um modelo de mistura é definir razões das proporções dos componentes e usar essas razões como variáveis em um modelo polinomial padrão. Em alguns experimentos, a razão das proporções dos componentes é mais interpretável do que a proporção deles mesmos. O modelo de razão apresenta uma interessante alternativa aos modelos de Scheffé e Backer, em que os modelos de razão descrevem um diferente tipo de curvatura (SNEE, 1973). Em particular, definindo a razão entre os componentes de mistura como variável ajuda a obter melhores resultados para um sistema de mistura. Adicionalmente, os modelos com termos inversos definidos pela equação (2.8) podem ser usados como um modelo razão (DRAPER; ST JOHN, 1977). Razões do tipo:

$$w_i = \frac{x_i}{x_q^*};\ i = 1,\ 2,\ \ldots,\ q-1, \tag{2.9}$$

em que x_q^* corresponde ao componente de mistura que provoca o chamado *efeito de borda*, podem ser usados como exemplos.

O *efeito de borda* é caracterizado em situações em que o comportamento da resposta sofre bruscas alterações quando um (ou mais) dos componentes de mistura envolvidos atinge proporções próximas da fronteira da região simplex, ou seja, esse componente tem valores próximos de zero. Draper e John (1977) apresentam um experimento que consistiu em obter a melhor formulação de quatro ingredientes na luminosidade de labaredas. Para tal foi instalado um delineamento

em Extreme Vértice e foram utilizados os componentes magnésio, nitrato de sódio, nitrato de estôncio e *binder*. Esse último componente assume proporções entre 0.03 e 0.08. Os modelos de Scheffé e os modelos com variáveis razão foram utilizados e o que proporcionou melhor ajuste foi o modelo com variáveis razão. Ainda, a inclusão de variáveis com comportamento não linear, como é o caso das variáveis razão, pode contemplar possíveis efeitos curvilíneos da resposta quando um dos componentes apresenta proporção perto de zero.

Retornando à equação (2.9), devem existir $q-1$ razões no conjunto. Também, cada razão deve conter pelo menos um dos componentes utilizados em, pelo menos, uma das outras razões pertencente ao conjunto. Dessa forma, considerando $q = 3$ e a restrição de mistura, o modelo linear com variáveis w é dado por

$$y = \beta_1 w_1 + \beta_2 w_2 + \beta_3 w_3,$$

sendo $w_1 = \frac{x_1}{x_3}$, $w_2 = \frac{x_2}{x_3}$ e $w_3 = \frac{x_3}{x_3} = 1$. Logo o modelo é reescrito como

$$y = \beta_3 + \beta_1 w_1 + \beta_2 w_2.$$

Algumas alternativas para modelos razão foram propostas, por exemplo, por Aitchison e Bacon-Shone (1984), que sugeriram usar $q-1$ variáveis razões na escala logarítmica, ou seja,

$$w_i = \log\left(\frac{x_i}{x_q^*}\right);\ i = 1,\ 2,\ \ldots,\ q-1. \tag{2.10}$$

Da mesma forma, como uma alternativa às transformações dadas nas equações (2.9) e (2.10), Akay e Tez (2011) propuseram o modelo com variáveis razão utilizando a transformação raiz quadrada, ou seja,

$$w_i = \sqrt{\frac{x_i}{x_q^*}};\ i = 1,\ 2,\ \ldots,\ q-1. \tag{2.11}$$

O comportamento das transformações (2.9) a (2.11) é ilustrado na Figura 9 a seguir para um sistema de mistura de dois componentes. Então, considerando um sistema onde exista efeito de borda, é preferível o modelo que capte esse tipo de efeito, como pode ser visto na Figura 9. Nessa mesma figura, observe que, se fosse utilizado o modelo linear de Scheffé, como o efeito de x_1 é linear, ocorreria uma subestimação para os valores de y.

Na Figura 9, as transformações nas equações (2.9) e (2.11) apresentam o comportamento semelhante quando $x_1 \leq 0,5$, enquanto que as transformações nas equações (2.10) e (2.11) têm o comportamento semelhante para $x_1 \geq 0,5$. Em outras palavras, a transformação (2.11) atua como um balanço comparado às equações (2.9) e (2.10). Pela Figura 9 pode-se fazer ainda uma analogia das variáveis razão com uma variável do modelo de Scheffé, observando que a curva que, a grosso modo, mais se assemelha com a curva formada pelo componente x_1 é a (2.11) considerando um sistema bem comportado. Dessa forma, utilizando-se um modelo de mistura com variáveis razão, é desejável aquele cujas variáveis tenham comportamento o mais próximo possível da variável não transformada.

Dessa forma, em extensão ao modelo apresentado em (2.8), os modelos polinomiais de variáveis razão de primeiro e segundo grau para misturas são dados por

$$E\left[Y\right] = \hat{\beta}_0 + \sum_{i=1}^{q-1} \hat{\beta}_i w_i \tag{2.12}$$

$$E\left[Y\right] = \hat{\beta}_0 + \sum_{i=1}^{q-1} \hat{\beta}_i w_i + \sum_{i=1}^{q-1} \sum_{i<j}^{q-1} \hat{\beta}_{ij} w_i w_j.$$

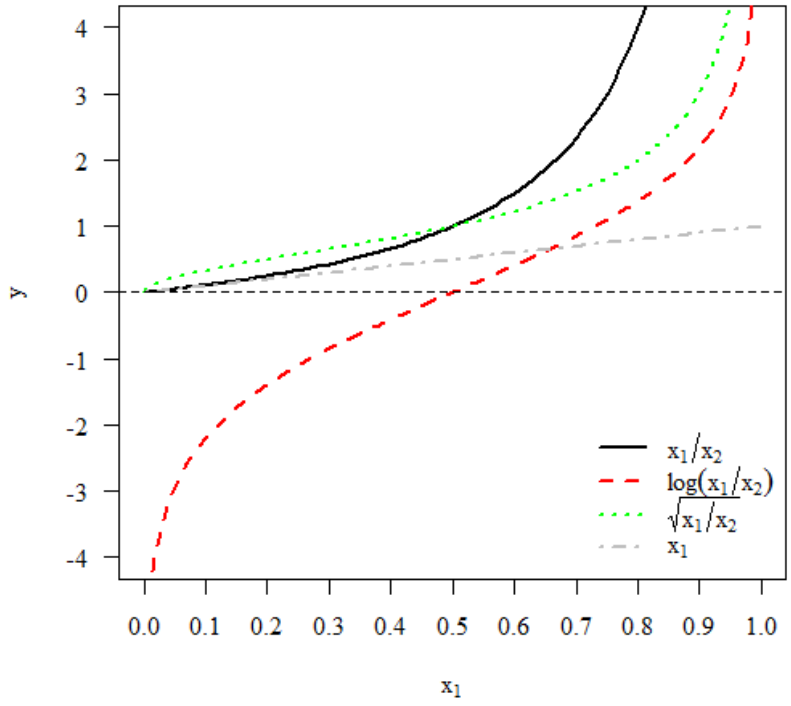


Figura 9 O comportamento das transformações para um sistema de mistura de dois componentes.

Uma observação que deve ser feita é sobre o fato de que as correlações entre os termos $w_i's$ serão menores do que as correlações entre os componentes de mistura x_j, uma vez que um termo constante é incluído no modelo. Esse fato pode ser observado na matriz de variâncias e covariâncias das estimativas dos parâmetros do modelo ajustado.

A transformação em pseudo-componentes não pode ser usada com modelos razão, pois pode existir pelo menos um componente de mistura igual a um

(SNEE, 1973). Portanto, com o objetivo de preservar o desempenho nas predições dos modelos quanto ao efeito dos componentes e permitir as comparações em escala comum, apenas preditores lineares com componentes originais devem ser utilizados.

2.2 Introdução ao método de Boosting

O método conhecido como Boosting foi inicialmente proposto por Shapire em 1990 com o propósito de solucionar problemas de classificação que envolviam muitas variáveis e observações, o que tipicamente ocorria na comunidade de aprendizado de máquinas. Dentro dessa comunidade, foi proposto um problema teórico chamado de problema de Boosting, que pode ser informalmente exposto da seguinte maneira:

> Seja Ω o espaço dos eventos da amostra (ou o espaço p dimensional contendo todos os possíveis vetores $\mathbf{x}$, com $\mathbf{x} = (\mathbf{x_1}, \mathbf{x_2}, \ldots, \mathbf{x_p})$ e p o número de variáveis medidas em cada objeto) e suponha que existe um método de classificação que é ligeiramente melhor do que uma escolha aleatória, para qualquer distribuição em Ω. Esse método é chamado de classificador fraco (*weak learner*). A existência de um classificador fraco implica a existência de um classificador forte (*strong learner*), com erro pequeno sobre todo o espaço Ω? (SHAPIRE, 1990, p. 199)

Esse problema foi resolvido por Schapire (1990), que mostrou que era possível obter um classificador forte a partir de um fraco. A partir de então, foram desenvolvidos vários algoritmos dentro do contexto de Boosting. Um dos mais recentes e bem sucedidos deles é o algoritmo conhecido como AdaBoost (Adaptative Boosting), que funciona perturbando a amostra de treinamento gerando a cada iteração (de forma determinística, mas adaptativa) uma distribuição sobre as observações da amostra, dando maior peso (maior probabilidade de pertencer à

amostra perturbada) às observações classificadas erroneamente no passo anterior. Existe um outro método de combinação de preditores, conhecido por *Bagging* (*Bootstrap Aggregating*), que funciona perturbando essa amostra de treinamento aleatoriamente por meio de reamostragem, gerando a cada iteração um classificador e o classificador final é obtido pela agregação desses classificadores (SHAPIRE; FREUND, 2012).

Desde o seu desenvolvimento como uma resposta a um problema teórico, os algoritmos do tipo Boosting têm recebido grande atenção, tanto na comunidade estatística quanto na de *machine learning*. A comunidade estatística busca entender como e por que o algoritmo Boosting funciona, abordando aspectos como consistência, enquanto na comunidade de *machine learning* a abordagem é mais focada nos próprios algoritmos e em sua funcionalidade (RUBESAM, 2004).

Jiang (2000) mostrou que o algoritmo AdaBoost é consistente, no sentido de que, durante o treinamento, ele gera uma sequência de classificadores com erro que converge para o erro do classificador de Bayes. Se for possível construir um classificador de Bayes, conforme o mesmo autor refere, pode-se conhecer o melhor classificador em termos do risco de Bayes. Contudo, na prática, a densidade de $\mathbf{X}$ dado k (a classe à qual o objeto pertence) e a densidade de $Y = k$ são desconhecidas, o que inviabiliza a obtenção do classificador de Bayes. Por outro lado, o classificador de Bayes pode fornecer, via simulação, o classificador que mais se aproxima do seu erro, uma vez que as densidades desconhecidas podem ser pressupostas.

Freund e Schapire (1996) apresentaram o primeiro algoritmo no contexto de Boosting, chamado de AdaBoost. Os autores fizeram uma análise do algoritmo em termos de limites para as probabilidades de erro na amostra de treinamento e nas amostras de teste (o erro de um classificador em casos novos é chamado na

literatura de aprendizagem de máquinas de erro de generalização). Um dos limites teóricos mostrados implica que o erro na amostra de treinamento decai exponencialmente com o número de iterações do algoritmo. Empiricamente, observa-se que, após algumas iterações, o erro na amostra de treinamento tende a zero, confirmando o resultado teórico.

Inicialmente, observou-se que, quando se continua a executar o algoritmo AdaBoost, o erro na amostra teste continua a decrescer, indicando que o algoritmo é resistente a super ajuste (*overftting*) (BUHLMANN; HOTHORN, 2007). O super ajuste é o problema que surge quando um modelo tem desempenho bom no conjunto de treinamento, mas em dados novos, que não foram usados no ajuste do modelo, tem desempenho ruim. Isso ocorre geralmente porque o modelo se torna complexo demais (ou seja, número excessivo de parâmetros) e passa a ajustar peculiaridades do conjunto de treinamento (BISHOP, 1995). Por exemplo, em regressão logística, a adição de variáveis sempre melhora o desempenho no conjunto usado para ajustar o modelo, mas em algum ponto isso começa a se tornar prejudicial e o desempenho em um conjunto de teste é ruim. Em redes neurais, se o algoritmo de otimização é executado indefinidamente, o erro sempre diminui no conjunto de treinamento, mas em certo ponto ele começa a aumentar no conjunto de teste. Existem métodos para determinar o ponto de parada nesse caso, como por exemplo o método de parada precoce (*early stopping*), que cessa a otimização quando o erro começa a aumentar no conjunto de teste (SCHAPIRE; FREUND, 2012).

Friedman, Hastie e Tibshirani (2001) mudaram totalmente o modo como Boosting é visto, pelo menos na comunidade estatística. Eles colocaram Boosting como uma aproximação do ajuste de um modelo aditivo na escala logística, usando a máxima verossimilhança da Bernoulli como critério. Ademais, sugeriram uma

aproximação mais direta, o que levou ao algoritmo LogitBoost, um algoritmo para ajustar regressão logística aditiva que dá resultados praticamente idênticos ao AdaBoost de Freund e Schapire.

2.2.1 Algoritmos Boosting utilizados na classificação binária

Serão apresentados a seguir dois algoritmos Boosting. O algoritmo AdaBoost não será utilizado neste trabalho, porém é necessária sua apresentação por ser um algoritmo precedente de outros algoritmos Boosting, inclusive o algoritmo utilizado neste trabalho, o algoritmo Gradiente Boosting de Friedman.

2.2.1.1 AdaBoost para duas classes

Será apresentada a seguir a versão do algoritmo AdaBoost para classificação binária, dada em Friedman, Hastie e Tibshirani (2001).

Suponha que temos um conjunto de treinamento $(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n)$, em que n é o tamanho do conjunto (amostra) de treinamento e y_i $(i = 1, \ldots, n)$ pertence a uma de duas possíveis classes, rotuladas por $\{-1, 1\}$. Defina $F(\boldsymbol{x}) = \sum_{m=1}^{M} c_m f_m(\boldsymbol{x})$, onde M é o número de vezes que o algoritmo é executado (iterações). O classificador base (passo 2(a) do algoritmo AdaBoost) retorna valores em $\{-1, 1\}$ e pode ser, por exemplo, uma árvore de regressão ou uma rede neural. Os valores c_m são constantes e a predição correspondente a cada valor de $\boldsymbol{x}$ é a função sinal de $F(\boldsymbol{x})$, ou seja, $sign(F(\boldsymbol{x}))$. A função $sign(\cdot)$ retorna 1 se $sign(\cdot) > 0$ e retorna -1 se $sign(\cdot) < 0$. O algoritmo AdaBoost ajusta classificadores base f_m em amostras ponderadas do conjunto de treinamento, dando maior peso, ou ponderação, aos casos que são classificados erroneamente, como pode ser

visto no passo 2(c). Os pesos são ajustados adaptativamente em cada iteração e o classificador final é uma combinação linear dos classificadores f_m.

A Figura 10 ilustra de maneira geral o funcionamento de um algoritmo Boosting para classificação binária.

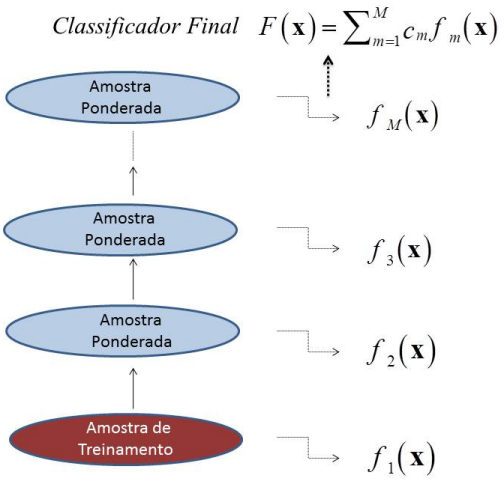


Figura 10 Algoritmo Boosting para classificação binária

O algoritmo AdaBoost pode ser esquematizado em três passos:

1. Dado $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)$ em que $\mathbf{x_i} \in \mathbf{X}$ e $y_i \in \{-1, +1\}$. Inicialize os pesos $W_i^m = 1/n$, $i = 1, 2, \ldots, n$.

2. Repita para $m = 1, 2, \ldots, M$:

   (a) Ajuste o classificador $f_m(\boldsymbol{x}) \in \{-1, 1\}$ usando os pesos W_i e os dados de treinamento;

   (b) Calcule:

$$\varepsilon_m = \frac{\sum_{i=1}^{n} W_i^m I\left[Y_i \neq f_m\left(X_i\right)\right]}{\sum_{i=1}^{n} W_i^m},$$

$$c_m = \frac{1}{2} \ln\left(\frac{1-\varepsilon_m}{\varepsilon_m}\right).$$

(c) Faça:

$$W_i^{m+1} = \frac{W_i^m}{Z^m} \times \begin{cases} e^{-c_m} \, se \; y_i = f_m\left(x_i\right) \\ e^{c_m} \, se \; y_i \neq f_m\left(x_i\right) \end{cases}$$

em que, Z^m é um fator de normalização,

$$Z^m = \sum\nolimits_{i=1}^{n} W_i^m e^{\left[-c_m \, y_i \, f_m(x_i)\right]}.$$

3. A predição é dada por $sign\left[F\left(\boldsymbol{x}\right)\right] = sign\left[\sum_{m=1}^{M} c_m \, f_m\left(\boldsymbol{x}\right)\right]$.

No algoritmo acima, ε_m representa a média ponderada dos erros com pesos $W = (W_i, \ldots, W_n)$. Em cada iteração, o algoritmo aumenta os pesos W_i das observações classificadas erroneamente por um fator que depende dos erros ε_m das observações do conjunto de treinamento (passo 2 (c)).

Em Liska (2012) é apresentada uma ilustração didática para melhor compreensão do algoritmo AdaBoost.

Friedman, Hastie e Tibshirani (2000) mostraram que o algoritmo AdaBoost pode ser derivado como algoritmo iterativo para ajustar um modelo aditivo logístico, otimizando um critério que até segunda ordem é equivalente à log-verossimilhança da binomial.

2.2.1.2 Algoritmo Gradiente Boosting de Friedman

Breiman (1998, 1999) mostrou que o algoritmo AdaBoost pode ser representado como um algoritmo do gradiente no espaço funcional, o qual pode-se denominar de Gradiente de Descida Funcional (FGD). Friedman, Hastie e Tibshirani (2000) e Friedman (2001) desenvolveram de forma mais geral uma estrutura estatística que leva a direta interpretação de Boosting como um método que pode ser utilizado para ajustar modelos de regressão. Na sua terminologia, trata-se de uma aproximação em modelagem aditiva *stagewise*. O termo *stagewise* em modelos de regressão está relacionado com processo de seleção de variáveis realizado em estágios. No contexto do algoritmo FGD, o termo carrega essa característica, mas também o fato de que o algoritmo adiciona contribuições individuais de cada parâmetro do modelo a cada iteração do algoritmo, ou seja, não necessariamente é selecionada uma variável a cada iteração do algoritmo e a mesma variável pode ser selecionada mais de uma vez em diferentes iterações do algoritmo.

No contexto de Boosting, o objetivo é estimar uma função de predição ótima $f^{*}(\cdot)$, também chamada de minimizador populacional, que é definido por

$$f^{*}(\cdot) = \arg\min_{g} E_{Y,X}\left[\rho\left(Y, g\left(\boldsymbol{X}\right)\right)\right] \tag{2.13}$$

em que, $\rho(\cdot,\ \cdot)$ é uma função perda que é assumida como sendo diferenciável com respeito a $g(\cdot)$. Na prática, trabalhamos com realizações $(y_i, \boldsymbol{x}_i^T)$, $i = 1, \ldots, n$, de $(\boldsymbol{y}, \boldsymbol{x}^T)$, e, por esse motivo, a esperança em (2.13) é desconhecida. Por essa razão, em vez de minimizar a perda esperada dado em (2.13), os algoritmos Boosting minimizam a perda média observada ou risco empírico, que é dada por

$$R\left[y, g\left(\mathbf{x}\right)\right] = \frac{1}{n}\sum_{i=1}^{n}\rho\left[y_i, g\left(\mathbf{x}_i\right)\right]. \tag{2.14}$$

Pode-se mostrar que minimizar o risco empírico, ou seja, a perda média observada, é equivalente a minimizar a expressão em (2.13). Para mostrar esse fato, será necessário fazer uma analogia com o estimador de máxima verossimilhança (MLE). Para tal, seja a função de verossimilhança $L\left(\theta|\, x_1, \ldots, x_n\right) = \prod_{i=1}^{n} f\left(x_i|\, \theta\right)$ e o logaritmo da função de verossimilhança

$$l\left(\theta\right) = \ln\left[L\left(\theta|\, x_1, \ldots, x_n\right)\right] = \sum_{i=1}^{n}\ln\left[f\left(x_i|\, \theta\right)\right].$$

O estimador de máxima verossimilhança é o valor θ que maximiza $L\left(\theta|\, x_1, \ldots, x_n\right)$. Sabe-se que o valor de θ que maximiza $L\left(\theta|\, x_1, \ldots, x_n\right)$ é o mesmo que maximiza $l\left(\theta\right)$, ou seja,

$$\hat{\theta}_{MLE} = \arg\max_{\theta} L\left(\theta|\, x_1, \ldots, x_n\right) = \arg\max_{\theta} l\left(\theta\right).$$

Um problema de maximização pode ser convertido em um problema de minimização utilizando-se da seguinte relação de equivalência,

$$\arg\max_{x} f\left(x\right) = \arg\min_{x} f\left(-x\right),$$

considerando $f\left(x\right) = x$. Dessa forma, temos que

$$\arg\max_{\theta}\sum_{i=1}^{n}\ln\left[f\left(x_i|\, \theta\right)\right] = \arg\min_{\theta} -\sum_{i=1}^{n}\ln\left[f\left(x_i|\, \theta\right)\right] = \hat{\theta}_{MLE}. \tag{2.15}$$

Agora, considere a média da quantidade em (2.15):

$$\arg\min_{\theta} \ \frac{1}{n}\sum_{i=1}^{n} -\ln\left[f\left(x_i|\,\theta\right)\right], \tag{2.16}$$

cujo mínimo ainda é $\hat{\theta}_{MLE}$, uma vez que a multiplicação de um escalar não influencia no problema de otimização. Pela Lei Forte dos Grandes Números (CASELLA; BERGER, 2011), a quantidade em (2.16) converge para a esperança

$$\arg\min_{\theta} \ E\left[-\ln\left[f\left(x|\,\theta\right)\right]\right]. \tag{2.17}$$

Ou seja, a expressão (2.16) converge para a mesma solução em (2.17) com alta probabilidade conforme $n \to \infty$. Isto quer dizer que o risco empírico em (2.16) converge para risco teórico em (2.17).

Para ilustrar esse fato, considere o seguinte exemplo organizado nos seguintes passos:

(i) Considere uma população cujos elementos tem distribuição Normal com média $\theta = 2$ e variância $\sigma^2 = 1$;

(ii) Considere uma sequência de tamanhos amostrais n_i, com $i = 1, 2, \ldots, 5000$. Para cada n_i, será gerada uma amostra de tamanho i de uma distribuição $N\left(2;\,1\right)$;

(iii) Considere uma sequência de parâmetros θ_j, no caso consideraremos $\theta \in (-10, 10)$. Considere o negativo do logaritmo neperiano da função de verossimilhança de n_i observações da distribuição Normal com parâmetros θ_j (desconhecido) e $\sigma^2 = 1$,

$$-l\left(\theta_j\right) = -\sum_{i=1}^{n} \ln\left[f\left(x_i|\,\theta_j\right)\right] = -\sum_{i=1}^{n} \ln\left[\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x_i-\theta_j)^2}\right];$$

(iv) Para cada θ_j no passo (iii), calcule $-l(\theta_j)$ utilizando o tamanho amostral n_i;

(v) Dos valores encontrados em (iv), selecione o menor dos $-l(\theta_j)$ e divida-o pelo tamanho amostral n_i. Esse é o risco empírico para o tamanho amostral n_i.

Do esquema acima, teremos uma sequência n_i, com $i = 1, 2, \ldots, 5000$, de riscos empíricos, conforme pode ser visto na Figura 11.

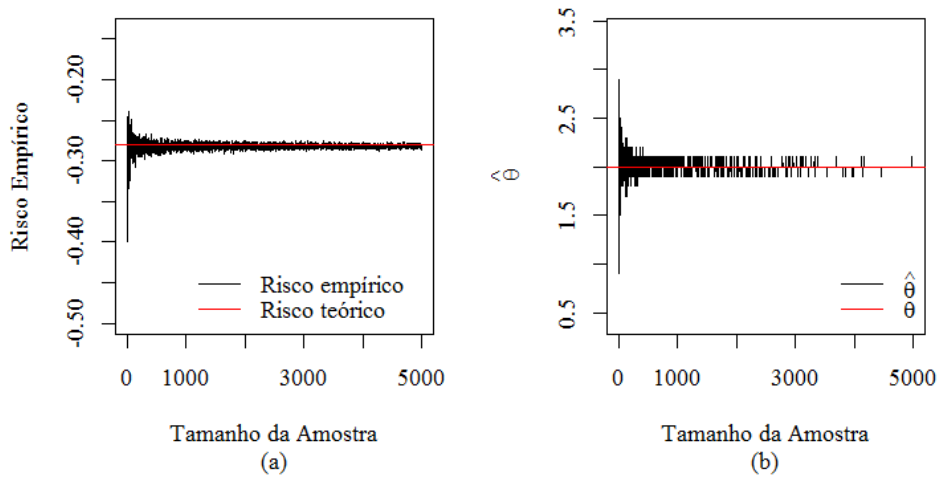


Figura 11 Simulação da convergência do risco empírico para o risco teórico conforme o tamanho amostral cresce (a) e convergência de $\hat{\theta}$ para θ conforme o tamanho amostral cresce (b).

Ainda, do exemplo acima, como trata-se de uma simulação, o valor esperado em (2.17) é conhecido e, no caso, é dado por -0.28, ou seja, o valor $\hat{\theta}$ que minimiza o risco teórico é $\hat{\theta} = 2$, como esperado uma vez que os elementos da população tem distribuição Normal com média $\theta = 2$ e variância $\sigma^2 = 1$. A Figura 12 ilustra esse fato mostrando que $\theta = 2$ é o valor de θ_j que minimiza a

quantidade $-l\left(\theta\right)$.

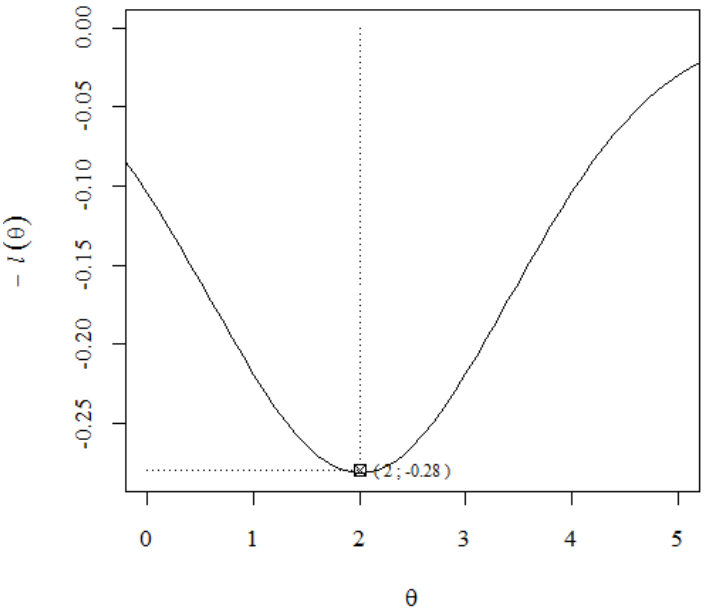


Figura 12 Risco teórico minimizado considerando a minimização do negativo do logaritmo neperiano da função de verossimilhança de n_i observações da distribuição Normal com parâmetros θ_j (desconhecido) e $\sigma^2 = 1$.

Portanto, do exemplo acima, mostramos de maneira visual que o risco empírico converge para o risco teórico.

Retornando à expressão 2.13, a minimização é feita em relação à $g\left(\cdot\right)$ e veremos mais adiante que $g\left(\cdot\right)$ é um procedimento base no caso de ajuste de modelos. O termo "procedimento base" vem da comunidade de aprendizado de máquinas e, no presente estudo, será simplesmente uma função linear (um polinômio) que relaciona as variáveis independentes à resposta. Mais ainda, no caso

de modelos lineares generalizados, o procedimento base é a função relacionada à componente sistemática do modelo.

Por exemplo, considerando a perda *erro quadrática* dada por,

$$\rho\left(y,\, g\right) = \left(y - g\right)^2,$$

e minimizando-a conforme 2.13, tem como minimizador populacional

$$f^*\left(\boldsymbol{x}\right) = E\left[\mathbf{Y}\,|\boldsymbol{X} = \boldsymbol{x}\right].$$

De maneira geral, dada uma função perda $\rho\left(y,\, g\right)$ e um procedimento base, o algoritmo a seguir foi proposto por Friedman (2001), também chamado de Algoritmo Gradiente Boosting de Friedman (Algoritmo GBF), e consiste dos seguintes passos:

(i) Inicialize $\hat{f}^{(0)}\left(\cdot\right)$ com um valor inicial. Escolhas comuns são

$$\hat{f}^{(0)}\left(\cdot\right) = \arg\min_{c} \frac{1}{n}\sum_{i=1}^{N} \rho\left(Y_i,\, c\right) \tag{2.18}$$

ou $\hat{f}^{(0)}\left(\cdot\right) = 0$. Na expressão (2.18) o argumento c é o valor inicial da função perda que minimiza $\hat{f}^{(0)}\left(\cdot\right)$. Por exemplo, para a *perda quadrática*, o valor de $\hat{f}^{(0)}\left(\cdot\right)$ que satisfaz a equação (2.18) é a média da variável resposta, ou seja $\bar{\boldsymbol{y}}$. Coloque $m = 0$.

(ii) Aumente m em 1. Calcule o gradiente negativo $-\frac{\partial}{\partial g}\rho\left(Y,\, g\right)$ e calcule em $\hat{f}^{(m-1)}\left(\boldsymbol{X}_i\right)$:

$$\mathbf{u}^{(m)} = \left(u_i^{(m)}\right)\Big|_{=1,\ldots,n} = -\frac{\partial}{\partial g\left(\boldsymbol{x}_i\right)}\rho\left(Y_i,\, g\left(\boldsymbol{x}_i\right)\right)|_{g(\boldsymbol{x}_i)=\hat{f}^{(m-1)}(\boldsymbol{x}_i)}$$

(iii) Ajuste o vetor gradiente negativo $u_1, \ldots, u_n$ para $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ por um procedimento base $\hat{g}^{(m)}(\cdot)$.

(iv) Atualize

$$\hat{f}^{(m)}(\cdot) = \hat{f}^{(m-1)}(\cdot) + v \cdot \hat{g}^{(m)}(\cdot)$$

em que $0 < v \leq 1$ é o fator *comprimento do passo*.

(v) Continue o processo de iteração entre os passos (ii) a (iv) até $m = M$, para alguma iteração de parada M.

A iteração de parada, um importante argumento de controle no algoritmo, pode ser determinada via validação cruzada ou algum critério de informação. Recomenda-se que o *comprimento do passo* v no passo (iv) seja pequeno, como por exemplo $v = 0,1$. Um menor valor de v tipicamente requer um maior número de iterações boosting e, consequentemente, maior tempo de computação. Quando se escolhe v "suficientemente pequeno", resultados empíricos mostram que a acurácia preditiva do modelo é a melhor dentre outros valores de v (BUHLMANN; HOTHORN, 2007).

2.2.1.3 Função Perda e Algoritmos Boosting

Como visto na seção anterior, o algoritmo Gradiente Boosting de Friedman exige que dois argumentos sejam definidos. Um desses argumentos é a função perda.

Vários algoritmos Boosting podem ser definidos especificando diferentes funções perda $\rho(\cdot, \cdot)$ e serão mostrados a seguir os algoritmos derivados de diferentes funções perdas. Considere o caso em que a resposta Y tem natureza

binária, ou seja, $Y \in \{0,1\}$ com $\pi(\boldsymbol{x}) = P[Y = 1 \,|\boldsymbol{X} = \boldsymbol{x}]$. Seguindo as recomendações de Buhlmann e Hothorn (2007) é conveniente codificar a resposta por $\tilde{Y} = 2Y - 1 \in \{-1, 1\}$. Ainda, segundo os mesmos autores, essa transformação proporciona eficiência computacional quando utilizado o *valor marginal*, que será explicado mais adiante. Considere o negativo da log-verossimilança da binomial como função perda:

$$\rho[y, \pi(\boldsymbol{x})] = -[y \ln \pi(\boldsymbol{x}) + (1 - y) \ln(1 - \pi(\boldsymbol{x}))] \tag{2.19}$$

por simplificação, a perda (2.19) será chamada daqui em diante de *perda binomial*. Sendo $\pi(\boldsymbol{x})$ dado por,

$$\pi(\boldsymbol{x}) = \frac{e^{g(\boldsymbol{x})}}{e^{g(\boldsymbol{x})} + e^{-g(\boldsymbol{x})}}, \tag{2.20}$$

tal que,

$$g(\boldsymbol{x}) = \frac{1}{2} \ln\left(\frac{\pi(\boldsymbol{x})}{1 - \pi(\boldsymbol{x})}\right) \tag{2.21}$$

é igual à metade do log da chance (*log-odds*). Observe que facilmente pode-se obter a equação (2.21) a partir de (2.20), e vice versa, sabendo que

$$\pi(\boldsymbol{x}) = \frac{e^{g(\boldsymbol{x})}}{e^{g(\boldsymbol{x})} + e^{-g(\boldsymbol{x})}} = \frac{e^{2g(\boldsymbol{x})}}{e^{2g(\boldsymbol{x})} + 1}.$$

Note-se que $\pi(\boldsymbol{x})$ assume valores contínuos no intervalo $(0, 1)$, uma vez que está associado com a probabilidade de ocorrer $Y = 1$ de um evento Bernoulli. Ainda, o componente $g(\boldsymbol{x})$ assume valores contínuos no intervalo $(-\infty, \infty)$. Logo, a equação (2.21) transforma as probabilidades da equação (2.20) em valores contínuos no intervalo $(-\infty, \infty)$. Da mesma forma, e equação (2.21) transforma valores

contínuos no intervalo $(-\infty,\infty)$ em valores contínuos no intervalo $(0,1)$.

Mas, em problemas de classificação, é necessário que se tenha uma forma de classificar um objeto numa determinada classe segundo sua probabilidade ou com base em $g(\boldsymbol{x})$. Para tal, considere $\tilde{y}g$, o chamado valor marginal, em que $\tilde{Y} = 2Y - 1 \in \{-1, 1\}$. Algumas observações sobre $\tilde{y}g$:

- Se $\tilde{y} = 1$ e $\pi(\boldsymbol{x}) > \frac{1}{2}$, implica que, pela equação (2.21), $g(\boldsymbol{x}) > 0$ e $\tilde{y}g > 0$. Isso quer dizer que o objeto tem maior probabilidade de ser classificado como sendo $\tilde{y} = 1$;

- Se $\tilde{y} = 1$ e $\pi(\boldsymbol{x}) < \frac{1}{2}$, implica que, pela equação (2.21), $g(\boldsymbol{x}) < 0$ e $\tilde{y}g < 0$. Isso quer dizer que o objeto tem menor probabilidade de ser classificado como sendo $\tilde{y} = 1$ e, nesse caso, seria classificado como $\tilde{y} = -1$;

- Se $\tilde{y} = -1$ e $\pi(\boldsymbol{x}) > \frac{1}{2}$, implica que, pela equação (2.21), $g(\boldsymbol{x}) > 0$ e $\tilde{y}g < 0$. Isso quer dizer que o objeto tem menor probabilidade de ser classificado como sendo $\tilde{y} = -1$ e, nesse caso, seria classificado como $\tilde{y} = 1$;

- Se $\tilde{y} = -1$ e $\pi(\boldsymbol{x}) < \frac{1}{2}$, implica que, pela equação (2.21), $g(\boldsymbol{x}) < 0$ e $\tilde{y}g > 0$. Isso quer dizer que o objeto tem maior probabilidade de ser classificado como sendo $\tilde{y} = -1$;

- Caso Se $\tilde{y} = 1$ ou $\tilde{y} = -1$ e $\pi(\boldsymbol{x}) = \frac{1}{2}$, implica que, pela equação (2.21), $g(\boldsymbol{x}) = 0$ e $\tilde{y}g = 0$. Isso causa uma indeterminação no sentido de que a observação deve ser classificada em qualquer uma das classes. Contudo, com probabilidade zero, o valor $g(\boldsymbol{x}) = 0$ ocorre.

Diante das considerações feitas, uma classificação incorreta ocorrerá se $\tilde{Y}g(\boldsymbol{X}) < 0$. Assim, a perda binomial utilizando o valor marginal é dada por

$$\rho\left(y, g\left(\boldsymbol{x}\right)\right) = \ln\left(1 + e^{-2\tilde{y}g}\right). \tag{2.22}$$

Convém ressaltar que a diferença entre as perdas (2.19) e (2.22) é que a perda (2.19) depende de $\pi\left(\boldsymbol{x}\right)$ e ao substituir $\pi\left(\boldsymbol{x}\right)$ (dado na equação (2.20)) em (2.19) e substituir Y por $\tilde{Y}$, obtém-se a perda em (2.22), que depende de g.

Pode-se mostrar que o minimizador populacional da perda binomial em (2.22) é dado por,

$$f^{*}\left(\boldsymbol{x}\right) = \frac{1}{2}\ln\left(\frac{\pi\left(\boldsymbol{x}\right)}{1 - \pi\left(\boldsymbol{x}\right)}\right), \tag{2.23}$$

em que, $\pi\left(\boldsymbol{x}\right)$ é como definido em (2.20).

Uma função perda alternativa à binomial é a perda exponencial (Figura 13), dada pela expressão (2.24).

$$\rho\left(y,\, g\right) = e^{-\tilde{y}g}, \tag{2.24}$$

cujo minimizador populacional pode ser mostrado como o mesmo para perda binomial (expressão 2.23) (FRIEDMAN; HASTIE; TIBSHIRANI, 2001). Diante das considerações feitas para o valor marginal $\tilde{y}g\left(x\right)$, se uma classificação incorreta ocorre, isto é, $\tilde{y}g < 0$, a função perda binomial ou exponencial tem maiores valores nessa região e, consequentemente, a solução se afasta do risco empírico ótimo (Figura 13).

Frente ao exposto, utilizar o algoritmo GBF com diferentes funções perdas leva a diferentes algoritmos Boosting. Quando usando a perda binomial (2.22), obtemos o algoritmo Binomial Boosting e, com a perda exponencial em 2.24, obtemos o algoritmo AdaBoost para estimação funcional. Friedman, Hastie e Tibshirani (2001) afirmam ainda que a perda binomial é uma aproximação da perda 0-1

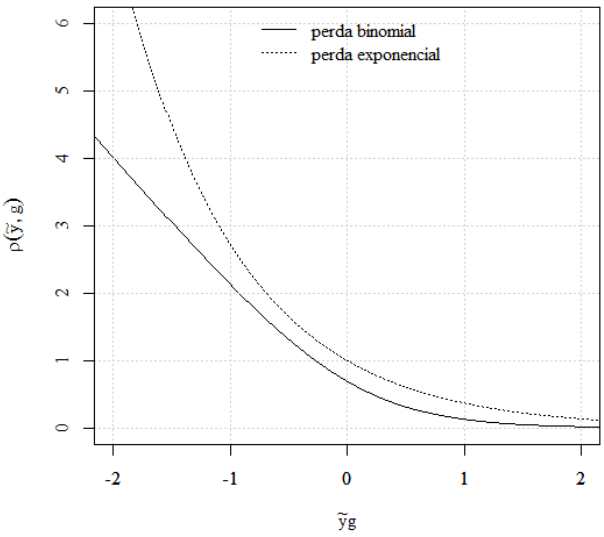


Figura 13 Funções perda binomial e exponencial como função do valor marginal $\tilde{y}g$

(função degrau), cujo minimizador populacional resulta no classificador de Bayes. Essa função é descontínua em zero e nesse caso não pode ser usada no algoritmo Gradiente Boosting de Friedman.

Importante ressaltar que a interpretação da estimativa Boosting $\hat{g}^{(m)}(\cdot)$ é feita como uma estimativa do minimizador populacional $f^*(\cdot)$. Dessa forma, os resultados do algoritmo Adaboost e Binomial Boosting correspondem as estimativas da metade do log da chance. Desse modo, as estimativas de probabilidade são dadas por

$$\hat{\pi}^{(m)}(\boldsymbol{x}) = \frac{e^{\hat{g}^{(m)}(\boldsymbol{x})}}{e^{\hat{g}^{(m)}(\boldsymbol{x})} + e^{-\hat{g}^{(m)}(\boldsymbol{x})}} \tag{2.25}$$

A razão da construção dessas estimativas de probabilidades estão basea-

das no fato de que Boosting com iteração de parada razoável é consistente, no sentido de que existe um número ideal de iterações do algoritmo Boosting tal que o risco empírico converge para o risco teórico (BARTLETT; TRASKIN, 2007). Esse número ideal de iterações será discutido na seção seguinte.

Como visto, diferentes algoritmos Boosting podem ser construídos especificando diferentes funções perdas. Para regressão com resposta $Y \in \mathbb{R}$, é conveniente utilizar a função perda erro quadrático, também conhecida como perda L_2, correspondendo ao algoritmo L_2Boosting, dentre outros (HOFNER et al., 2014).

2.2.1.4 Procedimento base para modelos lineares generalizados

Nas duas seções anteriores foi apresentado o Algoritmo Gradiente Boosting de Friedman e foi visto que, para que seja possível sua aplicabilidade, é necessário que dois argumentos sejam definidos: uma função perda e um procedimento base.

O termo *procedimento base* pode ser visto com outras denominações em diferentes textos na comunidade de Aprendizado de Máquinas. Por exemplo, quando algoritmos de classificação são propostos, é comum se referir a um procedimento base como *classificador fraco*. Também pode ser referido como *base aprendiz* (*base-learner*), e é importante salientar que é possível classificá-lo em três categorias de modelos: modelos lineares; modelos de suavização e árvores de decisão (NATEKIN; KNOLL, 2013). Existem ainda outras classes de modelos que necessitam que um procedimento base seja especificado, como por exemplo os Campos Aleatórios Markovianos (DIETTERICH; ASHENFELTER; BULATOV, 2004) ou Ondaletas (Wavelets) (VIOLA; JONES, 2001).

Nesta seção serão feitas considerações sobre o procedimento base utilizado para o ajuste de modelos lineares generalizados no contexto de Boosting.

Como definido na seção 2.2.1.3, um procedimento base é uma função que relaciona as variáveis independentes à resposta. Mais ainda, no caso de modelos lineares generalizados, o procedimento base é a função associada à componente sistemática do modelo.

Um modelo linear generalizado com covariáveis $\boldsymbol{x} = (x_1, \ldots, x_p)^T$ tem a forma

$$g(\mu) = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p,$$

com esperança condicional $\mu = E[Y \,|\boldsymbol{X}]$, função de ligação g e vetor de parâmetros $\boldsymbol{\beta}$.

Considere o seguinte procedimento base

$$\hat{g}(x) = \hat{\beta}^{(\hat{\lambda})} x^{(\hat{\lambda})}, \tag{2.26}$$

em que,

$$\hat{\beta}^{(j)} = \frac{\sum_{i=1}^{n} x_i^{(j)} u_i}{\sum_{i=1}^{n} \left(x_i^{(j)}\right)^2}, \tag{2.27}$$

em que, $u_1, \ldots, u_n$ é gradiente negativo de $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$, e

$$\hat{\lambda} = \underset{1 \leq j \leq p}{\arg\min} \sum_{i=1}^{n} \left(u_i - \hat{\beta}^{(j)} x_i^{(j)}\right)^2. \tag{2.28}$$

O procedimento em (2.26) é utilizado no passo (iii) do algoritmo GBF, descrito na seção 2.2.1.2 . O procedimento em (2.26) ajusta modelos lineares simples sem intercepto separadamente para cada coluna da matriz de delineamento ao vetor gradiente negativo, u_i, na expressão (2.27). Em outras palavras, para cada covariável ou variável de entrada j $(j = 1, \ldots, p)$, um modelo de regressão linear

simples sem intercepto é ajustado e, nesse caso, haverão p procedimentos bases ajustados na iteração m do algoritmo.

Dos p procedimentos bases ajustados, apenas o melhor modelo ajustado é usado no processo de atualização, ou seja, apenas o procedimento base que fornecer a menor quantidade em (2.28) será adicionado à corrente iteração (passo (iv)) do algoritmo.

No passo (ii) do algoritmo GBF, o vetor gradiente $\mathbf{u}^{(m)} = \left(u_i^{(m)}\right)\Big|_{=1,\ldots,n}$ é visto como uma estimativa do verdadeiro gradiente de R, o risco empírico definido em (2.14), que é adicionada à corrente estimativa $f^*(\cdot)$ em cada iteração do algoritmo. Consequentemente, o algoritmo GBF faz com que o vetor gradiente negativo percorra iterativamente no espaço de $g(\cdot)$ para minimizar R. Conforme Hofner et al. (2014), essa estratégia corresponde a substituir o clássico método Escore-Fisher em estimação por Máxima Verossimilhança de f^* pelo algoritmo gradiente descendente no espaço funcional de f^*, no caso o algoritmo GBF.

Diante do exposto, a função de predição ótima ou minimizador populacional $f^*(\cdot)$, em modelos lineares generalizados é a função $g(\mathbf{x})$ que retorna o menor risco empírico.

Uma observação importante sobre o algoritmo GBF é que utilizando o procedimento base em (2.26), automaticamente é realizado o processo de seleção de variáveis em um modelo de regressão múltipla. Por essa razão e utilizando o procedimento base em (2.26), diz-se que o procedimento de seleção de variáveis está embutido no algoritmo Gradiente Boosting de Friedman (FRIEDMAN, 2001).

Quando utilizado L_2Boosting com esse procedimento base, selecionamos em cada iteração uma variável preditora, não necessariamente uma diferente para cada iteração, e atualizamos a função linearmente:

$$\hat{f}^{(m)}(x) = \hat{f}^{(m-1)}(x) + v \cdot \hat{\beta}^{(\hat{\lambda}_m)} x^{(\hat{\lambda}_m)} \tag{2.29}$$

em que, $\hat{\lambda}_m$ denota o índice da variável preditora na iteração m, ou seja, a variável preditora (ou contribuição de uma determinada variável) que é selecionada na iteração m é aquela que apresentou o menor $\hat{\lambda}$ (2.28) dos p procedimentos bases ajustados no passo (iii) do algoritmo GBF. Alternativamente, a atualização dos coeficientes estimados é

$$\hat{\beta}^{(m)} = \hat{\beta}^{(m-1)} + v\hat{\beta}^{(\hat{\lambda}_m)} \tag{2.30}$$

A equação (2.30) indica que apenas o $\hat{\lambda}_m$-ésimo componente dos coeficientes estimados $\hat{\beta}^{(m)}$ (na iteração m) foi atualizado. Para cada iteração m, obtemos o ajuste de um modelo linear. Conforme m tende ao infinito, $\hat{f}^{(m)}(\cdot)$ converge para a solução de mínimos quadrados, que é única se a matriz de delineamento tem posto completo.

Um importante argumento do algoritmo GBF é o número de iterações de execução do algoritmo. Como visto, os passos (ii) a (iv) do algoritmo são executados até a m-ésima iteração. Uma forma de escolher o número ideal (ótimo) de iterações é escolher o m que minimiza algum critério de informação, como por exemplo o critério de informação de Akaike (HOFNER et al., 2014).

No Apêndice A* é apresentada uma ilustração didática do Algoritmo Gradiente Boosting de Friedman, bem como o processo de seleção de variáveis e a construção do modelo via Boosting.

Quando utilizado L_2Boosting com o procedimento base em 2.27, um valor inicial adequado para o passo (i) do algoritmo é calcular a média da variável resposta Y. Além disso, o vetor gradiente negativo é dado por

$$u_i = -\frac{\partial \rho\,(y,\, g)}{\partial g} = y - g \tag{2.31}$$

em que g é o procedimento base utilizado.

Quando usado Binomial Boosting com o procedimento base em (2.26), obtemos um ajuste, incluindo seleção de variáveis, de um modelo de regressão logística linear (BUHLMANN; HOTHORN, 2007). Um valor inicial adequado para esse algoritmo é calcular a frequência relativa de $Y = 1$ da amostra. O vetor gradiente negativo para a perda binomial é dado por

$$u_i = -\frac{\partial \rho\,(y,\, g)}{\partial g} = y_i - \frac{1}{1 + e^{-g}}, \tag{2.32}$$

em que, g é o procedimento base utilizado. Assim, o algoritmo Binomial Boosting utiliza a frequência relativa de $Y = 1$ e o vetor u_i para percorrer o espaço paramétrico do modelo proposto (BERK, 2008).

Foi mostrado na seção 2.2.1.2 que é possível fazer uma analogia entre o processo de otimização feita por máxima verossimilhança e o feito por Boosting e, diante disso, um questionamento pode ser feito: qual é o ganho em utilizar a metodologia de Boosting para ajustar, no caso, um modelo linear generalizado? As considerações a seguir ajudam a responder essa questão.

- A estimativa $\hat{\beta}$ de Máxima Verossimilhança é aquela que maximiza a probabilidade conjunta de a amostra ter ocorrido e, para isso, todas as variáveis e observações fazem parte do processo de otimização. A estimativa $\hat{\beta}$ de Boosting é aquela cujas contribuições de cada variável tornam mínimo o risco empírico. Note que essencialmente a diferença reside no fato de que em um processo todas variáveis e observações fazem parte do processo de otimização e no outro, apenas as variáveis "mais importantes", no sentido de tornar

mínimo o risco empírico, participam do processo de otimização.

- Conforme aumenta-se o número de variáveis a serem analisadas, a estimação por Máxima Verossimilhança pode ser comprometida, uma vez que operações com inversão de matrizes são necessárias. Isso pode ser visto nas seções seguintes quando são abordados os estimadores de Máxima Verossimilhança dos modelos de regressão logística e simplex. Esse problema não ocorre em Boosting, uma vez que cada variável é ajustada em um procedimento base.

- Boosting pode ser utilizado em casos em que a matriz de delineamento tem alta dimensão, ou seja, o número de observações é menor (ou até mesmo muito menor) que o número de variáveis. Buhlmann e Hothorn (2007) dão um exemplo com $n = 49$ amostras de pacientes com câncer de mama e $p = 7129$ níveis de expressão gênica e para essa situação utilizou o algoritmo Binomial Boosting.

- Existem situações em que a matriz de delineamento é esparsa, ou seja, existem elementos amostrais em que não se tem a medida de uma determinada variável. Nessas situações os algoritmos Boosting podem ser utilizados.

- Em conjuntos de dados que contêm um grande número de variáveis preditoras em relação ao tamanho amostral, a variância dos estimador de máxima verossimilhança é inflacionada. Esse problema tem como consequência o super ajuste, o que diminui a precisão na predição de uma modelo clássico de regressão. Esse fato pode ser verificado pela matriz de informação de Fisher do modelo. Nesse caso, ao método de Boosting pode ser utilizado, uma vez que este automaticamente realiza seleção de variáveis (SCHMID et al., 2013).

2.3 Regressão Logística

Nos modelos de regressão linear simples ou múltipla, a variável dependente Y é uma variável aleatória de natureza quantitativa. No entanto, em algumas situações, a variável dependente é qualitativa e expressa por duas ou mais categorias, ou seja, admite dois ou mais valores. Nesse caso, o método dos mínimos quadrados não oferece estimativas adequadas. Uma boa aproximação é obtida pela regressão logística que permite o uso de um modelo de regessão para se calcular ou prever a probabilidade de um evento específico (NELDER; WEDDERBURN, 1972).

As categorias ou valores que a variável dependente assume podem ser de natureza nominal ou ordinal. Em caso de natureza ordinal, há uma ordem natural entre as possíveis categorias e, então, tem-se o contexto da Regressão Logística Ordinal. Quando essa ordem não existe entre as categorias da variável dependente assume-se o contexto da Regressão Logística Nominal.

O seguinte exemplo ilustra uma situação em que a variável dependente possui natureza nominal. Suponha que se deseja estudar a toxicidade de uma certa droga e as categorias são: o animal morreu após administração da dose x $(Y = 1)$ ou o animal não morreu após administração da dose x $(Y = 0)$. Nesse contexto, dosagens $x_1 < x_2 < \ldots < x_m$ são fixadas. A dosagem x_i é administrada em uma quantidade c_i de animais. Após esse procedimento, ocorre um número y_i de mortes para cada i, com $1 \leq i \leq m$, em que m é o número dosagens administradas. Assume-se que $\pi(x)$ é a probabilidade que um animal escolhido aleatoriamente morra com a dosagem x. Dessa forma, y_i, são variáveis aleatórias independentes com distribuição binomial $Bin(c_i, \pi(x_i))$, com $i \in \{1, \ldots, m\}$. O objetivo aqui é encontrar um modelo no qual, para cada valor da variável independente x_i, é pos-

sível predizer a variável dependente $y\,(x_i)$, a qual é binomial com probabilidade de sucesso $\pi\,(x_i)$.

No contexto do experimento descrito na seção 2.1.1, nove dietas ($m = 9$), caracterizadas pelas variáveis x_1 (gordura), x_2 (carboidrato) e x_3 (fibra), foram avaliadas em grupos contendo 30 ratos ($c_i = 30$ para cada $i = 1, \ldots, 9$). Após administração das dietas, ocorreu um número y_i de ratos que tiveram a expressão de tumor mamário. Assim, seria razoável pensar que $\pi\,(x)$ é a probabilidade que um rato escolhido aleatoriamente tenha o tumor mamário com a dieta x_1, x_2 e x_3. Dessa forma, y_i, são variáveis aleatórias independentes com distribuição binomial $Bin\,(c_i,\, \pi\,(x_i))$.

2.3.1 Regressão Logística para dados binomiais

Antes de se iniciar a discussão sobre a regressão logística, é interessante fazer um breve comentário sobre Modelos Lineares Generalizados (MLG). Um modelo linear generalizado é especificado por três componentes: uma componente aleatória, a qual identifica a distribuição de probabilidade da variável dependente, uma componente sistemática, que especifica uma função linear entre as variáveis independentes e uma função de ligação, que descreve a relação matemática entre a componente sistemática e o valor esperado da componente aleatória (HOSMER; LEMESHOW, 1989).

Em outras palavras, a componente aleatória de um MLG consiste nas observações da variável aleatória Y, ou seja, com o vetor $\mathbf{y} = (y_1,\, y_2,\, \ldots,\, y_n)$.

A componente sistemática do MLG é definida através de um vetor $\boldsymbol{\eta} = (\eta_1,\, \eta_2,\, \ldots,\, \eta_n)$ que está associado ao conjunto das variáveis independentes por meio de um modelo linear $\boldsymbol{\eta} = \boldsymbol{x\beta}$, em que $\boldsymbol{x}$ é uma matriz que consiste nas

variáveis independentes das n observações e $\boldsymbol{\beta}$ é o vetor de parâmetros do modelo.

A terceira componente do MLG é a função de ligação entre as componentes aleatória e sistemática. Seja $\mu_i = E\left[Y_i \left| \boldsymbol{x}_i \right.\right]$, com $i \in \{1, \ldots, n\}$, então η_i é definida por $\eta_i = g\left(\mu_i\right)$, em que g é uma função monotônica e diferenciável. O fato de ser diferenciável é importante pelo fato de que no processo de estimação de $\boldsymbol{\beta}$, será necessário utilizar derivadas da função de ligação para compor o vetor gradiente e matriz de informação de Fisher. O fato de ser monotônica é importante no sentido de que, dado um valor particular para $\boldsymbol{x\beta}$, este corresponderá a um único μ e, além disso, se g é uma função monotônica, então ela tem uma função inversa, isto é, $\boldsymbol{x\beta} = g\left(\mu\right)$ implica que $g^{-1}\left(\boldsymbol{x\beta}\right) = \mu$.

Dessa forma, a função de ligação conecta os valores esperados das observações às variáveis explanatórias, para $i \in \{1, \ldots, n\}$, pela fórmula

$$g\left(\mu_i\right) = \beta_0 + \sum_{j=1}^{p} \beta_j \; x_{i\,j} \tag{2.33}$$

em que p é o número de variáveis independentes (covariáveis) no modelo.

É interessante comentar que, se a função g, em (2.33), for a função identidade, tem-se então o modelo de regressão linear.

Dependendo da natureza da componente aleatória de um MLG, existe um MLG adequado para cada situação. Se a componente aleatória for de natureza binária, os modelos *logit*, *probit* e *gompit* (*complemento log-log*) são adequados. Se a componente aleatória consiste do resultado de contagens, os modelos log-linear de Poisson e Binomial Negativo são candidatos. Para situações cuja resposta é contínua e assimétrica, os modelos Gama são candidatos (PAULA; TUDER, 1986).

Na sequência, apresenta-se o modelo de regressão logística binário, que é um caso particular dos modelos lineares generalizados, mais especificamente dos

modelos *logit*.

Para se analisar $\pi(\boldsymbol{x})$, tomam-se as observações independentes $x_1, x_2, \ldots, x_n$. Nesse contexto, é razoável assumir, como suposição inicial, que $\pi(\boldsymbol{x})$ é uma função monotônica com valores entre zero e um (inclusive zero e um), quando $\boldsymbol{x}$ varia na reta real, ou seja, $\pi(\boldsymbol{x})$ é uma função de distribuição de probabilidade.

Como $\pi(\cdot)$ varia entre zero e um, uma representação linear simples para π sobre todos os possíveis valores de $\boldsymbol{x}$ não é adequada, uma vez que os valores da forma linear estão no intervalo $(-\infty; +\infty)$. Nesse caso, uma transformação deve ser utilizada a fim de permitir que, para qualquer valor de $\boldsymbol{x}$, tenha-se um valor correspondente para $\pi(\cdot)$ no intervalo $[0; 1]$. Considere a transformação logística, também chamada de *logit*, logo

$$logit = \ln\left[\frac{\pi(\boldsymbol{x})}{1-\pi(\boldsymbol{x})}\right] = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p \tag{2.34}$$

A razão $\frac{\pi(\boldsymbol{x})}{1-\pi(\boldsymbol{x})}$ é chamada de chance (*odds*). Seja A um evento de interesse, logo a chance do evento A é a relação entre probabilidade de ocorrência de A e a probabilidade de não ocorrência de A. Suponha que a probabilidade de ocorrência de A é de 80%, então a chance de ocorrência desse evento é de 4 para 1, ou em porcentagem, de 400% (400 ocorrências para 100 não ocorrências). Da mesma forma, se um evento A tem chance de 0,25 (25% ou 1 para 4) de ocorrer, então a probabilidade de ocorrência de A é de 20%.

A chance varia na escala de $(0; +\infty)$. Então o logaritmo neperiano da chance (*ln odds*) varia em $(-\infty; +\infty)$. Na expressão (2.34), se $\pi(\boldsymbol{x}) = 0,5$, então $logit = 0$. Se $\pi(\boldsymbol{x}) < 0,5$, então $logit < 0$ e se $\pi(\boldsymbol{x}) > 0,5$, então $logit > 0$.

Exponenciando a expressão (2.34), tem-se que

$$e^{logit} = e^{\beta_0+\beta_1 x_1+\ldots+\beta_p x_p}$$

$$\frac{\pi(\boldsymbol{x})}{1-\pi(\boldsymbol{x})} = e^{\beta_0+\beta_1 x_1+\ldots+\beta_p x_p}$$

O inverso da função logit é a função logística, que é dada por

$$\pi(\boldsymbol{x}) = \frac{\exp(\beta_0+\beta_1 x_1+\ldots+\beta_p x_p)}{1+\exp(\beta_0+\beta_1 x_1+\ldots+\beta_p x_p)} \tag{2.35}$$

em que $\pi(\boldsymbol{x})$ varia em $[0;1]$. No caso de termos uma variável independente no modelo, x_1, se $\beta_1 > 0$, π é crescente e se $\beta_1 < 0$, π é decrescente. Quando x tende ao infinito, $\pi(x)$ tende a zero quando $\beta_1 < 0$ e tende a um quando $\beta_1 > 0$. Assim, dessa forma, define-se qualitativamente a função de ligação (vide Figura 14) necessária ao modelo, definido na equação (2.35). Caso $\beta_1 = 0$, a variável resposta Y é independente da variável X, logo $\pi(x)$ é constante. O caso $\beta_0 = 0$ e $\beta_1 = 0$ corresponde a $\pi(x) = 0,5$ (Figura 14).

O termo *regressão logística* vem do fato de esse tipo de modelo de regressão ser derivado da distribuição logística padrão, quando utilizado a função de ligação *logito*. Para tal, considere ainda uma variável independente no modelo. Se $\beta_0 = 0$ e $\beta_1 = -1$ então $\pi(x)$ é chamada de *função de distribuição logística padrão*. Sua respectiva *função de distribuição acumulada* (f.d.a.) é dada por

$$F(x) = \frac{\exp(-x)}{1+\exp(-x)}. \tag{2.36}$$

E, consequentemente, sua função de densidade de probabilidade (f.d.p.) e ilustrada na Figura 14(b), obtida por derivação da equação (2.36), é dada por

$$f(x) = \frac{\exp(-x)}{[1+\exp(-x)]^2}. \tag{2.37}$$

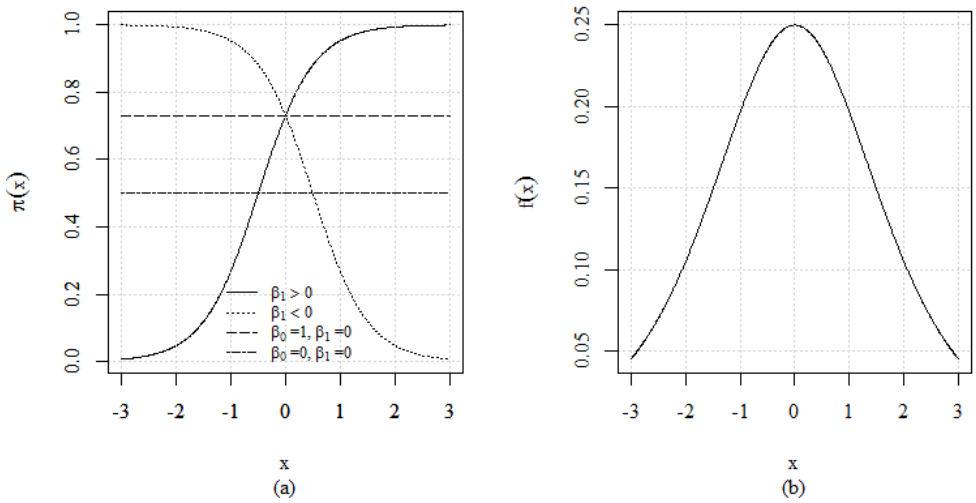


Figura 14 Ilustração do modelo logístico em uma variável independente considerando $\beta_0 = 1$ e $\beta_1 = 2$ quando $\beta_1 > 0$, $\beta_0 = 1$ e $\beta_1 = -2$ quando $\beta_1 < 0$ (a) e função densidade de probabilidade da distribuição logística padrão (b).

Caso haja reparametrização, na equação (2.36), dada por

$$\mu = \frac{-\beta_0}{\beta_1} \qquad \sigma = \frac{-1}{\beta_1},$$

tem-se que $\pi(\cdot)$ dada pela equação (2.36) assume a forma

$$\pi(x) = F\left(\frac{x-\mu}{\sigma}\right), \tag{2.38}$$

em que μ é o *centro de simetria* e σ é o *parâmetro de escala* da reparametrização.

Com base no exposto, pode-se definir o *modelo logístico* da seguinte forma: Seja $g(x) = \beta_0 + \beta_1 x$, para qualquer $x \in \mathbb{R}$. Seja $F(\cdot)$ a f.d.a. definida na equação (2.36) correspondente a $\beta_0 = 0$ e $\beta_1 = -1$. Ao reparametrizar-se $\mu = -\beta_0 (\beta_1)^{-1}$ e $\sigma = -(\beta_1)^{-1}$, para cada x_i, para $i \in \{1, \ldots, n\}$, temos que

$$F\left(\frac{x_i - \mu}{\sigma}\right) = \pi(x_i) = E[Y_i | x_i], \tag{2.39}$$

em que y_i representa uma amostra aleatória de tamanho um de $F(\cdot)$ (HOSMER; LEMESHOW, 1989). Desse modo, se $\mu = 0$ e $\sigma = 1$, o modelo em (2.39) corresponde à função de distribuição logística padrão. Essa distribuição é simétrica em torno de $\pi(x) = \frac{1}{2}$.

2.3.2 Estimação dos parâmetros do modelo de Regressão Logística para dados binomiais

Suponha que n observações binomiais independentes $y_1, \ldots, y_n$ são tais que a i-ésima observação, $i = 1, \ldots, n$, tem distribuição binomial com parâmetros n_i e p_i. Também suponha que o valor transformado da probabilidade de resposta para a i-ésima observação está relacionado a uma combinação linear de p variáveis explanatórias, isto é,

$$g(\pi_i) = \beta_0 + \beta_1 x_{1i} + \ldots + \beta_p x_{pi} = \mathbf{x}_i^T \boldsymbol{\beta} = \eta_i, \tag{2.40}$$

em que $g(\pi)$ pode ser a transformação logística de π, o "probit" de π, a transformação complemento log-log de π, entre outras. O componente linear deste modelo será denotado por η_i e a função g a função de ligação de um modelo linear generalizado.

O método usual para estimar $\boldsymbol{\beta}$ é via Máxima Verossimilhança. O logaritmo da função de verossimilhança para n observações binomiais é dado por

$$\begin{aligned} l(\boldsymbol{\beta}) &= \sum_{i=1}^{n} l_i(\pi_i) \\ &= \sum_{i=1}^{n} \left\{ \ln \left(\begin{array}{c} n_i \\ y_i \end{array} \right) + y_i \ln(\pi_i) + (n_i - y_i) \ln(1-\pi_i) \right\} \end{aligned} \tag{2.41}$$

Derivando $l(\boldsymbol{\beta})$ em relação a β_j, temos

$$\frac{\partial l(\boldsymbol{\beta})}{\partial \beta_j} = \sum_{i=1}^{n} \frac{\partial l_i(\pi_i)}{\partial \pi_i} \frac{\partial \pi_i}{\partial \eta_i} \frac{\partial \eta_i}{\partial \beta_j},$$

em que

$$\begin{aligned} \frac{\partial l_i(\pi_i)}{\partial \pi_i} &= y_i \frac{1}{\pi_i} + n_i \frac{1}{1-\pi_i}(-1) - y_i \frac{1}{1-\pi_i}(-1) \\ &= y_i \frac{1}{\pi_i} - n_i \frac{1}{1-\pi_i} + y_i \frac{1}{1-\pi_i} = \frac{y_i - n_i \pi_i}{\pi_i (1-\pi_i)}. \end{aligned}$$

para obter $\frac{\partial \pi_i}{\partial \eta_i}$, basta observar que $g(\cdot)$ depende de η_i por meio das probabilidades binomiais, logo

$$\begin{aligned} \frac{\partial \eta_i}{\partial \pi_i} = \frac{\partial g(\pi_i)}{\partial \pi_i} &= \frac{\partial}{\partial \pi_i} \left[\ln \left(\frac{\pi_i}{1-\pi_i} \right) \right] = \frac{1}{\frac{\pi_i}{1-\pi_i}} \frac{(1-\pi_i) + \pi_i}{(1-\pi_i)^2} \\ &= \frac{1-\pi_i}{\pi_i} \frac{1}{(1-\pi_i)^2} = \frac{1}{\pi_i (1-\pi_i)} \end{aligned}$$

e, como $g(\cdot)$ é diferenciável e possui inversa, temos que

$$\frac{\partial \pi_i}{\partial \eta_i} = \frac{1}{\frac{\partial \eta_i}{\partial \pi_i}} = \pi_i (1 - \pi_i)$$

Por fim, como $\eta_i = \beta_1 x_{1i} + \ldots + \beta_p x_{pi}$, temos que $\frac{\partial \eta_i}{\partial \beta_j} = x_{ji}$. Assim, o vetor escore, formado pelas derivadas parciais de primeira ordem da função em (2.41) é dado por

$$U_{\boldsymbol{\beta}}(l(\boldsymbol{\beta})) = \frac{\partial l(\beta)}{\partial \beta_j} = \sum_{i=1}^{n} \frac{y_i - n_i \pi_i}{\pi_i (1 - \pi_i)} \frac{1}{g'(\pi_i)} x_{ji}. \tag{2.42}$$

A expressão (2.42) pode ser escrita de forma matricial da seguinte maneira: seja $\mathbf{X}$ a matriz de ordem $(n \times p)$ das p variáveis explanatórias, $\mathbf{Y}$ o vetor dos valores y_i, $\mathbf{V}$ o vetor das probabilidades ajustadas com o i-ésimo elemento $\hat{\pi}_i$. Assim,

$$U_{\boldsymbol{\beta}}(l(\boldsymbol{\beta})) = \mathbf{X}^T (\mathbf{Y} - \mathbf{V}). \tag{2.43}$$

A matriz hessiana ($I_{\boldsymbol{\beta}}$) constituída pelas derivadas de segunda ordem de (2.41) em relação a cada parâmetro. Uma vez obtidos o vetor escore e a matriz hessiana, temos condições de montarmos a regra de Newton Raphson, dada por

$$\boldsymbol{\beta}^{(i+1)} = \boldsymbol{\beta}^{(i)} - \left(I_{\boldsymbol{\beta}}\right)^{-1} U_{\boldsymbol{\beta}}(l(\boldsymbol{\beta})) \tag{2.44}$$

em que $U_{\boldsymbol{\beta}}(l(\boldsymbol{\beta}))$ é o vetor gradiente e $\boldsymbol{\beta}^{(i)}$ representa um valor inicial para a primeira iteração do método de Newton-Raphson.

Um problema que pode surgir para o método de Newton-Raphson é que a inversa da matriz hessiana pode não existir. Em situações do tipo, o método de Escore-Fisher tem sido utilizado e esse método consiste em substituir a matriz

Hessiana no método de Newton-Raphson pela matriz de informação de Fisher esperada ($I_E(\boldsymbol{\beta})$). Pelo resultado de Wedderburn (1976), a matriz de informação de Fisher esperada é dada por

$$I_E(\boldsymbol{\beta}) = -E\left[\frac{\partial^2 l(\boldsymbol{\beta})}{\partial\beta_j \partial\beta_k}\right] = E\left[\frac{\partial l(\boldsymbol{\beta})}{\partial\beta_j}\frac{\partial l(\boldsymbol{\beta})}{\partial\beta_k}\right].$$

Nesse sentido, mostraremos essa derivação a seguir:

$$\begin{aligned}
-E\left[\frac{\partial^2 l(\boldsymbol{\beta})}{\partial\beta_j \partial\beta_k}\right] &= E\left[\frac{\partial l(\boldsymbol{\beta})}{\partial\beta_j}\frac{\partial l(\boldsymbol{\beta})}{\partial\beta_k}\right] \\
&= E\left[\sum_{i=1}^{n}\frac{y_i - n_i\pi_i}{\pi_i(1-\pi_i)}\frac{1}{g'(\pi_i)}x_{ji}\sum_{i=1}^{n}\frac{y_i - n_i\pi_i}{\pi_i(1-\pi_i)}\frac{1}{g'(\pi_i)}x_{ki}\right] \\
&= E\left[\sum_{i=1}^{n}\frac{(y_i - n_i\pi_i)^2}{[\pi_i(1-\pi_i)]^2}\frac{1}{[g'(\pi_i)]^2}x_{ji}x_{ki}\right] \\
&= \sum_{i=1}^{n}\frac{E\left[(y_i - n_i\pi_i)^2\right]}{[\pi_i(1-\pi_i)]^2}\frac{1}{[g'(\pi_i)]^2}x_{ji}x_{ki} \\
&= \sum_{i=1}^{n}\frac{n_i\pi_i(1-\pi_i)}{[\pi_i(1-\pi_i)]^2}\frac{1}{[g'(\pi_i)]^2}x_{ji}x_{ki} \\
&= \sum_{i=1}^{n}\frac{n_i}{\pi_i(1-\pi_i)}\frac{1}{[g'(\pi_i)]^2}x_{ji}x_{ki} \\
&= \sum_{i=1}^{n} n_i\pi_i(1-\pi_i)x_{ji}x_{ki}. \qquad (2.45)
\end{aligned}$$

A expressão (2.45) é obtida utilizando o fato de que as observações são independentes, para $i \neq j$, ou seja, $E[(y_i - n_i\pi_i)(y_j - n_j\pi_j)] = Cov[y_i, y_j] = 0, \text{ para } i \neq j$. Com isso, $E\left[(y_i - n_i\pi_i)^2\right] = Var[y_i] = n_i\pi_i(1-\pi_i)$.

Portanto, a matriz de informação de Fisher esperada é dada por

$$I_E(\boldsymbol{\beta}) = -E\left[\frac{\partial^2 l(\beta)}{\partial\beta_j \partial\beta_k}\right] = \sum_{i=1}^{n} n_i \pi_i (1-\pi_i) x_{ji} x_{ki}. \tag{2.46}$$

Matricialmente, a matriz de informação de Fisher esperada em (2.46) é dada por

$$I_E(\boldsymbol{\beta}) = -\mathbf{X}^T \mathbf{W} \mathbf{X}, \tag{2.47}$$

em que $\mathbf{W}$ é a matriz diagonal $n \times n$ dos pesos com o i-ésimo elemento da diagonal dado por $n_i \hat{\pi}_i (1-\hat{\pi}_i)$. O modelo de regressão logística faz parte da Família Exponencial, assim, em termos gerais, a matriz de informação de Fisher esperada é dada por

$$E\left[-\frac{\partial^2 l(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T}\right] = \phi \mathbf{X^T W X} \tag{2.48}$$

em que ϕ é o parâmetro de dispersão do modelo. No caso do modelo de regressão logística, escrevendo-o como um modelo da Família Exponencial, $\phi = 1$. Isso indica que no modelo de regressão logística a resposta está totalmente especificada e não existem parâmetros adicionais desconhecidos. Veremos mais adiante que o mesmo não ocorre na Regressão Simplex. Logo, a atualização pelo método de Escore-Fisher é dada por

$$\boldsymbol{\beta}^{(i+1)} = \boldsymbol{\beta}^{(i)} - \left[I_E(\boldsymbol{\beta})\right]^{-1} U_{\boldsymbol{\beta}}\left(l(\boldsymbol{\beta})\right)$$

$$\boldsymbol{\beta}^{(i+1)} = \boldsymbol{\beta}^{(i)} + \left(\mathbf{X^T W X}\right)^{-1} \mathbf{X^T}(\mathbf{Y} - \mathbf{V}) \tag{2.49}$$

O método de Quasi-Newton consiste em substituir a matriz Hessiana por $U_{\boldsymbol{\beta}}\left(l(\boldsymbol{\beta})\right) U_{\boldsymbol{\beta}}\left(l(\boldsymbol{\beta})\right)^T$, logo a atualização é dada por

$$\boldsymbol{\beta}^{(i+1)} = \boldsymbol{\beta}^{(i)} + \left[U_{\boldsymbol{\beta}}\left(l\left(\boldsymbol{\beta}\right)\right) U_{\boldsymbol{\beta}}\left(l\left(\boldsymbol{\beta}\right)\right)^T\right]^{-1} U_{\boldsymbol{\beta}}\left(l\left(\boldsymbol{\beta}\right)\right). \tag{2.50}$$

Ainda, um outro método que pode ser utilizado é o dos Mínimos Quadrados Reponderados, que consiste em obter iterativamente do seguinte modo

$$\boldsymbol{\beta}^{(m+1)} = \left[\mathbf{X}^T\mathbf{W}^{(m)}\mathbf{X}\right]^{-1}\mathbf{X}^T\mathbf{W}^{(m)}\mathbf{Y}^{*(m)}, \tag{2.51}$$

em que $\mathbf{Y}^* = (y_1^*, \ldots, y_n^*)^T$ é a variável dependente modificada na m-ésima iteração e é dada por

$$y_i^* = \eta_i + \frac{(y_i - n_i\pi_i)}{n_i\pi_i\left(1-\pi_i\right)}. \tag{2.52}$$

Mais detalhes desses métodos podem ser vistos em Paula (2013). A convergência do algoritmo de otimização é alcançada quando $\left|\hat{\boldsymbol{\beta}}^{(i+1)} - \hat{\boldsymbol{\beta}}^{(i)}\right| < \delta$, com δ sendo um nível de tolerância.

2.3.3 Técnicas de diagnóstico

Com o objetivo de detectar observações que influenciam no processo inferencial do modelo, serão apresentadas aqui as técnicas utilizadas para diagnosticar possíveis pontos discrepantes. Estudos de simulação têm sugerido o resíduo padronizado t_{D_i} para as análises de diagnóstico em MLG, uma vez que o mesmo tem apresentado nesses estudos propriedades similares àquelas do resíduo da regressão normal linear (WILLIAMS, 1984). Em particular, para os modelos binomiais, esse resíduo é expresso, para $0 < y_i < n_i$, na forma,

$$t_{D_i} = \pm\sqrt{\frac{2}{1-\hat{h}_{ii}}}\left[y_i \ln\left(\frac{y_i}{n_i\hat{\pi}_i}\right) + (n_i - y_i)\ln\left(\frac{n_i - y_i}{n_i - n_i\hat{\pi}_i}\right)\right]^{\frac{1}{2}} \tag{2.53}$$

em que, o sinal é o mesmo de $y_i - \hat{y}_i$ e $\hat{y}_i = n_i\hat{\pi}_i$. Se n_i for referente ao modelo binomial, y_i representa o número de sucessos ($Y = 1$) em n_i tentativas independentes. Se n_i for referente ao modelo Bernoulli, Y_i representa o evento de interesse $Y = 1$ em um ensaio e, nesse caso, $n_i = 1$. Quando $y_i = 0$ ou $y_i = n_i$, o componente do desvio padronizado assume as formas

$$t_{D_i} = -\frac{\{2n_i\,|\ln(1-\hat{\pi}_i)|\}^{1/2}}{\sqrt{1-\hat{h}_{ii}}} \tag{2.54}$$

ou

$$t_{D_i} = \frac{\{2n_i\,|\ln(\hat{\pi}_i)|\}^{1/2}}{\sqrt{1-\hat{h}_{ii}}}, \tag{2.55}$$

respectivamente.

Para se medir a influência das observações nas estimativas dos coeficientes, utilizamos a distância de Cook (LD) aproximada dada por

$$LD_i = \frac{1}{\left(1-\hat{h}_{ii}\right)^2}\frac{(y_i - n_i\hat{\pi}_i)^2}{n_i\hat{\pi}_i\,(1-\hat{\pi}_i)}. \tag{2.56}$$

Hosmer e Lemeshow (1989) observam que $\hat{h}_{ii}$ depende das probabilidades ajustadas $\hat{\pi}_i$, $i = 1, \ldots, k$, e consequentemente os resíduos t_{D_i} e a medida de influência LD_i também dependem. O valor $\hat{h}_{ii}$, também denominado de *leverage*, é dado por

$$\hat{h}_{ii} = n_i\hat{\pi}_i\,(1-\hat{\pi}_i)\,\boldsymbol{x}_i^T\left(\mathbf{X}^T\mathbf{W}\mathbf{X}\right)^{-1}\boldsymbol{x}_i \tag{2.57}$$

em que $\mathbf{W} = diag\left\{n_1\hat{\pi}_1\left(1-\hat{\pi}_1\right), \ldots, n_n\hat{\pi}_n\left(1-\hat{\pi}_n\right)\right\}$. Os autores mostraram ainda por um estudo numérico que o comportamento de $\hat{h}_{ii}$ numa regressão logística pode ser muito diferente do comportamento de $\hat{h}_{ii}$ na regressão linear para uma mesma matriz $\mathbf{X}$. Os resultados de $\hat{h}_{ii}$, t_{D_i} e LD_i são apresentados em gráficos, que são informativos quanto ao posicionamento dos pontos aberrantes e influentes com relação às probabilidades ajustadas. Nesses gráficos, os pontos mais afastados dos demais são candidatos a serem aberrantes e/ou influentes (PAULA, 1995).

A adequação do modelo linear generalizado binomial também pode ser verificada pela estimação do parâmetro de escala, ou dispersão, ϕ. Uma forma de estimar ϕ é pelo método dos momentos da estatística χ^2 de Pearson generalizada, dada por

$$\hat{\phi} = \frac{\left(\mathbf{Y}^{*(m)} - \hat{\boldsymbol{\eta}}\right)^T \mathbf{W}^{(m)} \left(\mathbf{Y}^{*(m)} - \hat{\boldsymbol{\eta}}\right)}{n-p}, \tag{2.58}$$

sendo $\mathbf{Y}^*$ o vetor de variáveis dependentes modificadas, conforme equação (2.52), $\hat{\eta}$ o vetor de estimativas dos preditores lineares, n o número de observações e p o número de parâmetros. Na equação (2.58), $\mathbf{Y}^*$ e $\mathbf{W}^{(m)}$ são obtidos da m-ésima iteração do método dos mínimos quadrados reponderados. A estimativa (2.58) é utilizada para checagem de sub ou superdispersão, em que valores inferiores ou superiores ao valor unitário evidenciam a sub ou super dispersão, respectivamente (NUNES; MORAIS; BUENO FILHO, 2004).

O gráfico normal de probabilidades para o resíduo t_{D_i} também pode ser utilizado para verificar a adequação do modelo de regressão logística, o qual indica se existem evidências de afastamento da suposição de distribuição binomial para a resposta. Consiste em gerar bandas de confiança por reamostragem, também chamado de envelope, e um ajuste adequado ocorre se todos os resíduos (ou grande parte deles) do modelo estiverem contidos nessas bandas de confiança. Mais deta-

lhes sobre o envelope simulado podem ser vistos em Atkinson (1985).

Atkinson (1981) propõe a construção por simulação de Monte Carlo de uma banda de confiança para o resíduos da regressão normal linear, a qual denominou envelope, e que permite uma melhor comparação entre os resíduos e os percentis da distribuição normal padrão. Williams (1987) discute, com base em estudos de simulação, a aproximação de forma padronizada proposta por Pregibon (1981) encontrando fortes evidências de concordância entre a distribuição empírica do componente de desvio padronizado e a distribuição normal padrão para vários modelos lineares generalizados e discute também a construção de envelopes em MLGs.

Enfim, a construção das bandas de confiança através de simulação, denominadas envelope e propostas por Atkinson (1985), pode ser feita seguindo os seguintes passos:

1. Geramos n observações da distribuição candidata considerando o parâmetro de dispersão ϕ^{-1}, as quais são armanezadas em $\mathbf{y} = (y_1, y_2, \ldots, y_n)^T$;

2. Ajustamos $\mathbf{y}$ contra $\mathbf{X}$ e obtemos r_i, $i = (1, \ldots, n)$;

3. Obtemos $t_i^* = r_i \Big/ (1 - h_{ii})^{1/2}$, $i = 1, \ldots, n$;

4. Repetimos os passos (1) - (3) K vezes. Logo teremos os resíduos gerados t_{ik}^*, $i = 1, \ldots, n$ e $k = 1, \ldots, K$;

5. Colocamos cada grupo de n resíduos em ordem crescente, obtendo $t_{(i)k}^*$, $i = 1, \ldots, n$ e $j = 1, \ldots, K$;

6. Obtemos os limites $t_{(i)I}^* = \min\limits_k t_{(i)k}^*$ e $t_{(i)S}^* = \max\limits_k t_{(i)k}^*$. Assim, os limites correspondentes ao i-ésimo resíduo serão dados por $t_{(i)I}^*$ e $t_{(i)S}^*$.

O procedimento acima foi implementado por Paula (2013) para vários modelos lineares generalizados e sugere $K = 100$ como o número de vezes para geração do envelope simulado.

2.4 Regressão Simplex

Como visto na seção anterior, a regressão logística pode ser utilizada em situações em que a resposta é proveniente de um evento Bernoulli ou a por meio da proporção de eventos $Y = 1$ em n ensaios.

Uma distribuição que pode ser utilizada para estudar uma variável resposta contínua e restrita ao intervalo $(0, 1)$ é a distribuição simplex (BANDORFF-NIELSEN; JORGENSES, 1991). A distribuição simplex faz parte dos modelos de dispersão (JORGENSEN, 1997), que estendem os modelos lineares generalizados.

Uma variável aleatória y que segue uma distribuição simplex com média $\mu \in (0, 1)$ e parâmetro de dispersão $\sigma^2 > 0$ tem função densidade dada por (2.59),

$$f\left(y; \mu, \sigma^2\right) = \frac{1}{\sqrt{2\pi\sigma^2 \left\{y\left(1-y\right)\right\}^3}} \exp\left\{-\frac{1}{2\sigma^2} d\left(y; \mu\right)\right\}, \; y \in \left(0, 1\right), \quad (2.59)$$

em que,

$$d\left(y; \mu\right) = \frac{\left(y-\mu\right)^2}{y\left(1-y\right)^2 \mu^2 \left(1-\mu\right)^2} \; . \quad (2.60)$$

A distribuição de y será denotada por $S\left(\mu, \sigma^2\right)$. A Figura 15 apresenta as diferentes densidades para y em diversos valores para os parâmetros μ e σ^2.

A variância de y é dada por

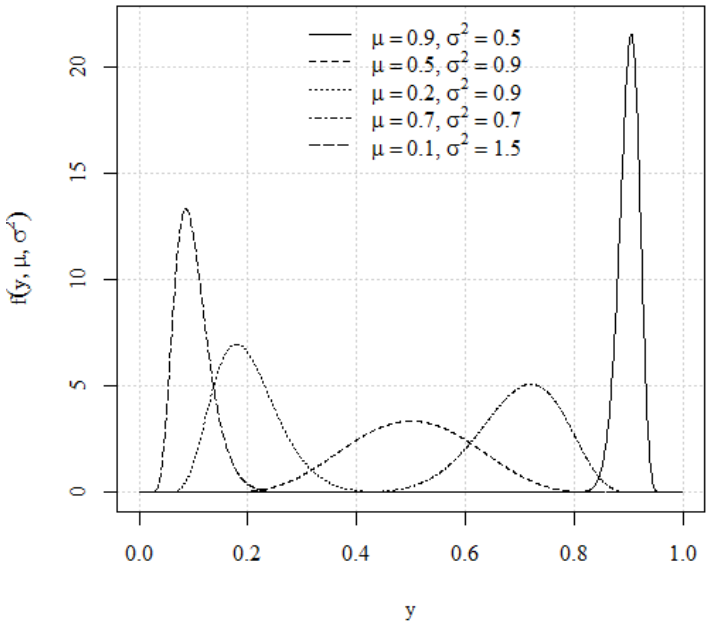


Figura 15 Ilustração da distribuição simplex em diferentes valores para os parâmetros μ e σ^2.

$$Var\left[Y\right] = \mu\left(1-\mu\right) - \frac{1}{\sigma\sqrt{2}}\exp\left\{\frac{1}{\sigma^2\mu^2\left(1-\mu\right)^2}\right\}\Gamma\left\{\frac{1}{2}, \frac{1}{2\sigma^2\mu^2\left(1-\mu\right)^2}\right\}, \tag{2.61}$$

em que $\Gamma\left\{a,b\right\}$ é a função gama incompleta definida por $\Gamma\left\{a,b\right\} = \int_b^\infty t^{a-1}e^{-t}dt$. A função de variância de y é dada por $V\left(\mu\right) = \mu^3\left(1-\mu\right)^3$.

Sejam $y_1, \ldots, y_n$ variáveis aleatórias independentes, com $y_i \sim S\left(\mu_i, \sigma^2\right)$, $i = 1, \ldots, n$. O modelo de regressão simplex é definido pela densidade da forma (2.59), sendo as médias μ_i modeladas por

$$g(\mu_i) = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ji} = \mathbf{x}_i^T \beta = \eta_i \tag{2.62}$$

em que $g(\cdot)$ é a função de ligação, estritamente monótona e duplamente diferenciável que transforma valores do intervalo $(0,1)$ nos reais, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)$ é o vetor dos parâmetros da regressão ($\boldsymbol{\beta} \in \mathbb{R}^p$), $\mathbf{x}_i^T = (x_{1i}, \ldots, x_{pi})$ são os valores conhecidos de p covariáveis e η_i é o preditor linear.

Pode-se observar que o modelo de regressão simplex acomoda variáveis respostas com variâncias não constantes, assim como o modelo de regressão beta, pois a variância de y_i é uma função de μ_i, definida em (2.61).

A seguir apresentaremos o procedimento para estimação dos parâmetros do modelo de regressão simplex. Considere uma amostra de n observações independentes. O logaritmo da função de verossimilhança é dado por

$$l\left(\boldsymbol{\beta}, \sigma^2\right) = \sum_{i=1}^{n} l_i\left(\mu_i, \sigma^2\right), \tag{2.63}$$

em que,

$$\begin{aligned} l_i\left(\mu_i, \sigma^2\right) &= -\frac{1}{2}\ln\left\{2\pi\sigma^2\left[y_i(1-y_i)\right]^3\right\} - \frac{1}{2\sigma^2} d\left(y_i; \mu_i\right) \\ &= -\frac{1}{2}\ln(2\pi) - \frac{1}{2}\ln\left(\sigma^2\right) - \frac{3}{2}\ln\left[y_i(1-y_i)\right] - \frac{1}{2\sigma^2} d\left(y_i; \mu_i\right), \end{aligned}$$

com $d(y_i; \mu_i)$ sendo o mesmo que a expressão (2.60) para a i-ésima observação. Derivando $l\left(\boldsymbol{\beta}, \sigma^2\right)$ em relação a β_i, $i = 1, \ldots, n$, temos

$$\frac{\partial l\left(\boldsymbol{\beta}, \sigma^2\right)}{\partial \beta_i} = \sum_{i=1}^{n} \frac{\partial l_i\left(\mu_i, \sigma^2\right)}{\partial \mu_i} \frac{d\mu_i}{d\eta_i} \frac{\partial \eta_i}{\partial \beta_i}, \tag{2.64}$$

em que

$$\frac{\partial l_i\left(\mu_i,\sigma^2\right)}{\partial \mu_i} = -\frac{1}{2\sigma^2} d'\left(y_i;\mu_i\right), \tag{2.65}$$

sendo $d'\left(y_i;\mu_i\right) = \frac{\partial d(y_i;\mu_i)}{\partial \mu_i}$.

Definindo

$$r_i = -\frac{1}{2} d'\left(y_i;\mu_i\right),\ i = 1,\ldots,n, \tag{2.66}$$

temos que

$$\frac{\partial l_i\left(\mu_i,\sigma^2\right)}{\partial \mu_i} = \sigma^{-2} r_i.$$

Assim, a função escore para $\boldsymbol{\beta}$ é dada por

$$\frac{\partial l\left(\boldsymbol{\beta},\sigma^2\right)}{\partial \beta_i} = \sum_{i=1}^{n} \sigma^{-2} r_i \frac{1}{g'\left(\mu_i\right)} x_{ii}, \tag{2.67}$$

em que, $\frac{d\mu_i}{d\eta_i} = \frac{1}{g'(\mu_i)}$ e $\frac{\partial \eta_i}{d\beta_i} = x_{ii}$. Em notação matricial, a função escore para $\boldsymbol{\beta}$ é dada por

$$U_{\boldsymbol{\beta}}\left[l\left(\boldsymbol{\beta},\sigma^2\right)\right] = \sigma^{-2}\mathbf{X}^T\mathbf{T}\mathbf{r}, \tag{2.68}$$

em que, $\mathbf{X}$ é uma matriz $n \times p$, cuja i-ésima linha é $\boldsymbol{x}_i^T$, $\mathbf{T} = diag\left\{\frac{1}{g'(\mu_1)},\ldots,\frac{1}{g'(\mu_n)}\right\}$ e $\mathbf{r} = (r_1,\ldots,r_n)$.

A função escore para σ^2 é obtida derivando $l\left(\boldsymbol{\beta},\sigma^2\right)$ em relação a σ^2,

$$\frac{\partial l\left(\boldsymbol{\beta},\sigma^2\right)}{\partial \sigma^2} = \sum_{i=1}^{n} \frac{\partial l_i\left(\mu_i,\sigma^2\right)}{\partial \sigma^2},$$

em que,

$$\frac{\partial l_i\left(\mu_i,\sigma^2\right)}{\partial\sigma^2}=-\frac{1}{2\sigma^2}+\frac{1}{2\sigma^4}d\left(y_i;\mu_i\right). \tag{2.69}$$

Logo, a função escore para σ^2 é dada por

$$U_{\sigma^2}\left[l\left(\boldsymbol{\beta},\sigma^2\right)\right]=-\frac{n}{2\sigma^2}+\frac{1}{2\sigma^4}\sum_{i=1}^{n}d\left(y_i;\mu_i\right) \tag{2.70}$$

A partir das segundas derivadas do logaritmo da função de verossimilhança definido em (2.61), pode-se obter a matriz de informação de Fisher. De (2.62), a segunda derivada de $l\left(\boldsymbol{\beta},\sigma^2\right)$ em relação a β_i e β_j é dada por

$$\begin{aligned}\frac{\partial^2 l\left(\boldsymbol{\beta},\sigma^2\right)}{\partial\beta_i\partial\beta_j}&=\sum_{i=1}^{n}\frac{\partial}{\partial\mu_i}\left(\frac{\partial l_i\left(\mu_i,\sigma^2\right)}{\partial\mu_i}\frac{d\mu_i}{d\eta_i}\frac{\partial\eta_i}{d\beta_i}\right)\frac{d\mu_i}{d\eta_i}\frac{\partial\eta_i}{d\beta_j}\\&=\sum_{i=1}^{n}\underbrace{\frac{\partial}{\partial\mu_i}\left(\frac{\partial l_i\left(\mu_i,\sigma^2\right)}{\partial\mu_i}\frac{d\mu_i}{d\eta_i}\right)}_{derivada\,do\,produto}\frac{d\mu_i}{d\eta_i}\frac{\partial\eta_i}{d\beta_j}x_{ii}\\&=\sum_{i=1}^{n}\left\{\frac{\partial^2 l_i\left(\mu_i,\sigma^2\right)}{\partial\mu_i^2}\frac{d\mu_i}{d\eta_i}+\frac{\partial l_i\left(\mu_i,\sigma^2\right)}{\partial\mu_i}\frac{\partial}{\partial\mu_i}\left(\frac{d\mu_i}{d\eta_i}\right)\right\}\frac{d\mu_j}{d\eta_j}x_{ii}x_{ij}.\end{aligned}$$

Pode-se mostrar que $E\left[\frac{\partial l_i\left(\mu_i,\sigma^2\right)}{\partial\mu_i}\right]=0$. Daí, temos que

$$E\left[\frac{\partial^2 l\left(\boldsymbol{\beta},\sigma^2\right)}{\partial\beta_i\partial\beta_j}\right]=\sum_{i=1}^{n}E\left[\frac{\partial^2 l_i\left(\mu_i,\sigma^2\right)}{\partial\mu_i^2}\right]\left(\frac{d\mu_j}{d\eta_j}\right)^2x_{ii}x_{ij}.$$

Da equação (2.65), obtemos

$$\frac{\partial^2 l_i\left(\mu_i,\sigma^2\right)}{\partial\mu_i^2}=-\frac{1}{2\sigma^2}d''\left(y_i;\mu_i\right),$$

em que, $d''(y_i;\mu_i) = \frac{\partial^2 d(y_i;\mu_i)}{\partial \mu_i^2}$.

Pode-se mostrar também que

$$\frac{1}{2}E\left[d''(y_i;\mu_i)\right] = \frac{3\sigma^2}{\mu_i(1-\mu_i)} + \frac{1}{\mu_i^3(1-\mu_i)^3}. \tag{2.71}$$

Dessa forma, temos que

$$E\left[\frac{\partial^2 l\left(\boldsymbol{\beta},\sigma^2\right)}{\partial\beta_i\partial\beta_j}\right] = -\frac{1}{\sigma^2}\sum_{i=1}^{n}\underbrace{\left\{\frac{3\sigma^2}{\mu_i(1-\mu_i)} + \frac{1}{\mu_i^3(1-\mu_i)^3}\right\}}_{a_i}\underbrace{\frac{1}{\left\{g'(\mu_i)\right\}^2}}_{\left(\frac{d\mu_j}{d\eta_j}\right)^2} x_{ii}x_{ij}. \tag{2.72}$$

Matricialmente, a expressão (2.72) é dada por

$$E\left[\frac{\partial^2 l\left(\boldsymbol{\beta},\sigma^2\right)}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T}\right] = -\frac{1}{\sigma^2}\boldsymbol{X}^T\boldsymbol{A}\boldsymbol{X}, \tag{2.73}$$

em que, $\boldsymbol{X}$ é a matriz de delineamento de ordem $n \times p$, $\mathbf{A} = diag\{a_1,\ldots,a_n\}$ e $a_i = \frac{3\sigma^2}{\mu_i(1-\mu_i)} + \frac{1}{\mu_i^3(1-\mu_i)^3}$.

A partir da expressão (2.64), pode-se obter a derivada de segunda ordem de $l\left(\boldsymbol{\beta},\sigma^2\right)$ em relação a $\boldsymbol{\beta}$ e σ^2, que é dada por

$$\begin{aligned}\frac{\partial^2 l\left(\boldsymbol{\beta},\sigma^2\right)}{\partial\beta_i\partial\sigma^2} &= \sum_{i=1}^{n}\frac{\partial}{\partial\sigma^2}\left[\sigma^{-2}r_i\frac{1}{g'(\mu_i)}x_{ii}\right] \\ &= -\frac{1}{\sigma^4}\sum_{i=1}^{n}r_i\frac{1}{g'(\mu_i)}x_{ii}.\end{aligned}$$

Como $E[r_i] = 0$, seque que

$$E\left[\frac{\partial^2 l\left(\boldsymbol{\beta},\sigma^2\right)}{\partial\beta_i\partial\sigma^2}\right]=0.$$

A segunda derivada em relação a σ^2 é obtida a partir da equação (2.69), logo

$$\begin{aligned}\frac{\partial^2 l\left(\boldsymbol{\beta},\sigma^2\right)}{\partial\sigma^4}&=\sum_{i=1}^{n}\frac{\partial l_i^2\left(\mu_i,\sigma^2\right)}{\partial\sigma^4}\\&=\sum_{i=1}^{n}\frac{\partial}{\partial\sigma^2}\left[-\frac{1}{2\sigma^2}+\frac{1}{2\sigma^4}d\left(y_i;\mu_i\right)\right]\\&=\frac{n}{2\sigma^4}-\frac{1}{\sigma^6}\sum_{i=1}^{n}d\left(y_i;\mu_i\right).\end{aligned}$$

Pode-se mostrar que $E\left[d\left(y_i;\mu_i\right)\right]=\sigma^2$, portanto

$$E\left[\frac{\partial^2 l\left(\boldsymbol{\beta},\sigma^2\right)}{\partial\sigma^4}\right]=-\frac{n}{2\sigma^4}. \tag{2.74}$$

Assim, a matriz de informação de Fisher, I_E, para $\left(\boldsymbol{\beta},\sigma^2\right)$ é dada por

$$I_E\left(\boldsymbol{\beta},\sigma^2\right)=\begin{pmatrix}E\left[-\frac{\partial^2 l(\boldsymbol{\beta},\sigma^2)}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T}\right] & E\left[-\frac{\partial^2 l(\boldsymbol{\beta},\sigma^2)}{\partial\beta_i\partial\sigma^2}\right]\\ E\left[-\frac{\partial^2 l(\boldsymbol{\beta},\sigma^2)}{\partial\beta_i\partial\sigma^2}\right] & E\left[-\frac{\partial^2 l(\boldsymbol{\beta},\sigma^2)}{\partial\sigma^4}\right]\end{pmatrix}$$

$$I_E\left(\boldsymbol{\beta},\sigma^2\right)=\begin{pmatrix}\frac{1}{\sigma^2}\mathbf{X}^T\mathbf{A}\mathbf{X} & 0\\ 0 & \frac{n}{2\sigma^4}\end{pmatrix}, \tag{2.75}$$

com $\mathbf{A}$ definido como em (2.72).

Os estimadores de máxima verossimilhança dos parâmetros $\boldsymbol{\beta}$ e σ^2 são

obtidos por meio das soluções das equações $U_{\boldsymbol{\beta}}\left(\boldsymbol{\beta}, \sigma^2\right) = 0$ e $U_{\sigma^2}\left(\boldsymbol{\beta}, \sigma^2\right) = 0$. Porém, somente o estimador para σ^2 possui forma fechada.

Seguindo essa ideia, igualando a zero a equação (2.70), o estimador de máxima verossimilhança para σ^2 é dado por

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} d\left(y_i; \hat{\mu}_i\right) \tag{2.76}$$

Para a estimação de $\boldsymbol{\beta}$, será necessário utilizar algum método de maximização numérica do logaritmo da função de verossimilhança definida em (2.63), como o método de Newton-Raphson, escore de Fisher ou quasi-Newton. A seguir, descreveremos o método de quasi-Newton. Este método é mais conhecido por BFGS, pois foi desenvolvido por Broyden, Fletcher, Goldfarb e Shanno (NOCEDAL; WRIGHT, 1999). Este método, com o mesmo princípio do algoritmo de Newton-Raphson, utiliza uma sequência de matrizes simétricas e positivas definidas $B^{(i)}$ em substituição da matriz $[\mathbf{I}_\theta]^{-1}$.

Denotando $U\left(\boldsymbol{\beta}\right) = U_{\boldsymbol{\beta}}\left(\boldsymbol{\beta}, \sigma^2\right)$, a forma iterativa para estimar $\boldsymbol{\beta}$ é dada por

$$\boldsymbol{\beta}^{(i+1)} = \boldsymbol{\beta}^{(i)} - \alpha^{(i)}\mathbf{B}^{(i)}U\left(\boldsymbol{\beta}^{(i)}\right),\ i = 0, 1, \ldots, \tag{2.77}$$

em que, $\alpha^{(i)}$ é um escalar determinado por algum procedimento de busca linear a partir de $\theta^{(i)}$ na direção $-B^{(i)}U\left(\theta^{(i)}\right)$ de forma que $f\left(y; \theta^{(i)}\right)$ cresça nessa direção. $B^{(i)}$ são matrizes simétricas e positivas definidas, dadas por

$$B^{(i+1)} = B^{(i)} - \frac{B^{(i)}s^{(m)}\left(s^{(m)}\right)^T B^{(i)}}{\left(s^{(m)}\right)^T B^{(i)}s^{(m)}} + \frac{y^{(m)}\left(y^{(m)}\right)^T}{\left(y^{(m)}\right)^T s^{(m)}},\ i = 0, 1, \ldots,$$

em que $s^{(i)} = \beta^{(i+1)} - \beta^{(i)}$ e $y^{(i)} = U\left(\beta^{(i+1)}\right) - U\left(\beta^{(i)}\right)$. Mais detalhes sobre o método também pode ser encontrado em Nocedal e Wright (1999).

O procedimento BFGS, descrito acima para estimar os parâmetros do modelo de regressão simplex, está implementado no software R por meio da função *optim*, que além desse otimizador, possui implementado também os otimizadores *Nelder-Mead*, gradiente conjugado e *simulating annealing*.

2.4.1 Técnicas de diagnóstico

Uma vez estimados os parâmetros de modelo de regressão simplex, existe o interesse em verificar a qualidade do ajuste do modelo corrente. A medida h_{ii}, o *leverage*, é utilizada para tal propósito. A matriz $\mathbf{H}$ de dimensão $n \times n$ para o modelo de regressão simplex é definida como

$$\mathbf{H} = \mathbf{A}^{1/2}\mathbf{X}\left(\mathbf{X}^T\mathbf{A}\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{A}^{1/2}, \tag{2.78}$$

em que $\mathbf{A} = diag\left\{a_1, \ldots, a_n\right\}$ e a_i é definido como em (2.72). Cada elemento de $\mathbf{A}$ depende de μ_i, logo os elementos de $\mathbf{A}$ correspondem aos valores ajustados $\hat{a}_i$ para cada observação.

Espinheira et al. (2008) e Ferrari e Cribari-Neto (2004) denominam o resíduo de um modelo de regressão beta como sendo padronizado ponderado. O mesmo é feito para um modelo de regressão simplex. Considerando σ^2 conhecido, o resíduo padronizado ponderado é definido da seguinte forma:

$$r_i^{pp} = \frac{\hat{r}_i}{\sqrt{\hat{s}_i\left(1 - \hat{h}_{ii}\right)}}. \tag{2.79}$$

Em que $\hat{r}_i = -\frac{1}{2}d'\left(y_i; \hat{\mu}_i\right)$, com $d'\left(y_i; \mu_i\right)$ definido em (2.65), $\hat{h}_{ii}$ é o i-ésimo

elemento da diagonal da matriz $\mathbf{H}$, $\hat{s}_i = V\hat{a}r\left(r_i\right) = \sigma^2\left[\frac{3\sigma^2}{\hat{\mu}_i(1-\hat{\mu}_i)} + \frac{1}{\hat{\mu}_i^3\left(1-\hat{\mu}_i^3\right)}\right]$ e $\hat{\mu}_i = g^{-1}\left(\mathbf{x}_i^T\hat{\boldsymbol{\beta}}\right)$. Na prática, como σ^2 é desconhecido, utilizamos a estimativa de máxima verossimilhança $\hat{\sigma}^2$.

2.5 Critérios de adequação de ajuste

Serão apresentados a seguir os critérios de adequação de ajuste utilizados neste trabalho.

2.5.1 Critérios de Informação de Akaike e Bayesiano

O Critério de informação de Akaike (AIC) proposto em Akaike (1974) é uma medida relativa da qualidade de ajuste de um modelo estatístico.

O AIC não é uma prova sobre o modelo, mas uma ferramenta útil na seleção de modelos. Para seu cálculo, não existe teste de hipóteses, significância e nem *valor-p*. É definido como:

$$AIC = -2\,l\left(\hat{\theta}\,|y\right) + 2p \tag{2.80}$$

em que $l\left(\hat{\theta}\,|y\right)$ é o logaritmo neperiano da função de verossimilhança do modelo em $\hat{\theta}$ e p é o número de parâmetros do modelo.

Schwarz (1978) propôs um critério conhecido como Critério de Informação Bayesiano (BIC), que corresponde à troca do fator 2, que é o peso do número de parâmetros, em (2.80) por $\ln(n)$, logo o BIC é dado por:

$$BIC = -2\,l\left(\hat{\theta}\,|y\right) + p\,\ln(n) \tag{2.81}$$

em que n é o número de observações da amostra.

Dado um conjunto de modelos ajustados aos dados, o modelo preferido é o que apresentar menor valor dos critérios acima, ou seja, quanto menor for o valor desses critérios melhor será o ajuste do modelo aos dados (AKAIKE, 1974).

Para modelos de regressão ajustados via algoritmo boosting, o cálculo do AIC e BIC é modificado para levar em consideração a função perda otimizada. Conforme Buhlmann e Hothorn (2007), o AIC é dado por:

$$AIC = 2\hat{\rho}\left(Y, g\right) + 2df\left(m\right) \tag{2.82}$$

em que $\hat{\rho}\left(Y, g\right)$ é a função perda avaliada na iteração m do algoritmo e $df\left(m\right)$ é o número de graus de liberdade na iteração m. De maneira similar, o BIC é dado por

$$BIC = 2\hat{\rho}\left(Y, g\right) + \ln\left(n\right) df\left(m\right). \tag{2.83}$$

A quantidade $df\left(m\right)$ não é obtida de maneira usual como em modelos de regressão. Vamos apresentar a ideia de graus de liberdade em algoritmo boosting iniciando pelo modelo mais simples, que é obtido quando utilizamos a perda quadrática e resulta no algoritmo L_2Boosting. Considere o procedimento base descrito em (2.26). Denote por

$$\mathbf{H}^{(j)} = \frac{\mathbf{X}^{(j)}\left(\mathbf{X}^{(j)}\right)^T}{\left\|\mathbf{X}^{(j)}\right\|^2},\ j = 1, \ldots, p,$$

a matriz chapéu (*hat*) de ordem $n \times n$ para o ajuste por mínimos quadrados utilizando a j-ésima variável preditora $\mathbf{X}^{(j)} = \left(X_1^{(j)}, \ldots, X_n^{(j)}\right)$ apenas. O denominador $\left\|x^{(j)}\right\|^2 = \mathbf{x^T x}$ denota a norma euclidiana para o vetor $\mathbf{x} \in \mathbb{R}^{\mathbf{n}}$ Então, a matrix chapéu do procedimento base em (2.26) é

$$\mathbf{H}^{(\hat{\lambda})} : (\mathbf{u_1}, \ldots, \mathbf{u_n}) \mapsto \hat{\mathbf{u}}_\mathbf{1}, \ldots, \hat{\mathbf{u}}_\mathbf{n},$$

ou seja, o vetor gradiente negativo contribui para a obtenção das estimativas dos betas que minimizam a soma de quadrados do resíduo, conforme pode ser visto na expressão (2.28). A matriz chapéu para o algoritmo L_2Boosting na iteração m é dada por:

$$\begin{aligned}\mathbf{B}^{(m)} &= \mathbf{B}^{(m-1)} + v\mathbf{H}^{(\hat{\lambda}_\mathbf{m})}\left(\mathbf{I} - \mathbf{B}^{(\mathbf{m-1})}\right) \\ &= \mathbf{I} - \left(\mathbf{I} - v\mathbf{H}^{(\hat{\lambda}_\mathbf{m})}\right)\left(\mathbf{I} - v\mathbf{H}^{(\hat{\lambda}_\mathbf{m-1})}\right)\cdots\left(\mathbf{I} - v\mathbf{H}^{(\hat{\lambda}_\mathbf{1})}\right),\end{aligned}$$

em que, $\hat{\lambda}_r \in \{1, \ldots, p\}$ denota o componente que é selecionado pelo procedimento base (2.26) na r-ésima iteração. Buhlmann e Hothorn (2007) salientam que $\mathbf{B}^{(m)}$ é dependente da variável resposta Y pelos componentes selecionados $\hat{\lambda}_r$, $r = 1, \ldots, m$. Por causa disso, $\mathbf{B}^{(m)}$ deve ser visto como uma aproximação da matriz chapéu e, conforme os mesmos autores, quando o algoritmo é finalizado em sua iteração ótima, a matriz $\mathbf{B}^{(m)}$ equivale à matriz $\mathbf{H}$ em estimação por mínimos quadrados ou máxima verossimilhança. Sendo assim, definem os graus de liberdade de um ajuste boosting na iteração m como

$$df\,(m) = tr\left(\mathbf{B}^{(m)}\right).$$

No caso de boosting, mesmo quando $v = 1$, $df\,(m)$ é diferente do que contar o número de variáveis que foram selecionadas na iteração m, uma vez que uma mesma variável pode ser selecionada em duas ou mais iterações do algoritmo. Diante do exposto, podemos estimar a variância do erro, $\sigma^2_\varepsilon = E\left[\varepsilon_i^2\right]$ em um modelo linear por

$$\sigma_{\varepsilon}^{2} = \frac{1}{n - df\,(m)} \sum_{i=1}^{n} \left(Y_i - \hat{f}^{(m)}\left(\mathbf{X}_i\right) \right)^2,$$

Além disso, podemos representar

$$\mathbf{B}^{(m)} = \sum_{j=1}^{p} \mathbf{B}^{(m)(j)}, \tag{2.84}$$

em que $\mathbf{B}^{(m)(j)}$ é a matriz chapéu (aproximada) que produz os valores ajustados para o j-ésimo preditor, ou seja, $\mathbf{B}^{(m)(j)}\mathbf{Y} = \mathbf{X}^{(j)}\hat{\beta}_j^{(m)}$. Assim, as matrizes $\mathbf{B}^{(m)(j)}$ de cada iteração podem ser obtidas de maneira iterativa pela seguinte atualização:

$$\begin{aligned} \mathbf{B}^{(m)\left(\hat{\lambda}_m\right)} &= \mathbf{B}^{(m-1)\left(\hat{\lambda}_m\right)} + v\mathrm{H}^{\left(\hat{\lambda}_m\right)} \left(\mathbf{I} - \mathbf{B}^{(m-1)} \right), \\ \mathbf{B}^{(m)(j)} &= \mathbf{B}^{(m-1)(j)}, \; para \; todo \; j \neq \hat{\lambda}_m. \end{aligned}$$

Então, temos uma decomposição do total de graus de liberdade em p termos:

$$\begin{aligned} df\,(m) &= \sum_{j=1}^{p} df^{(j)}\,(m), \\ df^{(j)}\,(m) &= tr\left(\mathbf{B}^{(m)(j)} \right). \end{aligned}$$

Os graus de liberdade individuais $df^{(j)}\,(m)$ são úteis para quantificar a complexidade da estimativa individual do coeficiente $\hat{\beta}_j^{(m)}$.

Tendo visto o processo de obtenção dos graus de liberdade em algoritmos boosting, podemos calcular o AIC e BIC, como apresentado nas equações (2.82) e (2.83). Além disso, uma boa iteração de parada m pode ser determinada.

No caso do algoritmo Binomial Boosting, ocorre uma modificação na derivação da matriz chapéu. Ao invés de (2.84), um argumento de linearização leva à seguinte recursão para uma matriz chapéu aproximada $\mathbf{B}^{(m)}$, assumindo $\hat{f}^{(0)}(\cdot) \equiv 0$:

$$\begin{aligned}
\mathbf{B}^{(1)} &= v\mathbf{W}^{(0)}\mathrm{H}^{\left(\hat{\lambda}_1\right)} \\
\mathbf{B}^{(m)} &= \mathbf{B}^{(m-1)} + 4v\mathbf{W}^{(m-1)}\mathrm{H}^{\left(\hat{\lambda}_m\right)}\left(\mathbf{I} - \mathbf{B}^{(m-1)}\right),\ m \geq 2, \\
\mathbf{W}^{(m)} &= diag\left\{\hat{\pi}^{(m)}(\mathbf{X}_i)\left(1 - \hat{\pi}^{(m)}(\mathbf{X}_i)\right)\right\};\ 1 \leq i \leq n.
\end{aligned}$$

Os graus de liberdade são dados por $df(m) = tr\left(\mathbf{B}^{(m)}\right)$. Logo, o AIC para a perda negativo da log-verossimilhança da binomial, definida em (2.19), fica dado por

$$AIC(m) = -2\sum_{i=1}^{n}\left[Y_i \ln\left(\hat{\pi}^{(m)}(\boldsymbol{X}_i)\right) + (1 - Y_i)\ln\left(1 - \hat{\pi}^{(m)}(\boldsymbol{X}_i)\right)\right] + 2df(m) \tag{2.85}$$

ou para o $BIC(m)$ com termo de penalidade $\ln(n)\, df(m)$.

2.5.2 Índice de Complexidade da Informação de Bozdogan (ICOMP)

Os critérios de informação apresentados na seção 2.5.1 são uma combinação entre uma medida de bondade de ajuste (como o negativo de duas vezes a log-verossimilhança) com uma medida de complexidade do modelo, como por exemplo, no AIC e BIC essa medida é dada em termos do número de parâmetros do modelo. Bozdogan (2000) fez críticas com relação à essas medidas de informação no sentido de que o termo que mede a complexidade (ou termo de penalidade) do modelo no AIC e BIC não leva em consideração a correlação entre

as variáveis, a linearidade e não linearidade dos parâmetros do modelo. Nesse sentido, propôs um critério, o índice de complexidade da informação de Bozdogan (ICOMP), que utiliza a matriz de informação de Fisher para compor o termo que mensura a complexidade do modelo, uma vez que ela contém informação sobre a correlação entre os parâmetros do modelo avaliado.

Por isso, o mesmo autor diz que critério ICOMP estende os critérios do tipo AIC, com uma penalidade devido ao aumento da complexidade do sistema, no sentido de que quanto maior o número de variáveis e a relação entre elas, mais complexo se torna o modelo. Em geral, para modelos univariados e multivariados o ICOMP é definido como

$$ICOMP = -2l\left(\hat{\theta}\,|y\right) + 2C\left[I^{-1}\left(\hat{\theta}\right)\right], \tag{2.86}$$

em que,

$$C\left[I^{-1}\left(\hat{\theta}\right)\right] = \frac{s}{2}\ln\left[tr\left(\frac{I^{-1}\left(\hat{\theta}\right)}{s}\right)\right] - \frac{1}{2}\ln\left[tr\left(I^{-1}\left(\hat{\theta}\right)\right)\right], \tag{2.87}$$

denota a complexidade da informação maximal de $I^{-1}\left(\hat{\theta}\right)$. Na equação 2.86, p é número de parâmetros, n o tamanho da amostra, $\hat{\theta}$ a estimativa dos parâmetros, $I^{-1}\left(\hat{\theta}\right) = V\hat{a}r\left[\hat{\theta}\right]$ a inversa da matriz de informação de Fisher, e $s = rank\left[I^{-1}\left(\hat{\theta}\right)\right]$.

Como dito, o critério ICOMP avalia a complexidade do modelo por meio da estrutura de correlação das estimativas dos parâmetros via $I^{-1}\left(\hat{\theta}\right)$. Os elementos da diagonal de $I^{-1}\left(\hat{\theta}\right)$ correspondem às variâncias estimadas dos parâmetros do modelo e indicam a sensibilidade dos parâmetros, enquanto que os elementos fora da diagonal são as covariâncias entre os parâmetros, as quais medem o grau

de colinearidade entre as colunas de $I^{-1}\left(\hat{\theta}\right)$ e revelam a dimensão da independência dos parâmetros. De acordo com esse critério, o melhor modelo dentro de um conjunto de modelos é o que minimiza o ICOMP (LANNING; BOZDOGAN, 2003).

Em estudos de simulação com modelos não lineares, seleção de variáveis em regressão multivariada e modelos de séries temporais, Bozdogan (2000) verificou que o critério ICOMP apresentou resultados superiores ao AIC em termos de seleção do verdadeiro modelo simulado.

2.5.3 Interpretação dos parâmetros e medindo o efeito dos componentes em experimentos de mistura

Em MLG´s, funções de ligação podem ser usados para interpretação dos parâmetros. Por exemplo, uma vez que uma chance estabelece uma relação com a função de ligação *logit*, a razão de chances é utilizada na interpretação dos parâmetros. Na regressão logística, β_i representa a mudança no logit que resultaria de uma mudança unitária em x_i, quando as outras variáveis estão fixas. Portanto, interpretações para os parâmetros usando a ligação *logit* é direta para os coeficientes exponenciados. Esses coeficientes exponenciados representam a razão de chances (HOSMER; LEMESHOW, 2000).

Na análise de experimentos de mistura com regressão logística, a razão de chances não pode ser usada como na regressão logística usual. Uma vez que a proporção dos componentes individuais x_i estão no intervalo entre 0 e 1, os outros componentes não podem se manter constantes devido às restrições dadas na expressão (2.1). Em outras palavras, se a quantidade do componente x_i aumenta, então a quantidade de todos os outros componentes diminuem, mas sua razão de

um para o outro permanece constante. Nesse caso, a função de ligação pode ser usada para interpretar os parâmetros. Se não existem restrições para os componentes da mistura como na expressão (2.2), β_i retorna o valor esperado da resposta de acordo com a escala *logit* do componente i. Isso é também a altura da superfície no i-ésimo vértice da região simplex. Similarmente, de acordo com a escala *logit*, os parâmetros β_i mostram o desvio na superfície que é definido pelo modelo de primeiro grau

$$logit\,(\pi) = \sum_{i=1}^{q} \beta_i x_i.$$

Se o β_i é positivo, acarreta um efeito sinergético sobre a resposta, caso contrário, ou seja, se o β_i é negativo, acarreta um efeito antagônico sobre a resposta.

Se existem restrições sobre os componentes da mistura como na expressão (2.2), o modelo com os componentes originais apenas mostra a superfície sobre a região experimental restrita. Nesse caso, uma interpretação direta sobre os parâmetros do modelo com os componentes originais não pode ser feita. Por outro lado, existe o interesse sobre o conhecimento do efeito de $\beta_{ij} x_i x_j$ (num modelo de segundo grau), ou outros termos, os quais são comumente adicionados aos termos lineares da mistura, sobre a resposta. Em situações do tipo, é mais adequado estudar o efeito de cada componente sobre a resposta do que interpretar os parâmetros. Nesse sentido, para estudar o comportamento de sistemas de mistura em mais detalhes será abordado o conceito de direção Cox (*Cox direction*) para os efeitos dos componentes sobre a resposta.

2.5.3.1 Direção Cox para gráfico de resposta traço

Gráficos de resposta traço, ou também denominados de *trace plots*, tem sido amplamente utilizados em mistura de experimentos para medir os efeitos dos componentes que compõem a mistura sobre a resposta na região experimental. Essencialmente, esses gráficos avaliam o comportamento da variável resposta conforme incrementos de cada componente ocorrem enquanto os outros componentes permanecem fixos. Contudo, como visto na seção anterior, em experimentos de mistura os componentes variam conjuntamente. Para contornar esse problema, os gráficos de traço são construídos utilizando a direção Cox (AKAY; TEZ, 2011).

A direção Cox do componente i é uma linha imaginária projetada da mistura de referência para o vértice $x_i = 1$. As proporções dos q componentes na mistura de referência é $\mathbf{c} = (c_1, c_2, \ldots, c_q)$, em que $\sum_{i=1}^{q} c_i = 1$. O ponto de referência $\mathbf{c}$ padrão é geralmente adotado como o centróide do experimento. Quando a proporção c_i do componente i é mudada por uma quantidade Δ_i na direção Cox, então a nova proporção se torna

$$x_i = c_i + \Delta_i. \tag{2.88}$$

Quando existem apenas restrições como na expressão (2.1) sobre os componentes de mistura, Δ_i varia no intervalo $[-c_i, 1 - c_i]$. Caso contrário, para restrições adicionais sobre os componentes como na expressão (2.2), Δ_i varia no intervalo $[L_i - c_i, U_i - c_i]$. As proporções dos $q - 1$ componentes restantes, resultante de c_i no i-ésimo componente, é

$$x_j = c_j \frac{1 - x_i}{1 - c_i},\ j = 1, 2, \ldots, q,\ j \neq i. \tag{2.89}$$

O fato de um incremento ocorrer em um componente x_i e ter os correspondentes compoentes em (2.89) pode ser explicado como se segue para o caso de $q = 3$. Considere uma mistura na região simplex (x_1, x_2, x_3), um ponto de referência $\mathbf{c} = (c_1, c_2, c_3)$ e os vértices $A = (1, 0, 0)$, $B = (0, 1, 0)$ e $C = (0, 0, 1)$. Considere ainda que um incremento Δ_1 ocorreu em x_1, ou seja, vamos obter a direção Cox para o componente x_1 com mistura de referência $\mathbf{c}$ em direção ao vértice A. A Figura 16 ilustra os incrementos Δ_i no ponto de referência para cada componente de mistura.

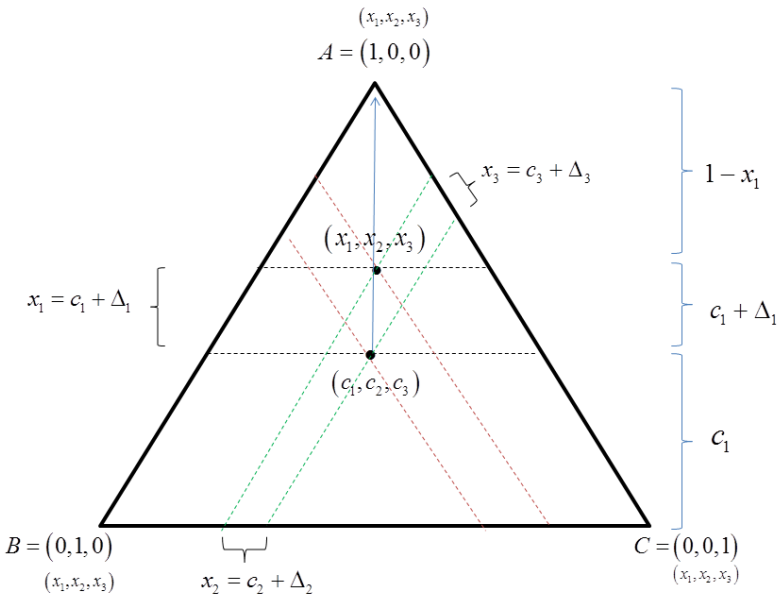


Figura 16 Ilustração da direção Cox para os incrementos Δ_i no ponto de referência $\mathbf{c} = (c_1, c_2, c_3)$ para cada componente de mistura.

Veja na Figura 16 que a partir de $x_1 = c_1 + \Delta_1$, todos os outros componentes sofreram alguma variação Δ. Note ainda que, para $x_1 = c_1 + \Delta_1$, três

segmentos são definidos, sendo um com comprimento $1 - x_1$, um com comprimento $c_1 + \Delta_1$ e outro com comprimento c_1. Considere agora outro componente de mistura, diga x_2, e, da mesma forma como foi feito para x_1, pode-se encontrar os segmentos $1 - x_2$, $1 - c_2$, x_2 e c_2. A relação entre esses segmentos e os obtidos por x_1 e c_1 são mostradas na Figura 17.

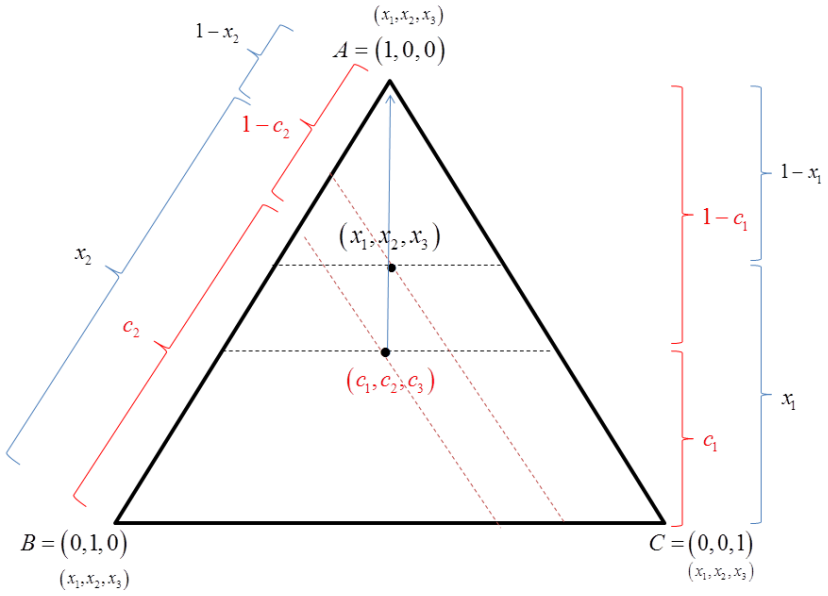


Figura 17 Ilustração da direção Cox considerando o componente de mistura x_2 e as relações entre os segmentos formados por esse componente, x_1 e o ponto de referência $\mathbf{c} = (c_1, c_2, c_3)$.

Para melhor visualização das relações entre os seguimentos referenciados, considere a Figura 18 que mostra apenas os segmentos envolvidos para os componentes x_1 e x_2. Note que, pelo Teorema de Tales [1], é possível estabelecer uma

[1] Se duas retas são transversais de um feixe de retas paralelas, então a razão(divisão) entre dois segmentos quaisquer de uma delas é igual à razão entre os segmentos correspondentes da outra.

relação entre os segmentos $1 - c_1$, c_2, $1 - x_1$ e x_2, ou seja,

$$\frac{x_2}{1 - x_1} = \frac{c_2}{1 - c_1} \Rightarrow x_2 = c_2 \frac{1 - x_1}{1 - c_1}.$$

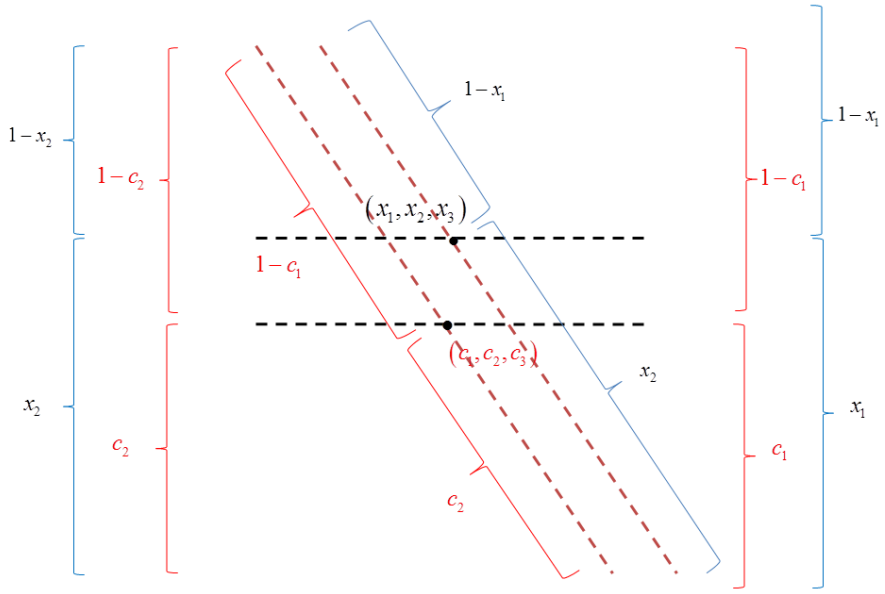


Figura 18 Ilustração da direção Cox mostrando apenas os segmentos envolvidos para os componentes x_1 e x_2, a fim de estabelecer uma relação entre os segmentos $1 - c_1$, c_2, $1 - x_1$ e x_2, pelo Teorema de Tales.

De maneira análoga, obtém-se

$$x_3 = c_3 \frac{1 - x_1}{1 - c_1}.$$

E o caso geral recai na equação (2.89). Note que a razão das proporções para os componentes j e k, em que x_j e x_k são definidos pela equação (2.89), é o mesmo valor que a razão dos componentes j e k na mistura referência, ou seja,

$\frac{x_j}{x_k} = \frac{c_j}{c_k}$. Em outras palavra, conforme a quantidade do componente x_i aumenta, a quantidade de todos os outros componentes diminui, mas sua razão de um para o outro permanece constante.

No caso de uma região experimental restrita ser um simplex regular, uma representação alternativa da direção Cox pode ser dada, considerando o fato de que $\frac{x_j}{x_k} = \frac{c_j}{c_k}$. Nesse caso, analogamente à razão das proporções dos componentes escolhidos ao longo dos eixos para um componente na direção Cox, com a ajuda da razão de componentes em qualquer ponto na região restrita, a mudança na resposta pode ser medida. Por exemplo, em um sistema de mistura com $q = 3$, ao longo que o eixo do componente x_i passa através do ponto de referência, os componentes x_j e x_k sendo $x_j/x_k = \rho_{x_i}$, são obtidos fazendo

$$\begin{aligned} \frac{x_j}{x_k} = \frac{c_j}{c_k} \Rightarrow x_j = \frac{c_j}{c_k} x_k \Rightarrow x_j = \frac{c_j}{c_k} c_k \frac{1-x_i}{1-c_i} \Rightarrow x_j = \rho_{x_i} c_k \frac{1-x_i}{1-c_i} \\ \Rightarrow x_j = \rho_{x_i} \frac{1-x_i}{\frac{c_j+c_k}{c_k}} \Rightarrow x_j = \frac{\rho_{x_i}(1-x_i)}{\rho_{x_i}+1}, \end{aligned} \tag{2.90}$$

em que, $1 - c_i = c_j + c_k$. Da mesma forma obtém-se x_k, daí

$$\begin{aligned} \frac{x_j}{x_k} = \frac{c_j}{c_k} \Rightarrow x_k = \frac{c_k}{c_j} x_j \Rightarrow x_k = \frac{c_k}{c_j} c_j \frac{1-x_i}{1-c_i} \Rightarrow x_k = \frac{1}{\rho_{x_i}} c_j \frac{1-x_i}{c_j+c_k} \\ \Rightarrow x_k = \frac{1}{\rho_{x_i}} \frac{1-x_i}{\frac{c_j+c_k}{c_j}} \Rightarrow x_k = \frac{1-x_i}{\rho_{x_i}+1}, \end{aligned} \tag{2.91}$$

em que, $L_i \leq x_i \leq U_i$ e ρ_{x_i} mostra a razão dos componentes exceto x_i no ponto referência.

Para fixar o conceito de direção Cox e a constância entre a razão de dois componentes, considere o seguinte exemplo numérico apresentado na Tabela 4

considerando o ponto de referência $\mathbf{c} = (0, 332, 0, 466, 0, 202)$ e incrementos para o componente x_1. Observe que $\rho_{x_1} = 2, 307 = \frac{c_2}{c_3} = \frac{x_2}{x_3}$ para qualquer incremento de $0, 1$ para x_1.

Tabela 4 Exemplo numérico da direção Cox para o componente x_1 com componentes x_2 e x_3 (denotados por x_2^* e x_3^*) obtidos pelas equações (2.90) e (2.91), considerando o ponto de referência $\mathbf{c} = (0, 332, 0, 466, 0, 202)$.

x_1	c_1	c_2	c_3	p_{x_1}	x_1	x_2^*	x_3^*	x_2^*/x_3^*
0	0,332	0,466	0,202	2,307	0	0,698	0,302	2,307
0,1	0,332	0,466	0,202	2,307	0,1	0,628	0,272	2,307
0,2	0,332	0,466	0,202	2,307	0,2	0,558	0,242	2,307
0,3	0,332	0,466	0,202	2,307	0,3	0,488	0,212	2,307
0,4	0,332	0,466	0,202	2,307	0,4	0,419	0,181	2,307
0,5	0,332	0,466	0,202	2,307	0,5	0,349	0,151	2,307
0,6	0,332	0,466	0,202	2,307	0,6	0,279	0,121	2,307
0,7	0,332	0,466	0,202	2,307	0,7	0,209	0,091	2,307
0,8	0,332	0,466	0,202	2,307	0,8	0,140	0,060	2,307
0,9	0,332	0,466	0,202	2,307	0,9	0,070	0,030	2,307
1	0,332	0,466	0,202	2,307	1,0	0,000	0,000	-

* Componentes de mistura obtidos de acordo com as equações (2.90) e (2.91).

Os pontos formados pelos componentes x_1, x_2 e x_3 na Tabela 4 formam uma linha na região simplex que começa no lado em que $x_1 = 0$ e, conforme incrementos de $\Delta = 0, 1$ ocorrem em x_1, os pontos correspondentes caminham em direção ao vértice $A = (1, 0, 0)$, em que $x_1 = 1$, como pode ser visto na Figura 19. Esses pontos formam a direção Cox. Analogamente pode-se fazer a direção Cox para os componentes x_2 e x_3.

Dessa forma, o valor da resposta predita para o preditor linear de primeiro grau, ao longo da direção Cox para o i-ésimo componente é dado por

$$logit\left(\hat{\pi}_{x_i}\right) = \hat{\beta}_i x_i + \hat{\beta}_j \frac{\rho_{x_i}\left(1 - x_i\right)}{\rho_{x_i} + 1} + \hat{\beta}_k \frac{1 - x_i}{\rho_{x_i} + 1},\ L_i \leq x_i \leq U_i. \qquad (2.92)$$

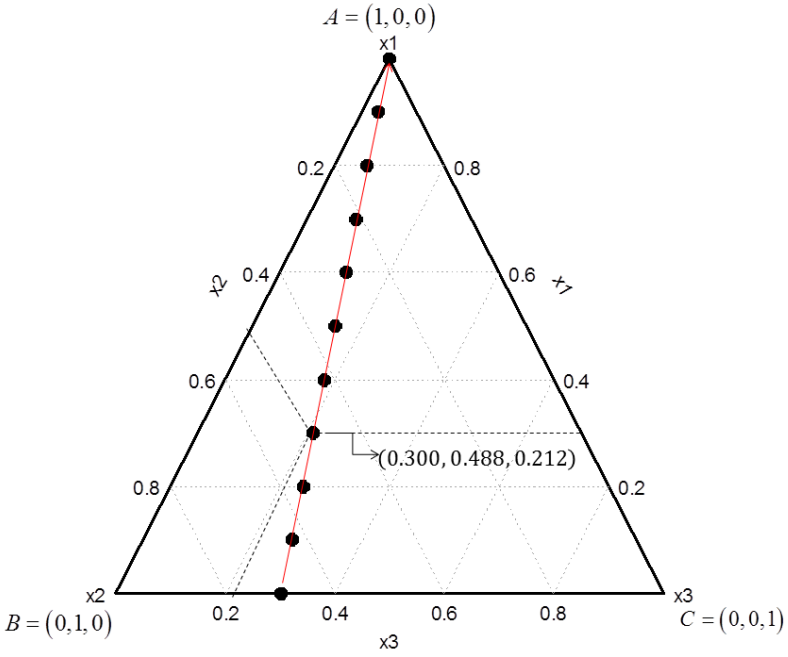


Figura 19 Ilustração da direção Cox com os pontos formados pelos componentes x_1, x_2^* e x_3^* (x_2^* e x_3^* obtidos de acordo com as equações (2.90) e (2.91)) na Tabela 4, os quais formam uma linha na região simplex com origem no lado em que $x_1 = 0$ e, conforme incrementos de $\Delta = 0,1$ ocorrem em x_1, os pontos correspondentes se direcionam ao vértice $A = (1, 0, 0)$, em que $x_1 = 1$.

Na equação (2.92), usando $logit\,(\hat{\pi}_{x_i})$, um traço resposta $\hat{\pi}_{x_i}$ para cada componente de mistura x_i é obtido. O traço resposta é um gráfico dos valores estimados da resposta, usando o modelo ajustado, ao longo da direção definido na equação (2.88). Esses traços representam o efeito da mudança de cada componente de mistura enquanto todos os outros componentes permanecem em razão constante. O modelo dado pela equação (2.92) pode ser expandido para diferentes preditores lineares.

2.5.3.2 Gráficos da razão de chances para os componente de mistura

A razão de chances é comumente utilizada para interpretação dos parâmetros em regressão logística por causa de sua fácil interpretação. A razão de chances pode ser utilizada também como uma medida relativa da chance de sucesso de um conjunto relativo a um outro conjunto. Em experimentos de mistura, técnicas gráficas baseadas em gráfico traço podem ser usadas para comparações do tipo. Considere um ponto qualquer $\mathbf{c} = (c_i, c_j, c_k)$ que é tomado como um grupo controle sobre a região experimental, a razão de chances é dada ao longo do eixo x_i por

$$\begin{aligned}\widehat{OR\,(x_i)} =& \frac{chance\, x_i}{chance\, controle} \\ =& \frac{\exp\left\{\hat{\beta}_i x_i + \hat{\beta}_j \frac{\rho_{x_i}(1-x_i)}{\rho_{x_i}+1} + \hat{\beta}_k \frac{1-x_i}{\rho_{x_i}+1}\right\}}{\exp\left\{\hat{\beta}_i c_i + \hat{\beta}_j c_j + \hat{\beta}_k c_k\right\}};\ L_i \leq x_i \leq U_i.\end{aligned}$$

De maneira mais simples, a equação acima pode ser reescrita como

$$\widehat{OR\,(x_i)} = \exp\left\{\hat{\beta}_i A + \hat{\beta}_j B + \hat{\beta}_k C\right\}, \tag{2.93}$$

em que $A = x_i - c_i$, $B = \frac{\rho_{x_i}(1-x_i)}{\rho_{x_i}+1} - c_j$ e $C = \frac{1-x_i}{\rho_{x_i}+1} - c_k$. Em uma comparação que toma o grupo controle como base, o grupo acima $chance\, x_i$ deve ser dado sobre diferentes eixos, uma vez que se os mesmos eixos são utilizados, intervalos de confiança não apropriados serão obtidos.

Em geral, a razão de chances é apenas uma estimativa pontual e assume valores entre zero e infinito. Se a razão de chances é igual a um, isso quer dizer que não existe diferença entre os grupos comparados, e se for diferente de um,

significa que existe diferença. Além disso, as razões preditas obtidas pela equação (2.93) formarão uma curva e a diferença entre os grupos comparados com a ajuda das curvas obtidas pode ser interpretada.

A precisão da razão de chances pode ser verificada pelo intervalo de confiança. A largura do intervalo de confiança reflete o tamanho da variabilidade inerente à razão de chances. Para calcular o erro padrão e o intervalo de confiança para a razão de chances, precisamos transformá-la para a escala logarítmica, que resulta em sua distribuição amostral sendo aproximadamente normal (HOSMER; LEMESHOW, 2000). Observe que a expressão para o logaritmo neperiano da razão das chances na equação (2.93) depende do componente de mistura x_i e o coeficiente estimado $\hat{\beta}_i$. O estimador do logaritmo neperiano da razão das chances é obtido trocando os parâmetros na equação (2.93) pelos seus estimadores. Utilizando métodos para calcular a variância de uma soma, obtemos o seguinte estimador:

$$\begin{aligned} \hat{Var}\left\{\ln\left[\widehat{OR\left(x_i\right)}\right]\right\} =& A^2\hat{Var}\left[\hat{\beta}_i\right] + B^2\hat{Var}\left[\hat{\beta}_j\right] + C^2\hat{Var}\left[\hat{\beta}_k\right] \\ &+ 2ABC\hat{ov}\left[\hat{\beta}_i, \hat{\beta}_j\right] + 2ACC\hat{ov}\left[\hat{\beta}_i, \hat{\beta}_k\right] \\ &+ 2BCC\hat{ov}\left[\hat{\beta}_j, \hat{\beta}_k\right]. \end{aligned} \tag{2.94}$$

Uma vez obtida a estimativa da variância do log da razão das chances estimado, podemos obter o estimador do intervalo de $100\,(1-\alpha)\,\%$ de confiança para o logaritmo neperiano da razão das chances, que é dado por

$$IC_{(1-\alpha)}\left\{\ln\left[\widehat{OR\left(x_i\right)}\right]\right\} = \left[\hat{\beta}_i A + \hat{\beta}_j B + \hat{\beta}_k C\right] \pm z_{(\frac{\alpha}{2})}\sqrt{\hat{Var}\left\{\ln\left[\widehat{OR\left(x_i\right)}\right]\right\}}, \tag{2.95}$$

em que, $\sqrt{\hat{Var}\left\{\ln\left[\widehat{OR\left(x_i\right)}\right]\right\}}$ é o erro padrão do estimador do logaritmo ne-

periano da razão de chances, $z_{(\frac{\alpha}{2})}$ é o $(\frac{\alpha}{2})$-ésimo quantil da distribuição normal padrão com α nível de significância.

Como visto na seção 2.2.1.2, o algoritmo Gradiente Boosting de Friedman depende apenas do vetor gradiente do modelo e, dessa forma, as variâncias e covariâncias dos parâmetros do modelo não podem ser obtidas. Para contornar esse problema, foi obtido o erro padrão do estimador do logaritmo neperiano da razão de chances pelo método de Monte Carlo, como descrito em Rizzo (2007). Para tal, considere θ como sendo o valor do logaritmo neperiano da razão de chances e $\hat{\theta}$ o seu estimador. O intervalo de confiança desejado foi obtido seguindo-se os seguintes passos:

(i) Gerar uma amostra de tamanho n da distribuição Simplex com as estimativas de Máxima Verossimilhança $\hat{\mu}$ e $\hat{\sigma}^2$. Para o modelo de mistura do tipo Scheffé, simular a amostra por meio do delineamento Centróide Simplex e para o modelo do tipo razão simular a amostra por meio do delineamento Vértice Extremo. Mais detalhes desses delineamentos podem ser vistos em Cornell (2002);

(ii) Da amostra simulada em (i), calcular $\hat{\theta}$;

(iii) Repetir os passos (i) e (ii) B vezes;

(iv) A partir do vetor $\hat{\boldsymbol{\theta}}^* = \left(\hat{\theta}^*_{(1)} \leq \hat{\theta}^*_{(2)} \leq \ldots \leq \hat{\theta}^*_{(B)}\right)$, para algum nível de significância α, o intervalo de confiança Monte Carlo $100 \times (1-\alpha)\,\%$ é dado por

$$IC_{(1-\alpha)}(\theta) = \left[\hat{\theta}^*_{(k_1)};\, \hat{\theta}^*_{(k_2)}\right], \tag{2.96}$$

em que $k_1 = (B+1)(\alpha/2)$ e $k_2 = (B+1)(1-\alpha/2)$ são os maiores inteiros não maiores que $(B+1)(\alpha/2)$ e $(B+1)(1-\alpha/2)$, respectivamente e $\hat{\theta}^*_{(k_1)}$ é o percentil $100\,(\alpha/2)\,\%$ e $\hat{\theta}^*_{(k_2)}$ o percentil $100\,(1-\alpha/2)\,\%$.

Os limites inferiores e superiores do estimador do intervalo de confiança para a razão de chances são obtidos exponenciando os limites nas equações (2.95) e (2.96). São preferidos os modelos cujos intervalos de confiança são mais estreitos. Portanto, métodos gráficos baseados em intervalos de confiança para a razão de chances podem ser utilizados para comparar os modelos.

Finalizando a metodologia proposta, para obtenção dos resultados serão utilizados os pacotes estatísticos *mboost*, *bbmle*, *VGAM* e *mixexp* do Sistema Computacional Estatístico R (R CORE TEAM, 2015), para realização das análises.

3 CONCLUSÃO

Na perspectiva de proporcionar o embasamento teórico para o desenvolvimento do que foi exposto nos artigos que compõem este trabalho, na primeira parte da tese realizou-se uma explanação considerando alguns tópicos que constituem a base teórica para a estatística. Para essa fundamentação foi organizada uma abordagem sobre os experimentos de mistura e os principais modelos de regressão utilizados. A distribuição simplex, pertencente à classe dos modelos lineares generalizados, representa a contribuição desta tese aos experimentos de mistura, bem como o emprego de algoritmo boosting. Ao final descreveu-se os critérios que possibilitarão as comparações dos modelos, tais como o AIC, BIC, ICOMP, gráficos de envelope simulado para os resíduos e gráficos das razões de chances. O referencial teórico apresentado possui a função de validar o construto da pesquisa, estabelecendo relações que serão consideradas na discussão dos resultados e na conclusão final do trabalho.

REFERÊNCIAS

AKAIKE, H. A new look at the statistical model identification, **IEEE Transactions on Automatic Control**, Boston, v. 19, n. 6, p. 716-723, 1974.

AKAY, K. U.; TEZ, M. Alternative modeling techniques for the quantal response data in mixture experiments, **Journal of Applied Statistics**, v. 38, n. 11, p. 2597 - 2616, 2011.

ATKINSON, A. C. **Plots, Transformations and Regression**, Oxford University Press, Oxford, 1985.

AITCHISON, J.; BACON-SHONE, J. Log contrast models for experiments with mixtures, **Biometrika**, v. 71, s. 2, p. 323 - 330, 1984.

BARTLETT, P.; TRASKIN, M. AdaBoost is consistent, **Journal of Machine Learning Resources**, v. 8, p. 2347 - 2368, 2007.

BECKER, N. G. Models for the response of a misture, **Journals of the Royal Statistical Society**, v. 30, p. 349 - 358, 1968.

BERK, R. A. **Statistical Learning from a Regression Perspective**, Springer Series in Statistics, 373 p., 2008.

BISHOP, C. M. **Neural Networks for Pattern Recognition**, Oxford University Press, 504 p., 1995.

BOX, G. E. P; DRAPER, N. E. **Response surfaces, mixtures, and ridge analyses**, Wiley Series in Probability an Statistics, 2º ed., 874 p., 2007.

BOZDOGAN, H. Akaike's Information Criterion and recent developments in Information Complexity, **Journal of Mathematical Psychology**, v. 44, p. 62 - 91, 2000. doi:10.1006/jmps.1999.1277

BREIMAN, L. et al. **Classification and regression Trees**, Chapman and Hall/CRC, 368 p., 1º ed., 1984.

BREIMAN, L. Arcing classifiers (with discussion), **The Annals of Statistics**, v. 26, n.3, p. 801 - 849, 1998.

BREIMAN, L. Prediction games and arcing algorithms, **Neural Computation**, v. 11, p. 1463 - 1517, 1999.

BUHLMANN, P.; HOTHORN, T. Boosting Algorithms: Regularization, Prediction and Model Fitting, **Statistical Science**, v. 22, n. 4, p. 477-505, 2007.

BUHLMANN, P.; YU, B. Boosting with the L_2 loss: Regression and classification, **Journal of the American Statistical Association**, v. 98, p. 324-338, 2003.

CAI, Y. D. et al. Using LogitBoost classifier to predict protein structural classes, **Journal of Theoretical Biology**, v. 238, p. 172-176, 2006.

CAO, D. S. et al. The Boosting: A new idea of building models, **Chemometrics and Intelligent Laboratory Systems**, v. 100, p. 1-11, 2010.

CASELLA, G.; BERGER, R. L. **Statistical Inference**, ed. 2, Cengage Learning , ISBN 9780495391876, 660 p., 2008.

CHEN, J. J.; LI, L. A.; JACKSON, C. D. Analysis of quantal response data from mixture expreriments, **Environmetrics**, v. 7, n. 5, p. 503-512, 1996.

CORNELL, J. A. **Experiments with Mixtures: Designs, Models, and the Analysis of Mixture Data**, 3º ed., John Wiley Sonc Inc., USA, 2002.

DETTLING, M.; BUHLMANN, P. Boosting for tumor classification with gene expression data, **Bioinformatics**, v. 19, n. 9, p. 1061-1069, 2003.

DIETTERICH, T. G.; ASHENFELTER, T. D. A.; BULATOV, Y., Training conditional random fields via gradient tree boosting, in **Proceedings of the 21st International Conference on Machine Learning (ICML)**, (Banff, AB), 2004. Disponível online em: `http://citeseerx.ist.psu.edu/viewdoc/`. doi=10.1.1.58.6703

DRAPER, N. R.; ST JOHN, R. C. A mixture model with inverse terms. **Technometrics**, v. 19, n. 1, p. 37 - 46, 1977.

ESPINHEIRA, P. L.; FERRARI, S. L. P.; CRIBARI-NETO, F. On beta regression residuals. **Journal of Applied Statistics**, Routledge, v. 35, n. 4, p. 407 - 419, 2008.

FERRARI, S. L. P.; CRIBARI-NETO, F. Beta regression for modelling rates and proportions. **Journal of Applied Statistics**, Taylor and Francis Group, v. 31, n. 7, p. 799 - 815, 2004.

FERREIRA, D. F. **Estatística Computacional em Java**. Editora UFLA, Lavras, ed. 1, 695 p., 2013.

FREUND, Y.; SCHAPIRE, R. E. Experiments with a new Boosting algorithm, **In: International Conference on Machine Learning.**, p. 148-156, 1996.

FRIEDMAN, J. Greedy function aproximation: A gradient boosting machine, **The Annals of Statistics**, v. 29, p. 1189 - 1232, 2001.

FRIEDMAN, J. H.; HASTIE, T. J.; TIBSHIRANI, R. J. Additive logistic regression: A statistical view of Boosting (with discussion), **The Annals of Statistics**, v. 28, p. 337 - 407, 2000.

FRIEDMAN, J. H.; HASTIE, T. J.; TIBSHIRANI, R. J. **The Elements of Statistical Learning**, Basel: Springer Verlag, 2001.

HANLEY, J. A. Receiver operating characteristic (ROC) methodology: the state of the art, **Critical Reviews in Diagnostic Imaging**, v. 29, n.3, p. 307 - 335,

1989.

HOFNER, B.; MAYR, A.; ROBINZONOV, N.; SCHMID, M. Model-based boosting in R: a hands-on tutorial using the R package mboost, **Computation Statistics**, v. 29, p. 3 - 35, 2014. doi 10.1007/s00180-012-0382-5

HOSMER, D. W.; LEMESHOW, S.; STURDIVANT, R. X. **Applied Logistic Regression**, 3º ed., John Wiley, New York, 528p., 2013.

JIANG, L. **Process consistency for adaboost**, Technical Report 05, Department of Statistics, Northwestern University, 2000.

KEARNS, M.; VALIANT, L. Crystographic limitations on learning Boolean formulae and finite automata, **Journal Assoc. Comput. Machinery**, v. 41, p. 67 - 95, 1994.

LANNING, M. J.; BOZDOGAN, H. Ordinal Logistic modeling using ICOMP as a goodness-of-fit criteria, Statistical Data Mining and Knowledge Discovery, Chapman Hall/CRC, USA, p. 353-371, 2003.

LISKA, G. R. **Classificação de dados em modelos com resposta binária via algoritmo boosting e regressão logística**. 2012. 1055 f. Dissertação (Mestrado em Estatística e Experimentação Agropecuária) - Departamento de Ciências Exatas da Universidade Federal de Lavras, Lavras, 2012.

MICHIE, D.; SPIEGELHALTER, D. J.; TAYLOR, C. C. **Machine Learning: Neural and Statistical Classification**, Ellis Horwood Series in Artificial Intelligence, 290 p., 1994.

NATEKIN, A.; KNOLL, A. Gradient boosting machines, a tutorial, **Frontier in Neurorobotics**, v. 7, p. 1-21, 2013.

NELDER, J. A.; WEDDERBURN, R. W. M. Generalized linear models, **Journal of the Royal Statistical Society Series A**, v. 135, n. 3, p. 370-384, 1972.

NOCEDAL, J.; WRIGHT, S. J. **Numerical optimization**, Springer-Verlag, New York, 1999.

NUNES, J. A. R; MORAIS, A. R.; BUENO FILHO, J. S. S. Modelagem da superdispersão em dados por um modelo linear generalizado misto, **Revista da Matemática e Estatística**, v. 22, n. 1, p. 55-70, 2004.

OLIVEIRA, M. S. **Um modelo de regressão beta: teoria e aplicações**, Dissertação de Mestrado, IME-USP, São Paulo, 2004.

PAULA, G. A. Influence and residuals in restricted generalized linear models, **Journal of Statistical Computation and Simulation**, v. 51, p. 315 - 352, 1995.

PAULA, G. A. **Modelos de regressão com apoio computacional**. São Paulo: Instituto de Matemática e Estatística, USP, 440 p., 2013.

PAULA, G. A.; TUDER, R. M. Utilização da regressão logística para aperfeiçoar o diagnóstico de processo infeccioso pulmonar, **Revista Ciência e Cultura**, v. 40, p. 1046-1050, 1986.

PIEPEL, G. F. Measuring component effects in constrained mixture experiments. **Technometrics**, v. 24, p. 29-39, 1982.

PREGIBON, D. Logistic regression diagnostics, **Annals of Statistics**, v. 9, p. 705-724, 1981.

R CORE TEAM (2015). **R: A language and environment for statistical computing**, R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/.

RIPLEY, B. D. **Pattern Recognittion and Neural Networks**, Cambridge University Press, ISBN 0 521 46086 7, 416 p., 1996.

RIZZO, M. L. **Statistical Computing with R**. ISBN 9781584885450, Chapman & Hall/CRC, 416 p., 2007.

RODRIGUES, T. B.; MACRINI, J. L. R.; MONTEIRO, E. C. Seleção de Variáveis e Classificação de Padrões por Redes Neurais como auxílio ao diagnóstico de Cardiopatia Isquêmica, **Pesquisa Operacional**, v. 28, n. 2, p. 285-302, 2008.

RUBESAM, A. **Estimação Não Paramétrica Aplicada a Problemas de Classificação via Bagging e Boosting**. 2004. 127 f. Dissertação (Mestrado em Estatística) - Instituto de Matemática, Estatística e Computação Científica da Universidade Estadual de Campinas, Campinas, 2004.

SCHAPIRE, R. E. The strength of weak learnability, **Machine learning**, v. 5, p. 197-227, 1990.

SCHAPIRE, R. E.; FREUND, Y. **Boosting: Foundations and Algorithms**, Massachusetts Institute of Technology, ISBN 9780262310413, 544p., 2012.

SCHEFFÉ, H. Experiments with mixtures, **Journals of the Royal Statistical Society**, v. 20, p. 344 - 360, 1958.

SCHMID, M.; WICKLER, F.; MALONEY, K. O.; MITCHELL, R.; MAYR, A. Boosted Beta regression, **PLOS ONE**, v. 8, n. 4, p. 1-15, 2013.

SCHWARZ, G. Estimating the dimensional of a model, **The Annals of Statistics**, Hayward, v. 6, n. 2, p. 461-464, 1978.

SCHONLAU, M. Boosted regression (Boosting): An introductory tutorial and a Stata plugin, **The Stata Journal**, v. 5, n. 3, p. 330-354, 2005.

SNEE, S. H. Techniques for the analysis of mixture data, **Technometrics**, v. 15, s. 3, p. 517 - 528, 1973.

VIOLA, P.; JONES, M., Rapid object detection using a boosted cascade of simple features, in **Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2001**, (Kauai, HI), 2001. doi: 10.1109/CVPR.2001.990517

WEDDERBURN, R. W. M. On the existence and uniqueness of the maximum likelihood estimates for certain generalized linear models, **Biometrika**, v. 68, p. 27-32, 1976.

WILLIAMS, D. A. Generalized linear model diagnostic using the deviance and single case deletion, **Applied Statistics**, v. 36, p. 181-191, 1987.

WILLIAMS, D. A. Residuals in generalized linear models, **In: Proceedings of the 12th. International Biometrics Conference**, Tokyo, p. 59-68, 1984.

SEGUNDA PARTE - ARTIGOS

ARTIGO 1 - Construção do intervalo de confiança para razão de chances utilizando Regressão Simplex em um Experimento de Mistura.

ARTIGO 2 - Acurácia e precisão de razões de chances obtidas por meio do modelo Boosted Simplex Regression aplicado em Experimentos de Mistura.

ARTIGO 1

Construção do Intervalo de Confiança para Razão de Chances Utilizando Regressão Simplex em um Experimento de Mistura

Versão preliminar de artigo - Sujeito a alterações pelo corpo editorial da revista

Revista: Statistical Methodology (IF: 0,637)

LAVRAS - MG

2016

RESUMO

Experimentos de mistura são frequentemente utilizados em muitos campos da ciência, como por exemplo em química, farmácia e indústria de produtos para consumidores, nos quais a resposta a ser analisada está em função de uma combinação de variáveis denominadas de componentes de mistura. Esses experimentos caracterizam-se por apresentar colinearidade e, quando a resposta é proporção, pode ocorrer o efeito da sub/super-dispersão. Nesse contexto, esse trabalho propõe uma nova abordagem na análise de dados em experimentos de mistura utilizando a distribuição simplex pertencendo a classe dos modelos lineares generalizados, denominado por regressão simplex. As vantagens desta nova abordagem é ilustrada em um experimento utilizado para estudar o efeito de diferentes dietas compostas por gordura, carboidrato e fibra sobre a expressão de tumor nas glândulas mamárias em ratos fêmeas. A comparação do modelo proposto com os existentes na literatura foi feita mediante os critérios de seleção de modelos AIC, BIC e ICOMP, gráficos de envelope simulado para os resíduos dos modelos ajustados, gráficos da razão de chances e seu respectivo intervalos de confiança. Concluiu-se que o modelo de regressão simplex apresentou melhor qualidade de ajuste e produziu intervalos de confiança para a razão de chances mais precisos.

Palavras-chave: Modelo Linear Generalizado, Proporção, Modelo de Mistura, Região Simplex.

1 INTRODUÇÃO

Um experimento de mistura consiste em otimizar uma variável resposta com a restrição de que

$$\sum_{i=1}^{q} x_i = 1, \tag{1.1}$$

em que x_i ($0 \leq x_i \leq 1$) representa a proporção do i-ésimo componente ($i = 1, 2, \ldots, q$), com q o número de componentes (BOX; DRAPER, 2007). Caso um componente tenha proporção igual a um, ou seja $x_i = 1$, os demais componentes terão proporção iguais a zero e a referida mistura recebe o nome de *mistura pura* para um determinado componente. O espaço que compõe os q componentes assume a forma de um simplex regular de dimensão $(q-1)$. No caso $q = 3$, a região simplex é um espaço de mistura triangular. Por questões econômicas e, ou, físicas, às vezes são impostas restrições adicionais sobre os componentes individuais

$$0 \leq L_i \leq x_i \leq U_i \leq 1;\ i = 1, 2, \ldots, q, \tag{1.2}$$

sendo L_i e U_i, respectivamente, os limites inferiores e superiores e a restrição (1.2) reduz a região restrita dada na equação (1.1).

Diante do exposto, a modelagem estatística é feita utilizando modelos polinomiais assumindo normalidade para a variável resposta (CORNELL, 2002). No caso de variáveis seguindo outras distribuições, tais modelos são adaptados utilizando-se modelos lineares generalizados (MLG). Especialmente, quando a variável resposta é binária ou binomial, o modelo de regressão binomial (logística) tem sido bastante utilizado, mas esse modelo não acomoda o efeito da sub ou super-dispersão que frequentemente ocorre em dados agrupados. Para tal, esse

trabalho propõe o uso do modelo de regressão simplex, que é um modelo da classe dos MLG com a vantagem de modelar a sub ou super-dispersão de dados binomiais.

A sub ou super-dispersão é caracterizada quando a variância observada é inferior ou superior, respectivamente, à variância esperada do modelo, o que influencia diretamente nas estimativas do modelo ajustado causando a deflação ou inflação da resposta predita (MYERS; MONTGOMERY, 1997). Mais detalhes dessa abordagem serão apresentados na seção 2.3.

Especialmente, quando a variável resposta é binária ou binomial, uma medida de grande interesse prático é a razão de chances e, nesse caso, os métodos convencionais de análise e interpretação dos parâmetros de um modelo de mistura não são adequados, uma vez que quando um componente é modificado, os outros componentes são alterados, devida a restrição unitária mencionada em 1.1 (AKAY; TEZ, 2007). Para esse problema, a análise do efeito dos componentes de mistura consiste no uso da Direção Cox, que é abordada na seção 2.5.1. Além disso, o conceito de Direção Cox permite a obtenção dos intervalos de confiança para a razão de chances em experimentos de mistura. A largura desses intervalos representa a precisão da estimativa de razão de chances e a precisão desses intervalos é diretamente afetada pela presença do efeito da colinearidade.

Os modelos de mistura são altamente afetados pela colinearidade, que é causada pela restrição em (1.1). Várias alternativas foram propostas na literatura para contornar esse problema, como por exemplo o uso de pseudo-componentes, termos inversos ou variáveis razão (AKAY; TEZ, 2011). Dentre elas, o uso de variáveis razão no preditor linear do modelo tem proporcionado maior redução da covariância entre os parâmetros do preditor linear do modelo avaliado e, consequentemente, maior redução no efeito da colinearidade.

Diante do exposto, o presente trabalho tem por objetivo empregar o uso do modelo de regressão simplex na análise de experimento de mistura em comparação ao uso da regressão logística, o qual é bastante comum na análise de variável resposta com distribuição binomial, também caracterizada por proporções ou resposta limitada. Nesse sentido, as vantagens da abordagem proposta foram ilustradas no experimento que consistiu em estudar o efeito de diferentes dietas compostas por gordura, carboidrato e fibra sobre a expressão de tumor nas glândulas mamárias em ratos fêmeas. Para tal, foram estimadas as razões de chances e os seus respectivos intervalos de confiança de acordo com um grupo de referência e controle para mensurar o efeito da resposta conforme um dos componentes de mistura sofre incrementos. Além disso, foram utilizados os critérios de seleção de modelos AIC, BIC e ICOMP, bem como os gráficos de envelope simulado para os resíduos dos modelos ajustados a fim de determinar a distribuição de probabilidade da resposta.

2 MATERIAL E MÉTODOS

2.1 Descrição do Experimento

Foram utilizados os dados disponibilizados por Akay e Tez (2011). Os dados de mistura são referentes a ocorrência de tumor em ratos. O experimento de mixtura foi conduzido para estudar os efeitos de calorias da dieta composta por gordura, carboidrato e fibra sobre a expressão (promoção) de tumor nas glândulas mamárias induzido por *Dimetil-benzathracene* (DMBA) em ratos fêmeas. Nesse experimento, ratos da raça Sprague-Dwaley de 38 a 42 dias de idade foram aleatoriamente atribuídos a nove grupos, sendo 30 animais por grupo. Ao dia 52, os animais foram administrados com 7,5 mg de DMBA em óleo de milho por um

tubo ao estômago. Uma semana depois do tratamento com DMBA, os animais em cada grupo foram alimentados com suas correspondentes dietas em igual quantidade de calorias totais, mas com diferentes níveis de gordura, carboidrato e fibra. As dietas foram administradas durante 26 semanas, o tempo total de duração do experimento.

Tabela 1 O número observado de tumores induzidos DMBA em glândulas mamárias de ratos tratados com diferentes proporções calóricas de fibra, gordura e carboidrato (componentes originais). Cada grupo é composto por 30 ratos e tem igual quantidade de calorias totais.

Grupo	Componentes Originais			Ratos c/ Tumor	Taxa de Tumor
	Gordura (x_1)	Carboidrato (x_2)	Fibra (x_3)		
1	0,175	0,775	0,050	17	0,567
2	0,153	0,820	0,027	15	0,500
3	0,133	0,863	0,004	17	0,567
4	0,491	0,470	0,039	24	0,800
5	0,440	0,538	0,022	21	0,700
6	0,390	0,607	0,003	23	0,767
7	0,701	0,267	0,032	18	0,600
8	0,638	0,343	0,019	23	0,767
9	0,576	0,421	0,003	26	0,867

Os nutrientes gordura, carboidrato e fibra foram administrados sobre todo o alcance das dietas que poderiam ser consideradas como sendo fisiológicas. Essas dietas tiveram uma combinação de baixo, médio e alto níveis de cada um dos três nutrientes, num total de nove grupos de teste. A proporção das três dietas foram restritas pelos seguintes limites inferiores e superiores: $0,133 \leq Gordura \leq 0,730$; $0,267 \leq Carboidrato \leq 0,864$ e $0,003 \leq Fibra \leq 0,600$.

A Tabela 1 contém as respostas das proporções de tumor observadas de nove grupos de dieta com diferentes proporções calóricas de gordura (x_1), carboidrato (x_2) e fibra (x_3). Na Tabela 1, as dietas 1 a 3 são constituídas de baixa

gordura e alto carboidrato, as dietas 4 a 6 são constituídas de valores médios de gordura e carboidrato, as dietas 7 a 9 são possuem alta gordura e baixo carboidrato. As dietas 1, 4 e 7 são altas em fibras, as dietas 2, 5 e 8 são médias em fibras e as dietas 3, 6 e 9 são baixas em fibras. Em todas as dietas, a gordura e o carboidrato são as duas maiores origens de calorias.

2.2 Modelos de regressão aplicados aos experimentos de mistura

A proposta de experimentos de mistura é construir um modelo apropriado que relacione a resposta aos componentes $x_1, x_2, \ldots, x_q$. Assume-se que a resposta de interesse η seja uma função das variáveis de mistura x_i, ou seja, $\eta = f(x_1, x_2, \ldots, x_q)$. Quando um experimento é feito, é natural assumir que as respostas observadas, denotadas por y_i para o i-ésimo valor $(i = 1, 2, \ldots, n)$ são função da média de η_i com uma variância constante σ^2 para todo $i = 1, 2, \ldots, n$. As respostas observadas contém o erro experimental aditivo ε_i, ou seja, $y_i = \eta_i + \varepsilon_i$, em que $\varepsilon_i \overset{iid}{\sim} N\left(0; \sigma^2\right)$. A forma funcional da resposta $E[Y] = f(x_1, x_2, \ldots, x_q)$ geralmente não é conhecida, mas em várias situações os modelos de aproximação polinomial de primeiro e segundo graus são utilizados.

Na Tabela (2), os modelos (2.1) e (2.2) também são conhecidos como os polinômios canônicos de primeiro e segundo grau, respectivamente, de Scheffé (SCHEFFÉ, 1958). Porém, existem situações em que a resposta apresenta mudanças extremas conforme o valor de um ou mais componentes tendam ao limite da região simplex. Para esses casos, os modelos de Scheffé não são os mais adequados por não contemplarem possíveis efeitos curvilíneos oriundos do comportamento extremo da resposta. Conforme Draper e St. John (1977), esse problema é resolvido com a inclusão de termos inversos no modelo, cuja expressão é obtida, por exemplo, com a inclusão do termo $\beta_{-i} x_i^{-1}$, em que β_{-i} representa o parâme-

Tabela 2 Classificação dos modelos de mistura mais usuais.

Modelo	$E[Y]$	
Linear de Scheffé	$\sum_{i=1}^{q} \hat{\beta}_i x_i$	(2.1)
Quadrático de Scheffé	$\sum_{i=1}^{q} \hat{\beta}_i x_i + \sum_{i=1}^{q} \sum_{i<j}^{q} \hat{\beta}_{ij} x_i x_j$	(2.2)
Linear com variáveis razão	$\hat{\beta}_0 + \sum_{i=1}^{q-1} \hat{\beta}_i w_i$	(2.3)
Quadrático com variáveis razão	$\hat{\beta}_0 + \sum_{i=1}^{q-1} \hat{\beta}_i w_i + \sum_{i=1}^{q-1} \sum_{i<j}^{q-1} \hat{\beta}_{ij} w_i w_j$	(2.4)

FONTE: Box e Draper (2007)

tro associado ao termo inverso i, para cada componente de mistura num modelo linear.

A inclusão de termos inversos no modelo permite ainda produzir estimativas dos parâmetros mais precisas do que as obtidas no modelo de Scheffé, considerando um sistema em que a resposta sofre mudanças extremas. Além desse fato, a colinearidade causada pela restrição em (1.1) é reduzida, no sentido de que a covariância entre as estimativas dos parâmetros do modelo serão menores quando considerando termos inversos (CORNELL; GORMAN, 2003). Caso a restrição em (1.1) fosse removida, o problema da colinearidade estaria resolvido, mas essa possibilidade descaracterizaria o contexto de experimentos de mistura.

Na tentativa de promover maior redução do efeito da colinearidade e melhor modelagem de efeitos curvilíneos, outros modelos foram propostos. Snee (1973) definiu um modelo razão com a inclusão de termos $w_i = \frac{x_i}{x_q^*}$ nos modelos (2.3) e (2.4), em que x_q^* corresponde ao componente de mistura que provoca o *efeito de borda*. Aitchison e Bacon-Shone (1984) propuseram usar variáveis razão na escala logarítmica, ou seja, incluir os termos $w_i = \log\left(\frac{x_i}{x_q^*}\right)$ e Akay e Tez

(2011) propuseram o modelo com variáveis razão utilizando a transformação raiz quadrada, ou seja, $w_i = \sqrt{\frac{x_i}{x_q^*}}$. Pelas equações (2.3) e (2.4), devem existir $q-1$ razões no conjunto e cada razão deve conter pelo menos um dos componentes utilizados em, pelo menos, uma das outras razões pertencente ao conjunto.

2.3 Analizando experimentos de mistura usando Modelos Lineares Generalizados

Suponha que n observações binomiais independentes $y_1, \ldots, y_n$ são tais que a i-ésima observação, $i = 1, \ldots, n$, tem distribuição binomial com parâmetros n_i e π_i. Também suponha que o valor transformado da probabilidade de resposta para a i-ésima observação está relacionado a uma combinação linear de q componentes de mistura, isto é, $g(\pi_i) = \mathbf{x}_i^T\boldsymbol{\beta} = \eta_i$, em que $g(\pi)$ pode ser a transformação logística de π ("logit"), o "probit" de π, a transformação complemento log-log de π, entre outras, e $\mathbf{x}_i^T\boldsymbol{\beta} = \eta_i$ qualquer um dos modelos na Tabela (2). O componente linear deste modelo será denotado por η_i, a função g a função de ligação do modelo linear generalizado e, nesse trabalho, adotou-se g como sendo a transformação logística. Nesse caso, temos que $\text{logit}(\pi_i) = \mathbf{x}_i\boldsymbol{\beta}$, o qual isolando π_i resulta no modelo de regressão logística

$$\pi_i = \frac{\exp\{\mathbf{x}_i\boldsymbol{\beta}\}}{1+\exp\{\mathbf{x}_i\boldsymbol{\beta}\}}. \tag{2.5}$$

O método usual para estimar $\hat{\boldsymbol{\beta}}$ é via Máxima Verossimilhança. O logaritmo da função de verossimilhança para n observações binomiais é dado por

$$l(\boldsymbol{\beta}) = \sum_{i=1}^{n} \left\{ \ln \left(\begin{array}{c} n_i \\ y_i \end{array} \right) + y_i \ln(\pi_i) + (n_i - y_i) \ln(1 - \pi_i) \right\} \tag{2.6}$$

Note que para encontrar o $\hat{\boldsymbol{\beta}}$ que maximiza e equação em (2.6), algum método númerico de otimização deve utilizado, uma vez que $l(\boldsymbol{\beta})$ depende de $\boldsymbol{\beta}$ por meio das probabilidades binomiais π_i. Para tal, o método dos Escores de Fisher pode ser utilizado. A regressão logística pode ser utilizada em situações em que a resposta é proveniente de um evento Bernoulli ou por meio da proporção de eventos $Y = 1$ em n ensaios (HOSMER; LEMESHOW, 2000). Uma distribuição que pode ser utilizada para estudar uma variável resposta contínua e restrita ao intervalo $(0, 1)$ é a distribuição simplex (BANDORFF-NIELSEN; JORGENSES, 1991). A distribuição simplex faz parte dos modelos de dispersão (JORGENSEN, 1997), que estendem os modelos lineares generalizados.

Uma variável aleatória y que segue uma distribuição simplex com média $\mu \in (0, 1)$ e parâmetro de dispersão $\sigma^2 > 0$ tem função densidade dada por

$$f\left(y; \mu, \sigma^2\right) = \frac{1}{\sqrt{2\pi\sigma^2 \{y(1-y)\}^3}} \exp\left\{ -\frac{1}{2\sigma^2} d(y; \mu) \right\}, \; y \in (0, 1), \tag{2.7}$$

em que,

$$d(y; \mu) = \frac{(y - \mu)^2}{y(1-y)^2 \mu^2 (1-\mu)^2}. \tag{2.8}$$

A distribuição de y será denotada por $S\left(\mu, \sigma^2\right)$. Considere $y_1, \ldots, y_n$ variáveis aleatórias independentes, sendo que cada $y_i \sim S\left(\mu_i, \sigma^2\right)$, $i = 1, \ldots, n$. O modelo de regressão simplex é definido pela densidade da forma (2.7), sendo as

médias μ_i modeladas por $g(\mu_i) = \mathbf{x}_i^T\boldsymbol{\beta} = \eta_i$, em que $g(\cdot)$ é a função de ligação, estritamente monótona e duplamente diferenciável que transforma valores do intervalo $(0, 1)$ nos reais, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_q)$ é o vetor dos parâmetros da regressão ($\boldsymbol{\beta} \in \mathbb{R}^q$), $\mathbf{x}_i^T = (x_{1i}, \ldots, x_{qi})$ são os valores conhecidos de q componentes de mistura e η_i é o preditor linear.

Uma característica do modelo de regressão simplex é que ele acomoda variáveis respostas com variâncias não constantes, assim como o modelo de regressão beta, pois a variância de y_i é uma função de μ_i. Além disso, o parâmetro de dispersão determina a forma da distribuição simplex, proporcionando maior flexibilidade na modelagem.

O procedimento para estimação dos parâmetros do modelo de regressão simplex é derivado de maneira similar ao feito para o modelo de regressão logística, com a diferença de que existe um parâmetro adicional no modelo, o σ^2. O logaritmo da função de verossimilhança para de n observações independentes é dado por

$$l\left(\boldsymbol{\beta}, \sigma^2\right) = \sum_{i=1}^{n} -\frac{1}{2}\ln(2\pi) - \frac{1}{2}\ln\left(\sigma^2\right) - \frac{3}{2}\ln\left[y_i\left(1 - y_i\right)\right] - \frac{1}{2\sigma^2}d\left(y_i; \mu_i\right). \tag{2.9}$$

Os estimadores de máxima verossimilhança dos parâmetros $\boldsymbol{\beta}$ e σ^2 são obtidos por meio da solução do sistema de equações homogêneo. Porém, somente o estimador para σ^2 possui forma fechada. Para a estimação de $\boldsymbol{\beta}$, será necessário utilizar algum método de maximização numérica do logaritmo da função de verossimilhança definida em (2.9), como por exemplo os métodos de Newton-Raphson ou dos escores de Fisher e suas variações.

2.3.1 Construção dos gráficos de envelope simulado em MLG

A adequação de ajuste de um MLG pode ser verificada pelo gráfico normal de probabilidades para o resíduo componente de desvio padronizado e indica se existem evidências de afastamento da suposição de distribuição candidata para a resposta. Esse gráfico consiste em gerar bandas de confiança por reamostragem, também chamado de envelope, e um ajuste adequado ocorre se todos os resíduos (ou grande parte deles) do modelo estiverem contidos nessas bandas de confiança. Mais detalhes sobre o envelope simulado podem ser vistos em Atkinson (1985). Em estudos de simulação, Williams (1987) encontrou fortes evidências de concordância entre a distribuição empírica do componente de desvio padronizado e a distribuição normal padrão para vários MLG's.

A construção do gráfico de envelope simulado para o modelo de regressão logística foi utilizando-se do seguinte procedimento:

(1) Gera-se n observações da distribuição uniforme $U\,(0;1)$, obtendo o vetor $\mathbf{U} = (u_1, \ldots, u_n)$ e em seguida o vetor dos desvios $\mathbf{v} = \mathbf{U} - \hat{\mathbf{y}}$, em que $\hat{\mathbf{y}} = (\hat{y}_1, \ldots, \hat{y}_n)^T$;

(2) Ajusta-se $\mathbf{v}$ à $\mathbf{X}$, a matriz de delineamentodo experimento, obtendo r_i, $i = 1, \ldots, n$, o resíduo componente do desvio;

(3) Obtém-se o resíduo componente do desvio padronizado $t_i^* = t_{D_i}; i = 1, \ldots, n$, conforme descrito em Hosmer e Lemeshow (2000);

(4) Repete-se os passos (1) até (3) m vezes, resultando os resíduos gerados t_{ij}^*, $i = 1, \ldots, n$ e $j = 1, \ldots, m$;

(5) Coloca-se cada grupo de n resíduos em ordem crescente, obtendo $t_{(i)j}^*$, $i =$

$1, \ldots, n$ e $j = 1, \ldots, m$;

(6) Obtém-se os limites $t^*_{(i)I} = \min_j t^*_{(i)j}$ e $t^*_{(i)S} = \max_j t^*_{(i)j}$. Assim, os limites correspondentes ao i-ésimo resíduo serão dados por $t^*_{(i)I}$ e $t^*_{(i)S}$.

O gráfico de envelope simulado para o modelo de regressão simplex foi obtido de maneira similar ao procedimento acima considerando algumas modificações. No passo 1 as observações geradas foram provenientes de uma distribuição simplex com parâmetro de posição $\hat{\mu}_i$ e parâmetro de dispersão $\hat{\sigma}^2$ e no passo 3 o resíduo $t^*_i = r^{pp}_i$, conforme Espinheira, Ferrari e Cribari-Neto (2008), foi utilizado.

2.4 Seleção de modelos e critérios de adequação de ajuste

O Critério de informação de Akaike (AIC) proposto em Akaike (1974), é uma medida relativa da qualidade de ajuste de um modelo estatístico. É definido como $AIC = -2\,l\left(\hat{\theta}\,|y\right) + 2p$, em que $l\left(\hat{\theta}\,|y\right)$ é o logaritmo neperiano da função de verossimilhança do modelo em $\hat{\theta}$ e p é o número de parâmetros do modelo.

Schwarz (1978) propôs um critério conhecido como Critério de Informação Bayesiano (BIC), que corresponde à troca do fator 2, que é o peso do número de parâmetros, por $\ln(n)$, logo o BIC é dado por $BIC = -2\,l\left(\hat{\theta}\,|y\right) + p\,\ln(n)$, em que n é o número de observações da amostra.

Os critérios de informação apresentados são uma combinação entre uma medida de qualidade de ajuste (como o negativo de duas vezes a log-verossimilhança) com uma medida de complexidade do modelo, como por exemplo, no AIC e BIC essa medida é dada em termos do número de parâmetros do modelo. Bozdogan (2000) propôs um critério, o índice de complexidade da informação de Bozdogan (ICOMP), que utiliza a matriz de informação de Fisher para compor o termo que

mensura a complexidade do modelo, uma vez que ela contém informação sobre a correlação entre os parâmetros do modelo avaliado.

Por isso, o mesmo autor diz que critério ICOMP estende os critérios do tipo AIC, com uma penalidade devido ao aumento da complexidade do sistema, no sentido de que quanto maior o número de variáveis e a relação entre elas, mais complexo se torna o modelo. O ICOMP é definido como

$$ICOMP = -2l\left(\hat{\theta}\,|y\right) + 2C\left[I^{-1}\left(\hat{\theta}\right)\right], \tag{2.10}$$

em que,

$$C\left[I^{-1}\left(\hat{\theta}\right)\right] = \frac{s}{2}\ln\left[tr\left(\frac{I^{-1}\left(\hat{\theta}\right)}{s}\right)\right] - \frac{1}{2}\ln\left[tr\left(I^{-1}\left(\hat{\theta}\right)\right)\right], \tag{2.11}$$

denota a complexidade da informação maximal de $I^{-1}\left(\hat{\theta}\right)$. Na equação 2.10, p é número de parâmetros, n o tamanho da amostra, $\hat{\theta}$ a estimativa dos parâmetros, $I^{-1}\left(\hat{\theta}\right) = V\hat{a}r\left[\hat{\theta}\right]$ a inversa da matriz de informação de Fisher, e $s = rank\left[I^{-1}\left(\hat{\theta}\right)\right]$. Os elementos da diagonal de $I^{-1}\left(\hat{\theta}\right)$ correspondem às variâncias estimadas dos parâmetros do modelo e indicam a sensibilidade dos parâmetros, enquanto que os elementos fora da diagonal são as covariâncias entre os parâmetros, as quais medem o grau de colinearidade entre as colunas de $I^{-1}\left(\hat{\theta}\right)$ e revelam a dimensão da independência dos parâmetros. De acordo com esse critério, o melhor modelo dentro de um conjunto de modelos é o que minimiza o ICOMP (LANNING; BOZDOGAN, 2003).

2.5 Interpretação dos parâmetros e medindo o efeito dos componentes em experimentos de mistura

Em MLG´s, funções de ligação podem ser usados para interpretação dos parâmetros. Em se tratando da transformação logística, β_i representa a mudança no logit que resultaria de uma mudança unitária em x_i, quando as outras variáveis estão fixas. Desse modo, as interpretações para os parâmetros usando a ligação *logit* são diretas para os coeficientes exponenciados. Esses coeficientes exponenciados representam a razão de chances (HOSMER; LEMESHOW, 2000).

Na análise de experimentos de mistura com transformação logística, a razão de chances não pode ser usada como na regressão logística usual, por exemplo, pois a proporção dos componentes individuais x_i estão no intervalo entre 0 e 1 e os outros componentes não podem se manter constantes devido às restrições dadas na expressão (1.1). Em outras palavras, se a quantidade do componente x_i aumenta, então a quantidade de todos os outros componentes diminuem, mas sua razão de um para o outro permanece constante. Para melhor entendimento desse conceito, deve-se utilizar a direção Cox.

2.5.1 Direção Cox para o gráfico de resposta traço

A direção Cox do componente i é uma linha imaginária projetada da mistura de referência para o vértice $x_i = 1$. As proporções dos q componentes na mistura de referência é $\mathbf{c} = (c_1, c_2, \ldots, c_q)$, em que $\sum_{i=1}^{q} c_i = 1$. O ponto de referência $\mathbf{c}$ padrão é geralmente adotado como o centróide do experimento. Quando a proporção c_i do componente i é alterada por uma quantidade Δ_i na direção Cox, então a nova proporção se torna

$$x_i = c_i + \Delta_i. \tag{2.12}$$

Quando existem apenas restrições como na expressão (1.1) sobre os componentes de mistura, Δ_i está no intervalo $[-c_i,\, 1 - c_i]$. Caso contrário, para restrições adicionais sobre os componentes como na expressão (1.2), Δ_i está no intervalo $[L_i - c_i,\, U_i - c_i]$. As proporções dos $q - 1$ componentes restantes, resultante de c_i no i-ésimo componente, é

$$x_j = c_j \frac{1 - x_i}{1 - c_i},\ j = 1,\, 2,\, \ldots,\, q,\, j \neq i. \tag{2.13}$$

No caso de uma região experimental restrita ser um simplex regular, uma representação alternativa da direção Cox pode ser formulada, considerando o fato de que $\frac{x_j}{x_k} = \frac{c_j}{c_k}$. Nesse caso, analogamente à razão das proporções dos componentes escolhidos ao longo dos eixos para um componente na direção Cox, com a ajuda da razão de componentes em qualquer ponto na região restrita, a mudança na resposta pode ser avaliada. Por exemplo, em um sistema de mistura com $q = 3$, ao longo que o eixo do componente x_i passa através do ponto de referência, os componentes x_j e x_k sendo $x_j/x_k = \rho_{x_i}$, são dados por

$$x_j = \frac{\rho_{x_i}\,(1 - x_i)}{\rho_{x_i} + 1} \quad \text{e} \quad x_k = \frac{1 - x_i}{\rho_{x_i} + 1},$$

em que, $1 - c_i = c_j + c_k$, $L_i \leq x_i \leq U_i$ e ρ_{x_i} mostra a razão dos componentes exceto x_i no ponto referência. Dessa forma, o valor da resposta predita para o preditor linear de primeiro grau, ao longo da direção Cox para o i-ésimo componente é dado por:

$$logit\left(\hat{\pi}_{x_i}\right) = \hat{\beta}_i x_i + \hat{\beta}_j \frac{\rho_{x_i}\left(1 - x_i\right)}{\rho_{x_i} + 1} + \hat{\beta}_k \frac{1 - x_i}{\rho_{x_i} + 1},\ L_i \leq x_i \leq U_i. \tag{2.14}$$

Na equação (2.14), exponenciando $logit\left(\hat{\pi}_{x_i}\right)$, os valores preditos de $\hat{\pi}_{x_i}$ para cada componente de mistura x_i são obtidos. Com esses valores, pode-se construir um gráfico que relaciona os incrementos de x_i com os valores de $\hat{\pi}_{x_i}$ e esse gráfico é chamado de *resposta traço* (*trace plot*). Esses traços representam o efeito da mudança de cada componente de mistura enquanto todos os outros componentes permanecem em razão constante. O modelo dado pela equação (2.14) pode ser expandido para diferentes preditores lineares (AKAY; TEZ, 2007).

2.5.2 Gráficos da razão de chances para os componente de mistura

A razão de chances é frequentemente utilizada para interpretação dos parâmetros em MLG's utilizando transformação logística, devida a sua fácil interpretação. É utilizada também como uma medida relativa da chance de sucesso de um conjunto em relação a um outro conjunto (CUMMINGS, 2009). Em experimentos de mistura, técnicas gráficas baseadas em resposta traço podem ser usadas para comparações do tipo. Considere um ponto qualquer $\mathbf{c} = (c_i,\ c_j,\ c_k)$ que é tomado como um grupo controle sobre a região experimental, a razão de chances é dada ao longo do eixo x_i por

$$\widehat{OR\left(x_i\right)} = \frac{chance\, x_i}{chance\, controle} = \frac{\exp\left\{\hat{\beta}_i x_i + \hat{\beta}_j \frac{\rho_{x_i}(1-x_i)}{\rho_{x_i}+1} + \hat{\beta}_k \frac{1-x_i}{\rho_{x_i}+1}\right\}}{\exp\left\{\hat{\beta}_i c_i + \hat{\beta}_j c_j + \hat{\beta}_k c_k\right\}};\ L_i \leq x_i \leq U_i. \tag{2.15}$$

De maneira mais simples, $\widehat{OR\left(x_i\right)} = \exp\left\{\hat{\beta}_i A + \hat{\beta}_j B + \hat{\beta}_k C\right\}$, em que $A = x_i - c_i$, $B = \frac{\rho_{x_i}(1-x_i)}{\rho_{x_i}+1} - c_j$ e $C = \frac{1-x_i}{\rho_{x_i}+1} - c_k$, é uma estimativa pontual e assume

valores entre zero e infinito. Se a razão de chances é igual a um, isso quer dizer que não existe diferença entre os grupos comparados, e se for diferente de um, significa que existe diferença. Além disso, as razões preditas obtidas pela equação (2.15) formarão uma curva que pode ser interpretada na comparação entre grupos.

A precisão da razão de chances pode ser verificada pelo intervalo de confiança e sua amplitude reflete sua variabilidade inerente. Para calcular o erro padrão e o intervalo de confiança para a razão de chances, é necessário transformar a equação (2.15) para a escala logarítmica e, nesse caso, segundo Hosmer e Lemeshow (2000) a distribuição amostral dessa medida é aproximadamente normal. Observe que a expressão para o logaritmo neperiano da razão das chances na equação (2.15) depende do componente de mistura x_i e o coeficiente estimado $\hat{\beta}_i$. Utilizando métodos para calcular a variância de uma soma, obtemos o seguinte estimador $V\hat{a}r\left\{\ln\left[\widehat{OR\left(x_i\right)}\right]\right\} = A^2V\hat{a}r\left[\hat{\beta}_i\right] + B^2V\hat{a}r\left[\hat{\beta}_j\right] + C^2V\hat{a}r\left[\hat{\beta}_k\right] + 2ABC\hat{o}v\left[\hat{\beta}_i, \hat{\beta}_j\right] + 2ACC\hat{o}v\left[\hat{\beta}_i, \hat{\beta}_k\right] + 2BCC\hat{o}v\left[\hat{\beta}_j, \hat{\beta}_k\right]$, em que as variâncias e covariâncias dos parâmetros do modelo ajustado são provenientes da inversa da matriz de informação de Fisher. Uma vez obtida a estimativa da variância do logaritmo neperiano da razão das chances estimado, pode-se obter o estimador do intervalo de $100\left(1-\alpha\right)\%$ de confiança para o logaritmo neperiano da razão das chances, que é dado por,

$$IC_{(1-\alpha)}\left\{\ln\left[\widehat{OR\left(x_i\right)}\right]\right\} = \left[\hat{\beta}_iA + \hat{\beta}_jB + \hat{\beta}_kC\right] \pm z_{(\frac{\alpha}{2})}\sqrt{V\hat{a}r\left\{\ln\left[\widehat{OR\left(x_i\right)}\right]\right\}}, \tag{2.16}$$

em que, $\sqrt{V\hat{a}r\left\{\ln\left[\widehat{OR\left(x_i\right)}\right]\right\}}$ é o erro padrão do estimador do logaritmo neperiano da razão de chances, $z_{(\frac{\alpha}{2})}$ é o $(\frac{\alpha}{2})$-ésimo quantil da distribuição normal padrão com α nível de significância.

Os limites inferiores e superiores do estimador do intervalo de confiança

para a razão de chances são obtidos exponenciando os limites na equação (2.16) e são preferidos os modelos cujos intervalos de confiança são mais estreitos. Portanto, métodos gráficos baseados em intervalos de confiança para a razão de chances podem ser utilizados para comparar os modelos.

Finalizando a metodologia proposta, para obtenção dos resultados foram utilizados os pacotes estatísticos, *bbmle* e *mixexp* do Sistema Computacional Estatístico R (R CORE TEAM, 2015).

3 RESULTADOS

3.1 Seleção do Modelo

Uma observação a ser feita é a de que os resultados da regressão logística descritos na Tabela 3 são idênticos aos apresentados por Akay e Tez (2011), que apresentaram o modelo com variáveis razão como uma melhor alternativa aos modelos de regressão logística propostos por Chen, Li e Jackson (1996). Esses autores usaram o modelo de regressão logística para modelar a proporção de ratos com tumor em uma determinada mistura de componentes (Tabela 1) e fizeram estudos considerando as variáveis em pseudo-componentes nos modelos polinomiais de Scheffé e Backer. Dessa forma, em comparação aos resultados obtidos por esses autores, os resultados da Tabela 3 apontam o modelo de regressão simplex como sendo melhor do que os propostos por Akay e Tez (2011). Em se tratando dos modelos com variáveis razão, notou-se que os menores valores dos critérios ICOMP, AIC e BIC foram obtidos. Portanto, o modelo de regressão simplex com variáveis razão foi o melhor dentre os modelos utilizados e esse modelo forneceu menores erros-padrão das estimativas dos parâmetros do modelo, indicando maior

precisão nas mesmas.

Tabela 3 Estimativas dos parâmetros para os modelos ajustados e resultados dos indicadores de qualidade de ajuste.

Modelo	Tipo	Parâmetro *	Estimativa	Erro Padrao	ICOMP	AIC	BIC
Regressão Logística							
M1	Scheffé	β_1	-1,2397	1,4044	44,14	44,14	44,93
		β_2	-0,9678	0,6436			
		β_3	-6,4352	8,3783			
		β_4	10,4474	4,3393			
M2	Razão	β_0	0,2895	0,2655	29,93	44,41	45,00
		β_1	0,1808	0,0542			
		β_2	-0,0767	0,0402			
Regressão Simplex							
M3	Scheffé	β_1	-0,8241	0,9422	-20,25	-18,49	-22,09
		β_2	-0,8896	0,501			
		β_3	-7,9549	6,4595			
		β_4	9,5299	2,8266			
M4	Razão	β_0	0,2959	0,2006	-34,41	-18,8	-21,51
		β_1	0,1779	0,0361			
		β_2	-0,0754	0,033			

* Parâmetros relacionados aos modelos quadrático de Scheffé, com preditor linear dado por $\eta = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_1 x_2$, e linear de Razão, com preditor linear dado por $\eta = \beta_0 + \beta_1 \sqrt{\frac{x_1}{x_3}} + \beta_2 \sqrt{\frac{x_2}{x_3}}$.

Smith (2005) fez críticas com relação à análise feita por Chen, Li e Jackson (1996) em termos de colinearidade. Por exemplo, a estimativa de parâmetro para o componente x_3 é maior que a dos outros parâmetros. Menard (2009) alerta que a razão para essa discrepância entre as estimativas dos parâmetros do modelo é devida à colinearidade. Esse fato pode ser observado na matriz de variâncias e covariâncias S_{M1} e S_{M3} dos modelos M1 e M3. Além disso, os elementos da diagonal de S_{M3} indicam que a raiz quadrada desses valores resultarão em menores erros-padrão das estimativas.

As covariâncias entre parâmetros nos modelos M2 e M4 são menores do que as dos modelos M1 e M3, conforme pode ser visto nas matrizes S_{M2} e S_{M4}, logo, supostamente a diferença entre os valores do ICOMP pode ser explicada,

pois, conforme menciona Bozdogan (2000), sistemas cujas covariâncias entre seus componentes são mais evidentes, tendem a ter maiores valores para o ICOMP e, por outro lado, menores covariâncias resultam em menores valores para o ICOMP. Além disso, o modelo M4 tem maior informação sobre os parâmetros, uma vez que as variâncias dos mesmos são menores do que as dos outros modelos.

$$S_{M1} = \begin{bmatrix} 1,9723 & 0,6798 & -3,7750 & -5,7802 \\ 0,6798 & 0,4143 & -2,5498 & -2,3440 \\ -3,7750 & -2,5498 & 70,1966 & 7,3203 \\ -5,7802 & -2,3440 & 7,3203 & 18,8293 \end{bmatrix}$$

$$S_{M3} = \begin{bmatrix} 0,8877 & 0,3440 & -2,7979 & -2,5038 \\ 0,3440 & 0,2510 & -1,7064 & -1,2038 \\ -2,7979 & -1,7064 & 41,7257 & 6,3243 \\ -2,5038 & -1,2038 & 6,3243 & 7,9899 \end{bmatrix}$$

$$S_{M2} = \begin{bmatrix} 0,0705 & -0,0053 & -0,0035 \\ -0,0053 & 0,0029 & -0,0015 \\ -0,0035 & -0,0015 & 0,0016 \end{bmatrix}$$

$$S_{M4} = \begin{bmatrix} 0,0403 & -0,0020 & -0,0022 \\ -0,0020 & 0,0013 & -0,0009 \\ -0,0022 & -0,0009 & 0,0011 \end{bmatrix}$$

Akay e Tez (2011) fizeram uma observação quanto à presença do efeito da sub ou super-dispersão em dados agrupados, como é o caso dos dados apresentados na Tabela 1. Nesse contexto, assume-se que das n_i observações em um determinado grupo (ou dieta), todas elas tenham a mesma probabilidade π_i de ter o atributo de interesse (a ocorrência de tumor mamário), então é razoável assu-

mir que a distribuição de Y_i seja binomial com parâmetros π_i e n_i. O caso de dados não agrupados pode ser considerado como sendo um caso especial, onde $n_1 = n_2 = \cdots = n_n = 1$.

Os autores mencionaram que esse fato deve ser levado em consideração na seleção do modelo. No trabalho de Chen, Li e Jackson (1996) esse fato foi desprezado e Akay e Tez (2011) estimaram os parâmetros de dispersão dos modelos M1 e M2 e encontraram os valores $\hat{\phi}_{M1} = 0,6786$ e $\hat{\phi}_{M2} = 0,9434$, respectivamente. Nesse caso, pode-se dizer que no modelo M1 o efeito da sub-dispersão está presente e, por isso, está mal especificado. Ao utilizar o modelo com variáveis razão (M2) a estimativa do parâmetro de dispersão está próximo do valor unitário, que é o valor assumido para o modelo de regressão logística usual. Com isso, diz-se que o modelo M2 controlou o efeito de sub-dispersão. No caso dos modelos M3 e M4, o modelo de regressão simplex naturalmente modela a dispersão e para os referidos modelos são dados por $\hat{\phi}_{M3} = 0,7262$ e $\hat{\phi}_{M4} = 0,7291$.

Examinando os pontos experimentais (Figura 1) e referindo-se ao delineamento apresentado na Tabela 1, observa-se que a resposta sofreu uma mudança nos valores próximos do limite inferior do componente x_3 (fibra). Nos pontos em que o componente x_1 (gordura) tem um valor alto, por exemplo nas dietas 4, 6, 8 e 9, observa-se um aumento no número observado de tumores. Contudo, nos pontos em que o componente x_1 assume valores pequenos, observa-se um decréscimo no número de tumores. Quando o componente x_2 tem grandes valores, por exemplo nas dietas 1, 2 e 3, o número de tumores é reduzido e quando ele assume pequenos valores, existe um aumento na resposta (Figura 1). Para modelar um comportamento dos componentes desse tipo, utilizar a razão entre os componentes é uma abordagem mais realística, uma vez que modelos com esses tipos de variáveis permitem modelar mudanças quando um dos componentes assume valores pequenos,

como é o caso de x_3 (fibra). Akay e Tez (2011) concluiram que o modelo com variáveis razão utilizando a transformação raiz quadrada (M2) foi o melhor dentre os utilizados.

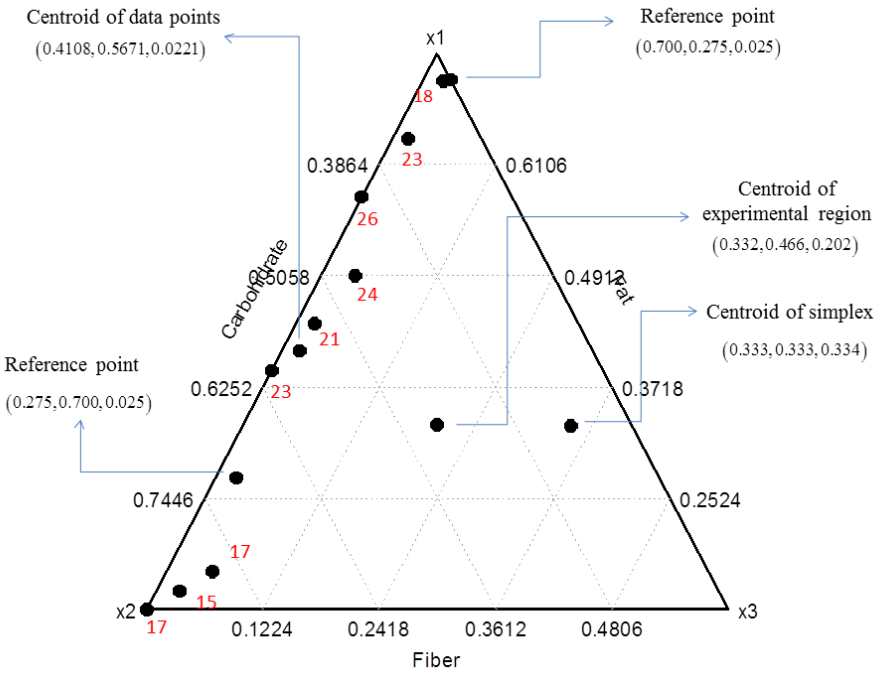
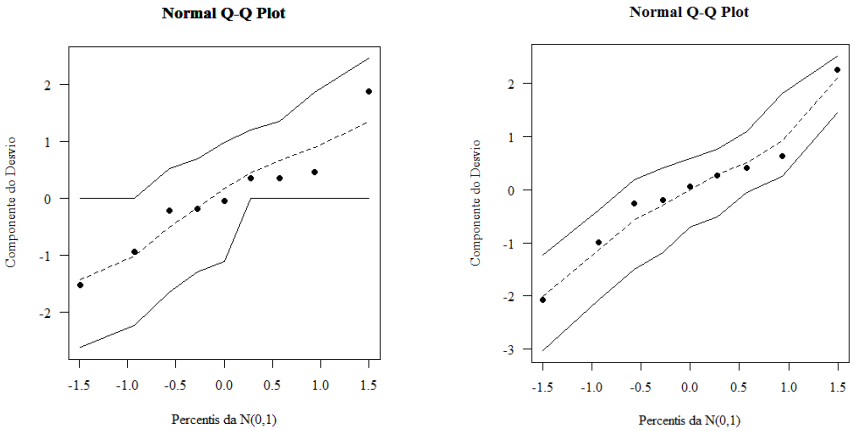


Figura 1 Região simplex restrita dos grupos de ratos tratados com diferentes proporções calóricas de fibra, gordura e carboidrato (componentes originais) e pontos de referência utilizados.

Diante do exposto, pode-se afirmar que o modelo de regressão simplex apresentou melhores indicadores de qualidade de ajuste para a proporção de tumor mamário em ratos fêmeas (Tabela 3). Para fins de comparação, foram discutidos os modelos M2 e M4, uma vez que o M2 foi o melhor dentre o proposto por Akay e Tez (2011) e o M4 o melhor dentre que consideraram a distribuição simplex. Dessa forma, os gráficos normais de probabilidade dos resíduos componente do



(a) M2 (b) M4

Figura 2 Envelopes simulados para diagnóstico de qualidade de ajuste dos modelos M2 (Regressão Logística do tipo razão - (b)) e M4 (Regressão Simplex do tipo razão - (b)).

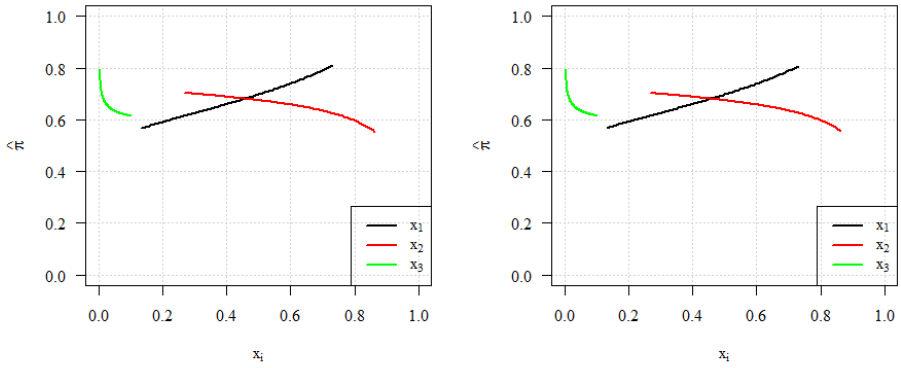
desvio para os modelos M2 e M4 afirmam que a suposição de resposta binomial, Figura 2(a), e/ou simplex, Figura 2(b), para a resposta analisada está adequada e o ajuste dos modelos foram satisfatórios (Figura 2).

3.2 Discussão dos modelos em relação aos gráficos dos efeitos dos componentes de mistura

A seguir serão apresentadas as abordagens gráficas dos modelos M2 e M4. Para os gráficos de resposta traço o ponto de referência na Direção Cox foi dado pelo centróide dos dados $\mathbf{c} = (0.4108,\ 0.5671,\ 0.0221)$, como pode ser visto na Figura 1.

Na Figura 3 (a), modelo M2, os componentes x_1 (gordura) e x_2 (carboidrato) tem efeito contrário sobre a resposta. Á medida que a proporção de gordura

aumenta a proporção esperada de tumor aumenta. Por outro lado, á medida que a proporção de carboidrato aumenta a proporção esperada de tumor diminui. O componente x_3 (fibra) tem mais efeito sobre a resposta do que os outros componentes, uma vez que sucessivos incrementos de fibra na dieta acarretam maior diminuição no número esperado de tumores. De maneira análoga, as mesmas conclusões podem ser feitas para o modelo M4.



(a) M2 (b) M4

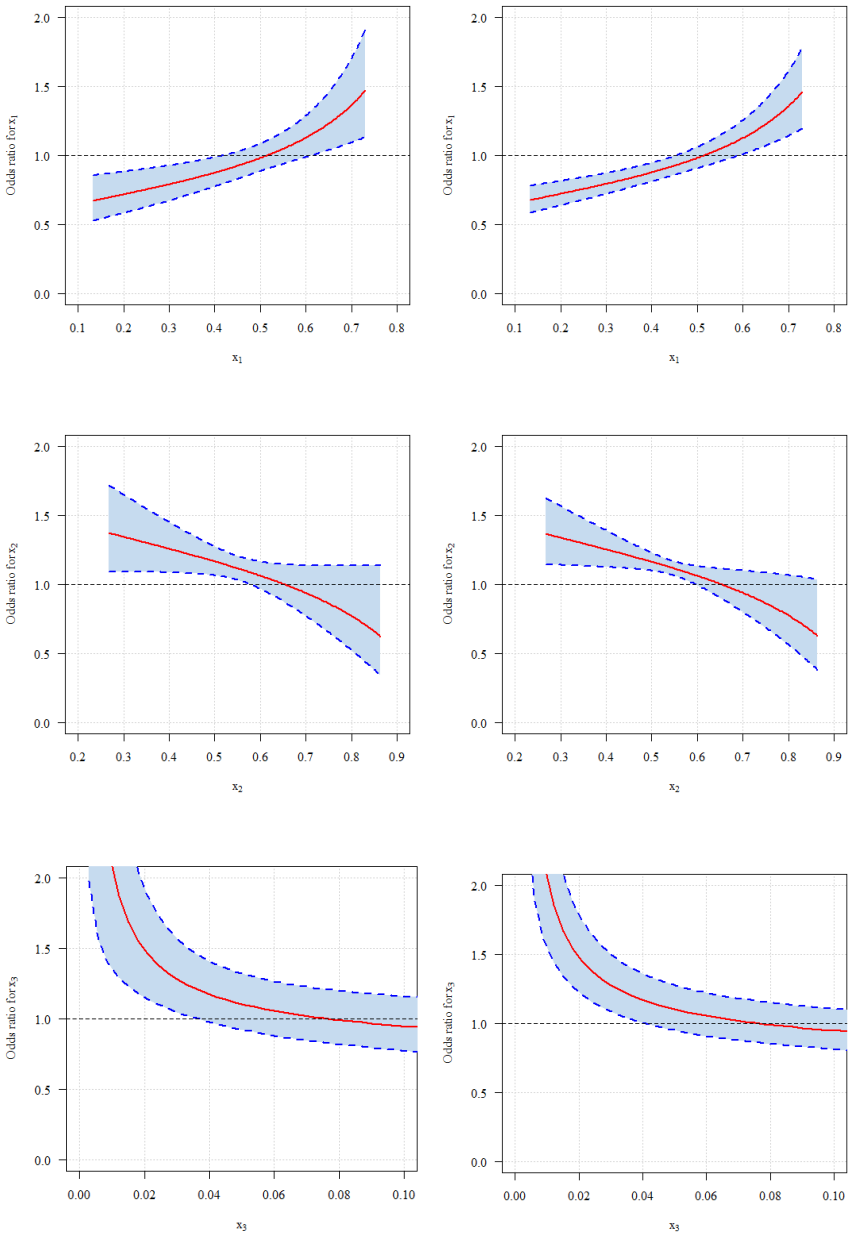
Figura 3 Gráficos de resposta traço dos modelos M2 (Regressão Logística do tipo razão - (a)) e M4 (Regressão Simplex do tipo razão - (d)) considerando o ponto $c = (0.4108, 0.5671, 0.0221)$, o centróide dos dados, como ponto de referência.

As Figuras 4 a 6 apresentam as razões de chances para diferentes pontos de referência em relação ao grupo controle para cada modelo avaliado. Para tal, foram considerados três diferentes pontos de referência $(0.7,\ 0.275,\ 0.025)$, $(0.275,\ 0.7,\ 0.025)$ e $(0.332,\ 0.466,\ 0.202)$. Os dois primeiros pontos de referência estão contidos na região onde os pontos amostrais estão. O terceiro ponto de referência é o centróide da região experimental restrita (Figura 1). Uma observação a ser feita é que deve-se ter cautela quanto às interpretações sobre o centróide da re-

gião restrita, uma vez que não foram amostrados dados nessa região. O grupo controle foi dado pelo centróide dos pontos amostrais $\mathbf{c} = (0.4108,\ 0.5671,\ 0.0221)$. O presente trabalho não se ateve às motivações biológicas para a escolha dos referidos pontos, mas sim ao fato de tal escolha ter sido feita estritamente por inspeção da região experimental e aplicabilidade em experimentos de mistura.

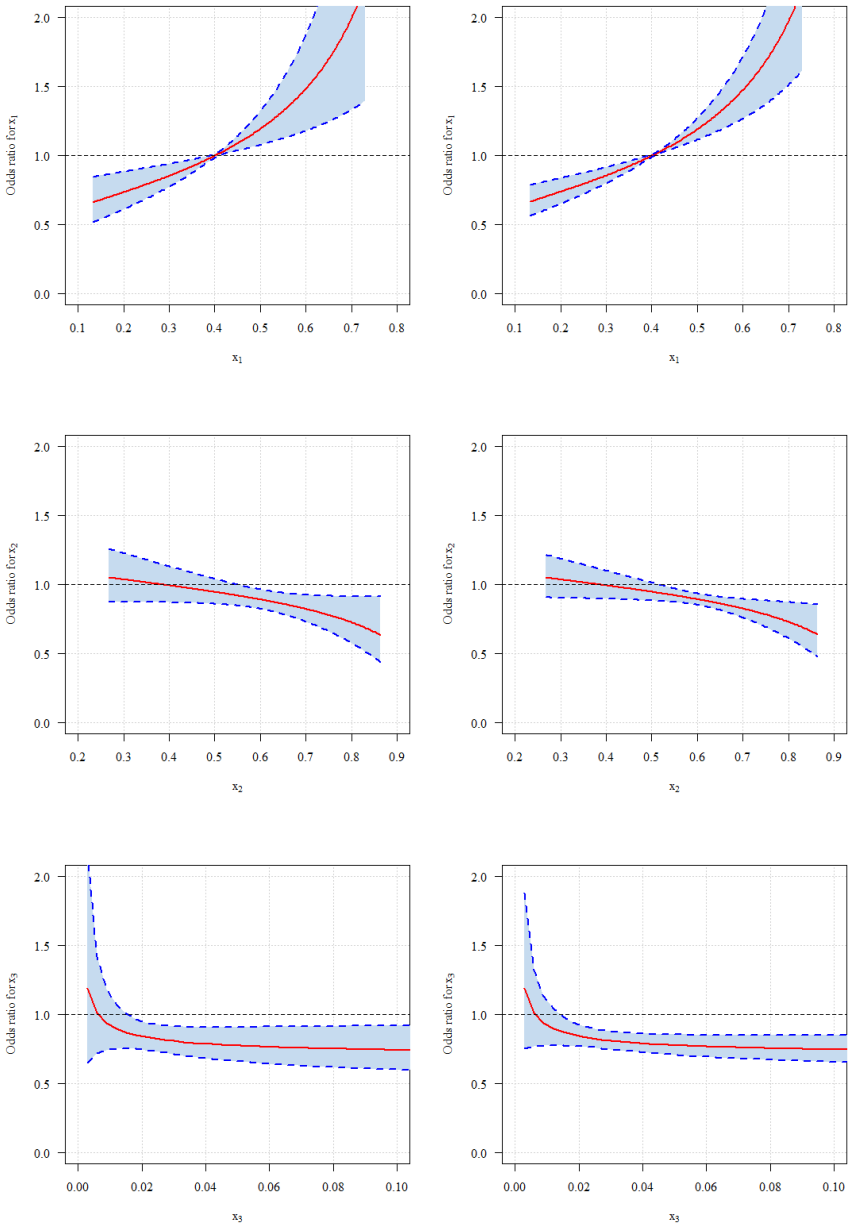
Pode ser observado que para o componente x_1, a chance de ocorrência de tumor mamário em ratos aumenta conforme ocorrem incrementos nesse componente. O respectivo intervalo de confiança de 95% do modelo M2 contém o valor 1, utilizado para comparar as razões de chances, nas quantidades de 0,4 a 0,6 aproximadamente. Logo pode-se dizer que embora a chance aumenta para as quantidades de 0,4 a 0,6, aproximadamente, de x_1, ela não é significativa, no sentido de que na população o componente x_1 (gordura) não influencia de maneira significativa na ocorrência de tumor mamário em ratos (Figura 4 (a)).

Em se tratando de estimativa pontual, as mesmas conclusões podem ser obtidas para o modelo M2 considerando o componente x_1, no entanto, observa-se que o seu respectivo intervalo de confiança de 95% para a razão de chances não contém o valor unitário em algumas quantidades de x_1 (Figura 4(b)), o que indica que esse componente influencia de maneira significativa na ocorrência de tumor mamário em ratos em quantidades diferentes daquelas explicadas pelo modelo M2. Esse fato é evidenciado pela amplitude do intervalo de confiança para a razão de chances desse componente, o qual é mais estreito no modelo M4 do que no modelo M2. Logo, o modelo M4 fornece estimativas de razão de chances do componente x_1 mais precisas do que o modelo M2 (Figura 4(b)).



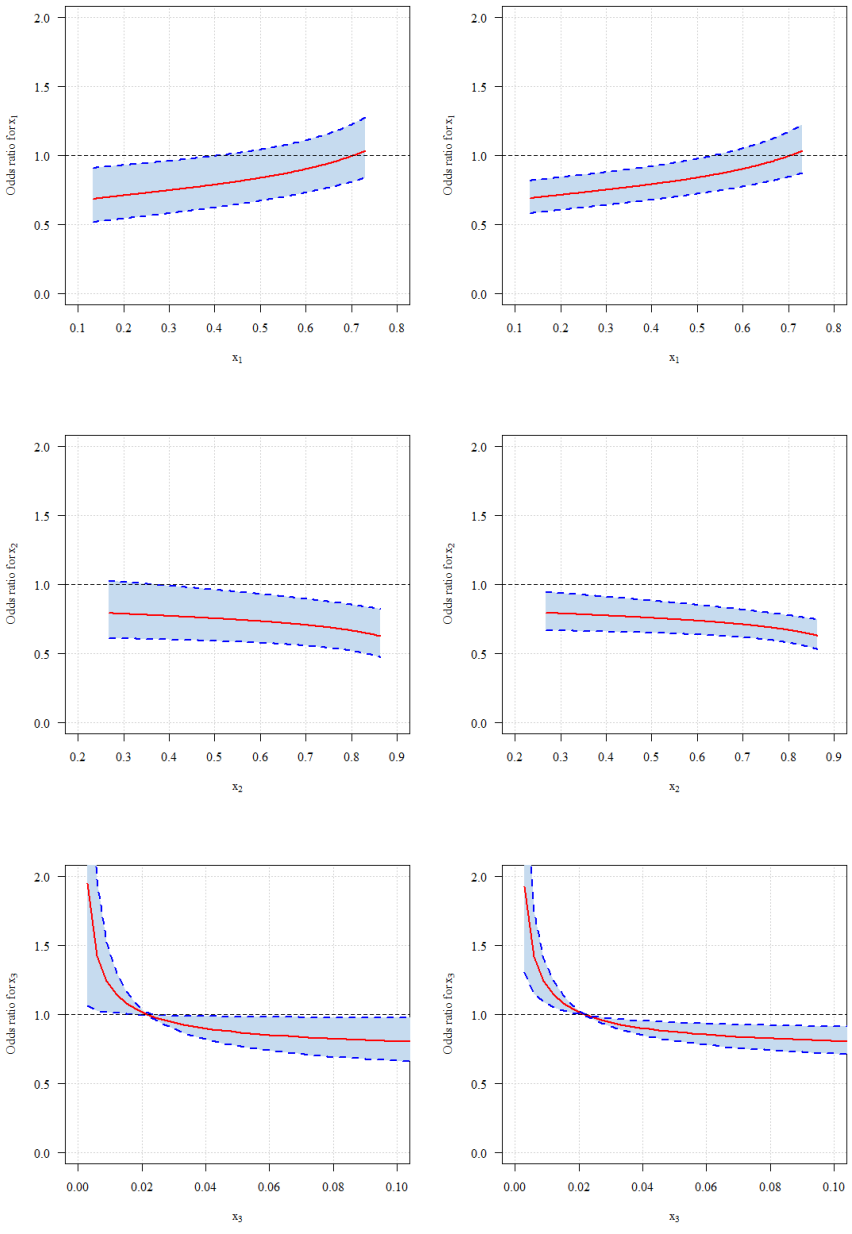
(a)M2 (b)M4

Figura 4 Razão de Chances e seus respectivos Intervalos de Confiança de 95% de confiança para os modelos M2 (Regressão Logística do tipo razão - (a)) e M4 (Regressão Simplex do tipo razão - (b)) no ponto de referência $(0.7, 0.275, 0.025)$ e grupo controle $(0.4108, 0.5671, 0.0221)$. Para visualização desses pontos veja Figura 1.



(a)M2 (b)M4

Figura 5 Razão de Chances e seus respectivos Intervalos de Confiança de 95% de confiança para os modelos M2 (Regressão Logística do tipo razão - (a)) e M4 (Regressão Simplex do tipo razão - (b)) no ponto de referência $(0.275, 0.700, 0.025)$ e grupo controle $(0.4108, 0.5671, 0.0221)$. Para visualização desses pontos veja Figura 1.



(a)M2 (b)M4

Figura 6 Razão de Chances e seus respectivos Intervalos de Confiança de 95% de confiança para os modelos M2 (Regressão Logística do tipo razão - (a)) e M4 (Regressão Simplex do tipo razão - (b)) no ponto de referência $(0.332, 0.466, 0.202)$ e grupo controle $(0.4108, 0.5671, 0.0221)$. Para visualização desses pontos veja Figura 1.

Dessa forma, pode-se afirmar que o modelo M4 proporcionou estimativas mais precisas de razão de chances do que o modelo M2 em todos os pontos de referência adotados (Figuras 4 a 6). Esse fato pode ser explicado pela inspeção das covariâncias entre os parâmetros dos modelos avaliados, como foi discutido na seção 3.1.

Portanto, conclui-se que considerar a proporção de incidência de tumor mamário em ratos como sendo uma variável aleatória com distribuição simplex e o uso de variáveis razão para estudar a relação entre os componentes de mistura gordura, carboidrato e fibra é uma alternativa viável a ser utilizada em relação ao modelo proposto por Akay e Tez (2011).

Uma informação importante que pode ser fornecida por um modelo de mistura é a mistura que proporciona a máxima ou mínima incidência de tumor, respeitando-se as restrições de cada componente. Reportando o modelo que apresentou os melhores indicadores de qualidade de ajuste, o modelo M4, este afirma que a máxima incidência de tumor esperada é de $90,06\%$ e a mistura que proporciona esse valor é formulada por $70,51\%$ de gordura, $29,18\%$ de carboidrato e $0,30\%$ de fibra.

Tabela 4 Misturas dos componentes x_1 (gordura), x_2 (carboidrato) e x_3 (fibra) que proporcionam a máxima e mínima incidência de tumor esperada ($\hat{y}$)

Modelo *	**Máximo**				**Mínimo**			
	x_1	x_2	x_3	$\hat{y}$	x_1	x_2	x_3	$\hat{y}$
M4	0,7244	0,2726	0,0030	0,9116	0,1336	0,8634	0,0030	0,5508

* M4 (Regressão Simplex do tipo razão)

Por outro lado, a mínima incidência de tumor esperada é de $56,30\%$ e a mistura que proporciona esse valor é formulada por $13,43\%$ de gordura, $86,25\%$ de carboidrato e $0,32\%$ de fibra (Tabela 4). Ou seja, as maiores diferença entre

os componentes que maximizam e minimizam a resposta ficou caracterizado pela proporção de gordura e carboidrato na mistura.

4 CONCLUSÕES

O modelo de regressão simplex apresentou ajuste satisfatório na análise do experimento de mistura que avaliou a incidência de tumor mamário em ratos fêmeas, sendo uma opção viável na análise de situações onde a resposta é limitada. O uso desse modelo também contempla a sub ou super-dispersão presente em dados agrupados.

Os intervalos de confiança para a razão de chances evidenciaram que diferentes escolhas dos pontos de referência afetam severamente a razão de chances e o respectivo intervalo de confiança. Portanto, considerar o modelo de regressão simplex na análise de experimentos de mistura com resposta limitada sob a presença de sub ou super-dispersão proporcionou estimativas para a razão de chances mais precisas e o modelo com termo raiz-quadrada produziu intervalos de confiança para a razão de chances mais estáveis em diferentes pontos de referência na região experimental em que exista efeito de borda.

REFERÊNCIAS

AKAIKE, H. A new look at the statistical model identification, **IEEE Transactions on Automatic Control**, Boston, v. 19, n. 6, p. 716-723, 1974.

AKAY, K. U.; TEZ, M. Analyzing mixture experiments via generalized linear models, **International Journal of Pure Applied Mathematics**, v. 36, n. 3, p. 373 - 390, 2007.

AKAY, K. U.; TEZ, M. Alternative modeling techniques for the quantal response data in mixture experiments, **Journal of Applied Statistics**, v. 38, n. 11, p. 2597 - 2616, 2011.

ATKINSON, A. C. Plots, Transformations and Regression, **Oxford University Press**, Oxford, 1985.

AITCHISON, J.; BACON-SHONE, J. Log contrast models for experiments with mixtures, **Biometrika**, v. 71, s. 2, p. 323 - 330, 1984.

BANDORFF-NIELSEN, O. E.; JORGENSEN, B. Some parametric models on the simplex. **Journal of Multivariate Analysis**, v. 39, p. 106 - 116, 1991.

BOX, G. E. P; DRAPER, N. E. **Response surfaces, mixtures, and ridge analyses**, Wiley Series in Probability an Statistics, 2º ed., 874 p., 2007.

BOZDOGAN, H. Akaike's Information Criterion and recent developments in Information Complexity, **Journal of Mathematical Psychology**, v. 44, p. 62 - 91, 2000. doi:10.1006/jmps.1999.1277

CHEN, J. J.; LI, L. A.; JACKSON, C. D. Analysis of quantal response data from mixture expreriments, **Environmetrics**, v. 7, n. 5, p. 503-512, 1996.

CORNELL, J. A. **Experiments with Mixtures: Designs, Models, and the Analysis of Mixture Data**, 3º ed., John Wiley Sonc Inc., USA, 2002.

CORNELL, J. A.; GORMAN, J. W. Two new mixture models: living with collinearity but removing its influence, **Journal of Quality Technology**, v. 35, n. 1, p. 78-88, 2003.

CUMMINGS, P. Methods for estimating adjusted risk ratios, **The Stata Journal**, v. 9, n. 2, p. 175-196, 2009.

DRAPER, N. R.; ST JOHN, R. C. A mixture model with inverse terms. **Technometrics**, v. 19, n. 1, p. 37 - 46, 1977.

ESPINHEIRA, P. L.; FERRARI, S. L. P.; CRIBARI-NETO, F. On beta regression residuals. **Journal of Applied Statistics**, Routledge, v. 35, n. 4, p. 407 - 419, 2008.

HOSMER, D. W.; LEMESHOW, S. **Applied Logistic Regression**, 2º ed., John Wiley, New York, 2000.

JORGENSEN, B. **The theory of dispersion models**, Chapman and Hall, London, 1997.

LANNING, M. J.; BOZDOGAN, H. **Ordinal Logistic modeling using ICOMP as a goodness-of-fit criteria**, Statistical Data Mining and Knowledge Discovery, Chapman Hall/CRC, USA, p. 353-371, 2003.

MENARD, S. **Logistic Regression: From Introductory to Advanced Concepts and Applications**, Sage Publications, Inc., USA, 2009.

MYERS, R. H.; MONTGOMERY, D. C. A tutorial on generalized linear models, **Journal of Quality Technology**, v. 29, n. 3, p. 274-291, 1997.

NOCEDAL, J.; WRIGHT, S. J. **Numerical optimization**, Springer-Verlag, New York, 1999.

PIEPEL, G. F. Measuring component effects in constrained mixture experiments. **Technometrics**, v. 24, p. 29-39, 1982.

R CORE TEAM (2015). **R: A language and environment for statistical computing**, R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/.

SCHWARZ, G. Estimating the dimensional of a model, **The Annals of Statistics**, Hayward, v. 6, n. 2, p. 461-464, 1978.

SMITH, W. F. **Experimental Design for Formulation**, ASA-SIAM Series on Statistics and Applied Probability, SIAM, Philadelphia, ASA, Alexandria, VA,

2005.

SNEE, S. H. Techniques for the analysis of mixture data, **Technometrics**, v. 15, s. 3, p. 517 - 528, 1973.

WILLIAMS, D. A. Generalized linear model diagnostic using the deviance and single case deletion, **Applied Statistics**, v. 36, p. 181-191, 1987.

UNIVERSIDADE FEDERAL DE LAVRAS

ARTIGO 2

Acurácia e precisão de razões de chances obtidas por meio do modelo Boosted Simplex Regression aplicado em Experimentos de Mistura

Versão preliminar de artigo - Sujeito a alterações pelo corpo editorial da revista

Revista: Biometrical (IF: 0,945)

LAVRAS - MG

2016

RESUMO

Devido à sua ampla aplicação, muitos estudos tem sido feitos na construção de experimentos de mistura de componentes e determinação de modelos que contemplem diferentes naturezas para a resposta a ser otimizada e, entre elas, as restritas ao intervalo $(0, 1)$, comumente referenciadas por proporção. Decorrente a natureza dos dados, uma alternativa a ser utilizada é dada por ajustar o modelo de regressão simplex, o qual apresenta como vantagens em relação à regressão logística o fato de acomodar o efeito da sub ou super dispersão e contemplar estrutura de variância não-binomial. Contudo, ao utilizar o método da máxima verossimilhança, surgem problemas relacionados a colinearidade e tamanho amostral reduzido. Para contornar esse problema, o objetivo desse trabalho consistiu em propor um modelo denominado por Boosted Simplex Regression, o qual foi avaliado em termos de acurácia e precisão para a razão de chances, bem como um procedimento para construção de envelopes simulados. Para ilustrar a eficiência do método, foi ultilizado um conjunto de dados de um experimento de mistura sobre incidência de tumor mamário em ratos e nas comparações foram utilizados os critérios de seleção de modelos AIC e BIC, gráficos de envelope simulado para os resíduos dos modelos ajustados, gráficos da razão de chances e seu respectivo intervalos de confiança. Concluiu-se que o modelo *Boosted Simplex Regression* se ajustou satisfatoriamente e produziu intervalos de confiança para a razão de chances acurados e ligeiramente mais precisos do que sua versão clássica.

Palavras-chave: Modelo Linear Generalizado, Algoritmo Boosting, Modelo de Mistura, Região Simplex.

1 INTRODUÇÃO

O algoritmo AdaBoost proposto inicialmente por Freund e Shapire (1996) em problemas de classificação tem atraído muita atenção na comunidade de aprendizado de máquinas, bem como em áreas relacionadas à estatística. Várias versões desse algoritmo foram comprovadas como sendo bastante promissoras em termos da acurácia preditiva (LISKA et al., 2015). Friedman, Hastie e Tibshirani (2000) mostraram que o algoritmo AdaBoost pode ser visto como um algoritmo gradiente no espaço funcional, inspirado pela otimização numérica e estatística e Friedman (2001) mostrou que o algoritmo Boosting pode ser utilizado para ajuste de modelos (CAO et al., 2010).

Em se tratando de ajuste de modelos, o trabalho de Friedman, Hastie e Tibshirani (2000) abriu uma nova perspectiva quanto ao uso do algoritmo boosting para além dos problemas de classificação, que foi seu emprego em modelos de regressão, como mencionado em Buhlmann e Yu (2003). Nesse contexto, basicamente, o ajuste de um modelo de regressão poderá ser feito em termos da otimização de uma função perda no espaço funcional dos parâmetros do modelo avaliado. Além disso, os parâmetros dependem de um procedimento base, que pode ser uma função linear aditiva, a qual relaciona a resposta do modelo de regressão com as covariáveis.

No que diz respeito à estimação de parâmetros, essa abordagem traz vantagens em relação à estimação por máxima verossimilhança de parâmetros. Em conjuntos de dados que contém um grande número de variáveis preditoras em relação ao tamanho amostral, a variância dos estimador de máxima verossimilhança é inflacionada. Esse problema tem como consequência o super ajuste, o que diminui a precisão na predição de uma modelo clássico de regressão. Esse fato pode

ser verificado pela matriz de informação de Fisher do modelo. Em situações do tipo, o método de Boosting pode ser utilizado, uma vez que este automaticamente realiza seleção de variáveis (SCHMID et al., 2013).

Dadas essas especificações, os delineamentos usuais de mistura, como o Centróide-Simplex ou Vértice Extremo, são caracterizados por apresentarem tamanho amostral reduzido. Para tais delineamentos, o experimento de mistura consiste em otimizar uma variável resposta com a restrição de que

$$\sum_{i=1}^{q} x_i = 1, \tag{1.1}$$

em que x_i ($0 \leq x_i \leq 1$) representa a proporção do i-ésimo componente ($i = 1, 2, \ldots, q$), com q o número de componentes (BOX; DRAPER, 2007). O espaço que compõe os q componentes assume a forma de um simplex regular de dimensão $(q-1)$. No caso $q = 3$, a região simplex é um espaço de mistura triangular. Por questões econômicas e, ou, físicas, às vezes são impostas restrições adicionais sobre os componentes individuais

$$0 \leq L_i \leq x_i \leq U_i \leq 1;\ i = 1, 2, \ldots, q, \tag{1.2}$$

sendo L_i e U_i, respectivamente, os limites inferiores e superiores e a restrição (1.2) reduz a região restrita dada na equação (1.1).

A modelagem estatística usual desses experimentos é feita utilizando modelos polinomiais assumindo normalidade para a variável resposta (CORNELL, 2002). No caso de variáveis seguindo outras distribuições, tais modelos são adaptados utilizando-se modelos lineares generalizados (MLG). Especialmente, quando a variável resposta é proporção, o modelo de regressão binomial (logística) tem sido bastante utilizado, porém convém ressaltar que em geral, esse modelo não

acomoda o efeito da sub ou super-dispersão que frequentemente ocorre em dados agrupados.

No tocante a estimação da razão de chances, para esses experimentos os métodos convencionais de análise e interpretação dos parâmetros de um modelo de mistura não são adequados, uma vez que quando um componente é modificado, os outros componentes são alterados, devida a restrição unitária em 1.1. Para esse problema, a análise do efeito dos componentes de mistura consiste no uso da Direção Cox, descrita a seguir, adaptados para estimar os intervalos de confiança para a razão de chances em experimentos de mistura.

1.1 Direção Cox para o gráfico de resposta traço em Experimentos de Mistura

A direção Cox do componente i é uma linha imaginária projetada da mistura de referência para o vértice $x_i = 1$. As proporções dos q componentes na mistura de referência é $\mathbf{c} = (c_1, c_2, \ldots, c_q)$, em que $\sum_{i=1}^{q} c_i = 1$. O ponto de referência $\mathbf{c}$ padrão é geralmente adotado como o centróide do experimento. Quando a proporção c_i do componente i é alterada por uma quantidade Δ_i na direção Cox, então a nova proporção se torna

$$x_i = c_i + \Delta_i. \tag{1.3}$$

Quando existem apenas restrições como na expressão (1.1) sobre os componentes de mistura, Δ_i está no intervalo $[-c_i, 1 - c_i]$. Caso contrário, para restrições adicionais sobre os componentes como na expressão (1.2), Δ_i está no intervalo $[L_i - c_i, U_i - c_i]$. As proporções dos $q-1$ componentes restantes, resultante de c_i no i-ésimo componente, é

$$x_j = c_j \frac{1 - x_i}{1 - c_i},\ j = 1,\ 2,\ \ldots,\ q,\ j \neq i. \tag{1.4}$$

No caso de uma região experimental restrita ser um simplex regular, uma representação alternativa da direção Cox pode ser formulada, considerando o fato de que $\frac{x_j}{x_k} = \frac{c_j}{c_k}$. Nesse caso, analogamente à razão das proporções dos componentes escolhidos ao longo dos eixos para um componente na direção Cox, com a ajuda da razão de componentes em qualquer ponto na região restrita, a mudança na resposta pode ser avaliada. Por exemplo, em um sistema de mistura com $q = 3$, ao longo que o eixo do componente x_i passa através do ponto de referência, os componentes x_j e x_k sendo $x_j/x_k = \rho_{x_i}$, são dados por

$$x_j = \frac{\rho_{x_i}(1 - x_i)}{\rho_{x_i} + 1} \quad \text{e} \quad x_k = \frac{1 - x_i}{\rho_{x_i} + 1},$$

em que, $1 - c_i = c_j + c_k$, $L_i \leq x_i \leq U_i$ e ρ_{x_i} mostra a razão dos componentes exceto x_i no ponto referência. Dessa forma, o valor da resposta predita para o preditor linear de primeiro grau, ao longo da direção Cox para o i-ésimo componente é dado por

$$logit\left(\hat{\pi}_{x_i}\right) = \hat{\beta}_i x_i + \hat{\beta}_j \frac{\rho_{x_i}(1 - x_i)}{\rho_{x_i} + 1} + \hat{\beta}_k \frac{1 - x_i}{\rho_{x_i} + 1},\ L_i \leq x_i \leq U_i. \tag{1.5}$$

Na equação (1.5), exponenciando $logit\left(\hat{\pi}_{x_i}\right)$, os valores preditos de $\hat{\pi}_{x_i}$ para cada componente de mistura x_i são obtidos. Com esses valores, pode-se construir um gráfico que relaciona os incrementos de x_i com os valores de $\hat{\pi}_{x_i}$ e esse gráfico é chamado de *resposta traço* (*trace plot*). Esses traços representam o efeito da mudança de cada componente de mistura enquanto todos os outros

componentes permanecem em razão constante. O modelo dado pela equação (1.5) pode ser expandido para diferentes preditores lineares (AKAY; TEZ, 2007).

1.2 Gráficos da razão de chances para os componente em Experimentos de Mistura

Considere um ponto qualquer $\mathbf{c} = (c_i, c_j, c_k)$ que é tomado como um grupo controle sobre a região experimental, a razão de chances é dada ao longo do eixo x_i por

$$\widehat{OR(x_i)} = \frac{chance\, x_i}{chance\, controle} = \frac{\exp\left\{\hat{\beta}_i x_i + \hat{\beta}_j \frac{\rho_{x_i}(1-x_i)}{\rho_{x_i}+1} + \hat{\beta}_k \frac{1-x_i}{\rho_{x_i}+1}\right\}}{\exp\left\{\hat{\beta}_i c_i + \hat{\beta}_j c_j + \hat{\beta}_k c_k\right\}}; \; L_i \leq x_i \leq U_i. \tag{1.6}$$

A expressão 1.6 pode ser reescrita por $\widehat{OR(x_i)} = \exp\left\{\hat{\beta}_i A + \hat{\beta}_j B + \hat{\beta}_k C\right\}$, em que $A = x_i - c_i$, $B = \frac{\rho_{x_i}(1-x_i)}{\rho_{x_i}+1} - c_j$ e $C = \frac{1-x_i}{\rho_{x_i}+1} - c_k$, a qual corresponde a uma estimativa pontual, assumindo valores entre zero e infinito. Se a razão de chances é igual a um, isso quer dizer que não existe diferença entre os grupos comparados, e se for diferente de um, significa que existe diferença. Além disso, as razões preditas obtidas pela equação (1.6) formarão uma curva que pode ser interpretada na comparação entre grupos.

A precisão da razão de chances pode ser verificada pelo intervalo de confiança e sua amplitude reflete sua variabilidade inerente. Para calcular o erro padrão e o intervalo de confiança para a razão de chances, é necessário transformar a equação (1.6) para a escala logarítmica e, nesse caso, segundo Hosmer, Lemeshow e Sturdivant (2013) a distribuição amostral dessa medida é aproximadamente normal. Observe que a expressão para o logaritmo neperiano da razão das chances na equação (1.6) depende do componente de mistura x_i e o coeficiente estimado $\hat{\beta}_i$. Utilizando métodos para calcular a variância de uma soma, obtemos o seguinte

estimador

$$\begin{aligned} V\hat{a}r\left\{\ln\left[\widehat{OR\left(x_i\right)}\right]\right\} =& A^2V\hat{a}r\left[\hat{\beta}_i\right] + B^2V\hat{a}r\left[\hat{\beta}_j\right] + C^2V\hat{a}r\left[\hat{\beta}_k\right] \\ &+ 2ABC\hat{o}v\left[\hat{\beta}_i,\hat{\beta}_j\right] + 2ACC\hat{o}v\left[\hat{\beta}_i,\hat{\beta}_k\right] \\ &+ 2BCC\hat{o}v\left[\hat{\beta}_j,\hat{\beta}_k\right], \end{aligned} \tag{1.7}$$

eem que as variâncias e covariâncias dos parâmetros do modelo ajustado são provenientes da inversa da matriz de informação de Fisher. Uma vez obtida a estimativa da variância do logaritmo neperiano da razão das chances estimado, pode-se obter o estimador do intervalo de $100\,(1-\alpha)\,\%$ de confiança para o logaritmo neperiano da razão das chances, que é dado por,

$$IC_{(1-\alpha)}\left\{\ln\left[\widehat{OR\left(x_i\right)}\right]\right\} = \left[\hat{\beta}_iA + \hat{\beta}_jB + \hat{\beta}_kC\right] \pm z_{(\frac{\alpha}{2})}\sqrt{V\hat{a}r\left\{\ln\left[\widehat{OR\left(x_i\right)}\right]\right\}}, \tag{1.8}$$

em que, $\sqrt{V\hat{a}r\left\{\ln\left[\widehat{OR\left(x_i\right)}\right]\right\}}$ é o erro padrão do estimador do logaritmo neperiano da razão de chances, $z_{(\frac{\alpha}{2})}$ é o $(\frac{\alpha}{2})$-ésimo quantil da distribuição normal padrão com α nível de significância.

Diante do exposto, o presente trabalho tem por objetivo empregar o uso do modelo de regressão simplex na análise de experimento de mistura, considerando-se que a variável resposta é proporção, e que tem a vantagem de modelar a sub ou super-dispersão de dados. Será proposto o algoritmo *Boosted Simplex Regression*, a versão boosting da regressão simplex, o qual será avaliado em termos de acurácia e precisão para razões de chances. Nesse sentido, as vantagens da abordagem proposta foi ilustrada em um experimento descrito na seção 2.1, que consistiu em

estudar o efeito de diferentes dietas compostas por gordura, carboidrato e fibra sobre a expressão de tumor nas glândulas mamárias em ratos fêmeas.

2 MATERIAL E MÉTODOS

2.1 Descrição do experimento

Seguindo o experimento descrito por Akay e Tez (2011), a variável resposta referiu-se a ocorrência de tumor em ratos, com o propósito de estudar os efeitos de calorias da dieta composta por gordura, carboidrato e fibra sobre a expressão (promoção) de tumor nas glândulas mamárias induzido por *Dimetil-benzathracene* (DMBA) em ratos fêmeas. Os dados são apresentados na Tabela 1.

Tabela 1 O número observado de tumores induzidos DMBA em glândulas mamárias de ratos tratados com diferentes proporções calóricas de fibra, gordura e carboidrato (componentes originais). Cada grupo é composto por 30 ratos e tem igual quantidade de calorias totais.

Dietas	Componentes Originais			Ratos c/ Tumor	Proporção de Tumor (y)
	Gordura (x_1)	Carboidrato (x_2)	Fibra (x_3)		
1	0,175	0,775	0,050	17	0,567
2	0,153	0,820	0,027	15	0,500
3	0,133	0,863	0,004	17	0,567
4	0,491	0,470	0,039	24	0,800
5	0,440	0,538	0,022	21	0,700
6	0,390	0,607	0,003	23	0,767
7	0,701	0,267	0,032	18	0,600
8	0,638	0,343	0,019	23	0,767
9	0,576	0,421	0,003	26	0,867

A proporção das três dietas foram restritas pelos seguintes limites inferiores e superiores: $0,133 \leq Gordura \leq 0,730$; $0,267 \leq Carboidrato \leq 0,864$ e $0,003 \leq Fibra \leq 0,600$. A Tabela 1 contém as respostas das proporções de

tumor observadas dos nove grupos de dieta com diferentes proporções calóricas de gordura (x_1), carboidrato (x_2) e fibra (x_3). As dietas 1 a 3 são constituídas de baixa gordura e alto carboidrato, as dietas 4 a 6 são constituídas de valores médios de gordura e carboidrato, as dietas 7 a 9 são possuem alta gordura e baixo carboidrato. As dietas 1, 4 e 7 são altas em fibras, as dietas 2, 5 e 8 são médias em fibras e as dietas 3, 6 e 9 são baixas em fibras. Em todas as dietas, a gordura e o carboidrato são as duas maiores origens de calorias.

2.2 Modelos de regressão aplicados aos experimentos de mistura

A proposta de experimentos de mistura é construir um modelo apropriado que relacione a resposta aos componentes $x_1, x_2, \ldots, x_q$. Assume-se que a resposta de interesse η seja uma função das variáveis de mistura x_i, ou seja, $\eta = f(x_1, x_2, \ldots, x_q)$. Quando um experimento é feito, é natural considerar que as respostas observadas, denotadas por y_i para o i-ésimo valor $(i = 1, 2, \ldots, n)$ são funções da média de η_i com uma variância constante σ^2 para todo $i = 1, 2, \ldots, n$. As respostas observadas contém o erro experimental aditivo ε_i, ou seja, $y_i = \eta_i + \varepsilon_i$, em que $\varepsilon_i \overset{iid}{\sim} N\left(0; \sigma^2\right)$. A forma funcional da resposta $E[Y] = f(x_1, x_2, \ldots, x_q)$ geralmente não é conhecida, mas em várias situações os modelos de aproximação polinomial de primeiro e segundo graus são utilizados.

Os modelos linear e quadrático de Scheffé são os mais comumente utilizados em experimentos de misturae (SCHEFFÉ, 1958). Contudo, para as respostas dos componentes que se aproximam da fronteira do simplex, esses modelos não são viáveis (DRAPER; ST. JOHN, 1977; SNEE, 1973; AITCHISON; BACONSHONE, 1984; CORNELL; GORMAN, 2003). Assim sendo, para que os efeitos curvilíneos sejam contemplados mediante ao comportamento extremo das respostas, ou também chamado de *efeito de borda*, procedeu-se com a incorporação dos

termos inversos definidos pelas variáveis razão $w_i = \sqrt{\frac{x_i}{x_q^*}}$, em que x_q^* corresponde ao componente de mistura que provoca o *efeito de borda* (AKAY; TEZ, 2011). Desta forma os modelos (2.3) e (2.4) são especificados na Tabela 2.

Tabela 2 Classificação dos modelos de mistura mais usuais.

Modelo	$E[Y]$	
Linear de Scheffé	$\sum_{i=1}^{q} \hat{\beta}_i x_i$	(2.1)
Quadrático de Scheffé	$\sum_{i=1}^{q} \hat{\beta}_i x_i + \sum_{i=1}^{q} \sum_{i<j}^{q} \hat{\beta}_{ij} x_i x_j$	(2.2)
Linear com variáveis razão	$\hat{\beta}_0 + \sum_{i=1}^{q-1} \hat{\beta}_i w_i$	(2.3)
Quadrático com variáveis razão	$\hat{\beta}_0 + \sum_{i=1}^{q-1} \hat{\beta}_i w_i + \sum_{i=1}^{q-1} \sum_{i<j}^{q-1} \hat{\beta}_{ij} w_i w_j$	(2.4)

FONTE: Box e Draper (2007)

2.3 Adaptação da distribuição Simplex e os modelos de mistura em MLG

Uma distribuição que pode ser utilizada para estudar uma variável resposta contínua e restrita ao intervalo $(0, 1)$ é a distribuição simplex (BANDORFF-NIELSEN; JORGENSEN, 1991; JORGENSEN, 1997).

Considerando que a resposta do experimento, representada pela variável aleatória y com média $\mu \in (0, 1)$ e parâmetro de dispersão $\sigma^2 > 0$ tem função de densidade dada por

$$f\left(y; \mu, \sigma^2\right) = \frac{1}{\sqrt{2\pi\sigma^2 \left\{y\left(1-y\right)\right\}^3}} \exp\left\{-\frac{1}{2\sigma^2} d\left(y; \mu\right)\right\}, \; y \in \left(0, 1\right), \quad (2.5)$$

em que,

$$d\left(y;\mu\right)=\frac{\left(y-\mu\right)^{2}}{y\left(1-y\right)^{2}\mu^{2}\left(1-\mu\right)^{2}}. \tag{2.6}$$

Dada uma amostra aleatória $y_1,\ldots,y_n$ de variáveis aleatórias independentes, sendo que cada $y_i \sim S\left(\mu_i,\sigma^2\right)$, $i=1,\ldots,n$. O modelo de regressão simplex é definido pela densidade da forma (2.5), para as médias μ_i caracterizadas por $g\left(\mu_i\right)=\mathbf{x}_i^T\boldsymbol{\beta}=\eta_i$, sendo $g\left(\cdot\right)$ a função de ligação "logit", $\boldsymbol{\beta}=\left(\beta_1,\ldots,\beta_p\right)$ é o vetor dos parâmetros da regressão ($\boldsymbol{\beta}\in\mathbb{R}^p$), $\mathbf{x}_i^T=\left(x_{1i},\ldots,x_{pi}\right)$ são os valores conhecidos das p covariáveis no modelo de mistura. Assim, η_i corresponde ao preditor linear, definido por um dos modelos descritos na Tabela (2).

Seguindo essas especificações, utilizou-se o método da máxima verossimilhança, considerando a log-verossimilhança de cada observação definida por:

$$l_i\left(\mu_i,\sigma^2\right)=-\frac{1}{2}\ln\left(2\pi\right)-\frac{1}{2}\ln\left(\sigma^2\right)-\frac{3}{2}\ln\left[y_i\left(1-y_i\right)\right]-\frac{1}{2\sigma^2}d\left(y_i;\mu_i\right). \tag{2.7}$$

Logo, segundo o logaritmo da função de verossimilhança para de n observações independentes, dado por $l\left(\boldsymbol{\beta},\sigma^2\right)=\sum_{i=1}^{n}l_i\left(\mu_i,\sigma^2\right)$, os estimadores de máxima verossimilhança dos parâmetros $\boldsymbol{\beta}$ e σ^2 são obtidos por meio da solução do sistema de equações homogêneo derivado de $l\left(\boldsymbol{\beta},\sigma^2\right)$.

2.3.1 Algoritmo Boosting aplicado ao modelo de Regressão Simplex

O propósito do algoritmo Gradiente Boosting de Friedman consiste em estimar uma função de predição ótima $f^*\left(\cdot\right)$, definida por

$$f^*\left(\cdot\right)=\arg\min_{\eta}E_{Y,\mathbf{X}}\left[\rho\left\{Y,g\left[\eta\left(\mathbf{X}\right)\right]\right\}\right] \tag{2.8}$$

onde, $\rho(\cdot,\,\cdot)$ é uma função perda. Desta forma, o algoritmo boosting para ajuste de MLG proposto for Friedman (2001), foi adaptado para a regressão simplex e denominado de *Boosted Simplex Regression*, descrito nos passos a seguir:

(i) Inicialize $\hat{f}^{(0)}(\cdot)$ com um valor inicial e $m = 0$. Escolhas comuns são

$$\hat{f}^{(0)}(\cdot) = \arg\min_{c} \frac{1}{n}\sum_{i=1}^{n} \rho(Y_i,\, c)$$

ou $\hat{f}^{(0)}(\cdot) = 0$. A função perda é o negativo da log-verossimilhança da densidade da distribuição simplex, ou seja,

$$\rho\{Y, g[\eta(\mathbf{X})]\} = -l\left(Y, g^{-1}(\eta(\mathbf{X})), \sigma^2\right),$$

como na expressão 2.7.

(ii) Aumente m em 1. Calcule o gradiente negativo $-\frac{\partial}{\partial g}\rho(Y,\, g)$ e calcule em $\hat{f}^{(m-1)}(\boldsymbol{X}_i)$:

$$\mathbf{u}^{(m)} = \left(u_i^{(m)}\right)|_{i=1,\ldots,n} = -\frac{\partial}{\partial g}\rho\{Y_i, g[\eta(\mathbf{X}_i)]\}\Big|_{g(\eta(\mathbf{X}_i))=\hat{f}^{(m-1)}(\mathbf{X}_i)},$$

em que o vetor gradiente $\mathbf{u}^{(m)} = \left(u_i^{(m)}\right)\Big|_{i=1,\ldots,n}$ representa uma estimativa do verdadeiro gradiente da esperança em (2.8). Logo, o algoritmo *Boosted Simplex Regression* faz com que o vetor gradiente negativo percorra iterativamente no espaço de $g(\cdot)$ para minimizar o risco empírico. Segundo Hofner et al. (2014), no algoritmo Gradiente Boosting de Friedman essa estratégia corresponde a substituir o clássico método Escore-Fisher em estimação por Máxima Verossimilhança de f^* pelo algoritmo gradiente descendente no espaço funcional de f^*.

(iii) Ajuste o vetor gradiente negativo $u_1, \ldots, u_n$ para $\mathbf{X}_1, \ldots, \mathbf{X}_n$ por um procedimento base $\hat{h}^{(m)}(\cdot)$, dado por $\hat{h}^{(m)} = \hat{\beta}^{[\hat{\lambda}]} x^{[\hat{\lambda}]}$, em que

$$\hat{\beta}^{(j)} = \frac{\sum_{i=1}^{n} x_i^{(j)} u_i}{\sum_{i=1}^{n} \left(x_i^{(j)}\right)^2}. \tag{2.9}$$

Note que são ajustados modelos lineares simples sem intercepto separadamente para cada coluna da matriz de delineamento ao vetor gradiente negativo, u_i, (2.9), ou seja, para cada covariável ou componente de mistura j ($j = 1, \ldots, p$), sendo o j-ésimo parâmetro do modelo de mistura da Tabela 2. Logo, haverão p procedimentos bases ajustados na iteração m do algoritmo.

Escolher o componente $\hat{\lambda}$ com melhor ajuste ao vetor gradiente negativo de acordo com o critério dos mínimos quadrados. Ou seja, selecione a covariável $\hat{h}_{(j)}^{(m)[\hat{\lambda}]}$ definida por

$$\hat{\lambda} = \underset{1 \leq j \leq p}{\arg\min} \sum_{i=1}^{n} \left(u_i^{(m-1)} - \hat{h}_{i(j)}^{(m)}\right)^2 \tag{2.10}$$

em que, $\hat{h}_{(j)}^{(m)} = \left(\hat{h}_{i(j)}^{(m)}\right)_{i=1,\ldots,n}$ são os valores ajustados dos procedimentos bases $\hat{h}_{i(j)}^{(m)}$ para as observações $i = 1, \ldots, n$ na m-ésima iteração. Convém ressaltar que dados os p procedimentos bases ajustados, o melhor modelo ajustado é utilizado no processo de atualização, ou seja, apenas o procedimento base que fornecer a menor quantidade em (2.10) será adicionado à corrente iteração (passo (iv)) do algoritmo. Por esse motivo, diz-se que o procedimento de seleção de variáveis está embutido no algoritmo (NATEKIN; KNOLL, 2013).

(iv) Atualize o preditor aditivo

$$\hat{f}^{(m)}(\cdot) = \hat{f}^{(m-1)}(\cdot) + v \cdot \hat{h}_{(j)}^{(m)[\hat{\lambda}]}$$

em que $0 < v \leq 1$ é o fator *comprimento do passo*. Recomenda-se que v seja pequeno, como por exemplo $v = 0,1$ (BUHLMANN; HOTHORN, 2007).

(v) Continue o processo de iteração entre os passos (ii) a (iv) até $m = M$, para alguma iteração de parada M.

A iteração de parada é um importante argumento de controle e, nesse trabalho será determinada pelos critérios de informação de Akaike (AIC) e bayesiano (BIC). Dessa forma, a função de predição ótima é aquela cujo AIC e, ou, BIC é mínimo (HOFNER et al., 2014).

Diante do exposto, a função de predição ótima ou minimizador populacional $f^*(\cdot)$, em modelos lineares generalizados é a função $g(\eta(\mathbf{X}))$ que retorna o menor risco empírico.

Seguindo essas especificações descritas nos passos (i) até (v), utilizou-se o pacote *mboost*, sendo esse adaptado para implementar a regressão simplex (BUHLMANN; HOTHORN, 2007).

2.3.2 Construção dos gráficos de envelope simulado para o modelo de Regressão Simplex

Após ajuste do modelo, a construção do gráfico de envelope simulado para o modelo de regressão simplex foi feita utilizando-se do seguinte procedimento (WILLIAMS, 1987):

(i) Gerou-se n observações da distribuição $S\left(\hat{\mu}_i, \hat{\sigma}^2\right)$;

(ii) Obteve-se o resíduo t_i^*, o resíduo padronizado ponderado (ESPINHEIRA; FERRARI; CRIBARI-NETO, 2008);

(iii) Repete-se os passos (1) - (2) k vezes, resultando os resíduos gerados t_{ij}^*, $i = 1, \ldots, n$ e $j = 1, \ldots, k$;

(iv) Coloca-se cada grupo de n resíduos em ordem crescente, obtendo $t_{(i)j}^*$, $i = 1, \ldots, n$ e $j = 1, \ldots, k$;

(v) Obtém-se os limites $t_{(i)I}^* = \min_j t_{(i)j}^*$ e $t_{(i)S}^* = \max_j t_{(i)j}^*$. Assim, os limites correspondentes ao i-ésimo resíduo foram dados por $t_{(i)I}^*$ e $t_{(i)S}^*$.

De maneira semelhante, foram construídos os gráficos de envelope simulado para os modelos ajustados via algoritmo boosting. Para tal, a obtenção dos resíduos t_i^*, que dependem de h_{ii}, também chamado de "leverage", no passo (2), foi utilizado o procedimento de obtenção da matriz $\mathbf{H}$ descrito em Buhlmann e Hothorn (2007).

2.3.3 Procedimento numérico para obter intervalo de confiança para razão de chances no modelo *Boosted Simplex Regression*

Considere $\hat{\theta}$ como sendo o estimador de $\ln\left[\widehat{OR\left(x_i\right)}\right]$. O intervalo de confiança para a razão de chances foi obtido pelo método de Monte Carlo (RIZZO, 2007), seguindo-se os seguintes passos:

(i) Gerar uma amostra de tamanho n da distribuição Simplex com as estimativas

de máxima verossimilhança $\hat{\mu}$ e $\hat{\sigma}^2$, sendo

$$\hat{\mu}_i = \frac{\exp\left\{\boldsymbol{x}_i\hat{\boldsymbol{\beta}}\right\}}{1+\exp\left\{\boldsymbol{x}_i\hat{\boldsymbol{\beta}}\right\}} \quad \text{e} \quad \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} d\left(y_i;\hat{\mu}_i\right),$$

em que, $\boldsymbol{x}_i\hat{\boldsymbol{\beta}}$ é o preditor linear do MLG definido na Tabela 2, $\hat{\boldsymbol{\beta}}$ a estimativa de máxima verossimilhança de $\boldsymbol{\beta}$ e $d\left(y_i;\hat{\mu}_i\right)$ de acordo com a equação (2.6). Para o modelo do tipo razão gerar a amostra por meio do delineamento Vértice Extremo. Mais detalhes desse delineamentos podem ser vistos em Cornell (2002);

(ii) Da amostra gerada em (i), calcular $\hat{\theta}$;

(iii) Repetir os passos (i) e (ii) B vezes;

(iv) A partir do vetor $\hat{\boldsymbol{\theta}}^* = \left(\hat{\theta}^*_{(1)} \leq \hat{\theta}^*_{(2)} \leq \ldots \leq \hat{\theta}^*_{(B)}\right)$, para algum nível de significância α, o intervalo de confiança Monte Carlo com $100\times(1-\alpha)\,\%$ de confiança é dado por

$$IC_{(1-\alpha)}\left(\theta\right) = \left[\hat{\theta}^*_{(k_1)};\, \hat{\theta}^*_{(k_2)}\right], \tag{2.11}$$

em que $k_1 = (B+1)(\alpha/2)$ e $k_2 = (B+1)(1-\alpha/2)$ são os maiores inteiros não maiores que $(B+1)(\alpha/2)$ e $(B+1)(1-\alpha/2)$, respectivamente e $\hat{\theta}^*_{(k_1)}$ é o percentil $100\,(\alpha/2)\,\%$ e $\hat{\theta}^*_{(k_2)}$ o percentil $100\,(1-\alpha/2)\,\%$.

Os limites inferiores e superiores do estimador do intervalo de confiança para a razão de chances são obtidos exponenciando os limites nas equações (1.8) e (2.11).

Finalizando a metodologia proposta, para obtenção dos resultados foram

utilizados os pacotes estatísticos, *bbmle*, *mixexp* e *mboost* do Sistema Computacional Estatístico R (R CORE TEAM, 2015).

3 RESULTADOS

3.1 Seleção do Modelo e desempenho do algoritmo Boosting

Os resultados descritos na Tabela 3 correspondem às estimativas dos parâmetros dos modelos de regressão simplex e *boosted simplex regression* ajustados em função do delineamento de mistura (Tabela 1) e os critérios de informação utilizados. As estimativas dos parâmetros não apresentaram discrepâncias entre elas e os critérios de informação foram similares. Ambos resultados são melhores do que os apresentados por Akay e Tez (2011) (modelos M1 e M2), que ajustaram o modelo de regressão logística com variáveis razão como uma melhor alternativa aos modelos de regressão logística propostos por Chen, Li e Jackson (1996), os quais fizeram estudos considerando as variáveis em pseudo-componentes nos modelos polinomiais de Scheffé e Backer, considerando o mesmo experimento. Assim, por meio dessa comparação há evidências estatísticas para afirmar que o modelo de regressão simplex é uma alternativa viável e promissora de ser aplicada a experimentos de mistura.

Convém ressaltar que em experimentos de mistura a seleção de variáveis em um modelo linear não faz sentido porque obrigatoriamente todos os componentes de mistura devem fazer parte do modelo. Assim, o uso de critérios de informação dar-se-á na comparação de modelos em diferentes métodos de estimação, sendo os métodos da máxima verossimilhança e boosting empregados nesse trabalho. Nesse caso, sucessivos ajustes do modelo são feitos e seleciona-se aquele

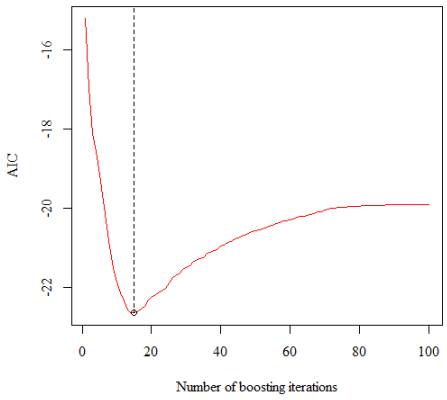
Tabela 3 Estimativas dos parâmetros para os modelos ajustados, resultados dos indicadores de qualidade de ajuste e número ótimo de iterações (M) do modelo *Boosted Simplex Regression*.

Modelo	Tipo	Parâmetro *	Estimativa	Erro Padrão	M	AIC	BIC
Regressão Simplex							
M1	Razão	β_0	0,2959	0,2006	-	-18,80	-21,51
		β_1	0,1779	0,0361			
		β_2	-0,0754	0,0330			
Boosted Simplex Regression							
M2	Razão	β_0	0,3260	-	15	-22,65	-22,44
		β_1	0,1691	-			
		β_2	-0,0712	-			
Regressão Logística							
M3	Razão	β_0	0,2895	0,2655	-	44,41	45
		β_1	0,1808	0,0542			
		β_2	-0,0767	0,0402			

* Parâmetros relacionados ao modelo linear de Razão, com preditor linear dado por $\eta = \beta_0 + \beta_1\sqrt{\frac{x_1}{x_3}} + \beta_2\sqrt{\frac{x_2}{x_3}}$.

que retornar o menor AIC, o que implica em eliminar as variáveis com menor contribuição para a explicação da resposta. Essa tarefa pode ser bastante árdua quando se tem um grande número de componentes, uma vez que para o modelo de regressão simplex isso deve ser feito manualmente. Uma vantagem do algoritmo *boosted simplex regression* é sua capacidade de realizar automaticamente a seleção de variáveis. No entanto, para fins de comparação, foram consideradas as mesmas covariáveis no ajuste via boosting do que pelo método da máxima verossimilhança. Isso implica que os modelos ajustados via boosting foram aqueles que retornam o mínimo AIC ou BIC dado o conjunto de covariáveis de acordo com a Tabela 3, conforme pode ser visto na Figura 1.

Smith (2005) fez críticas com relação à análise feita por Chen, Li e Jackson (1996) em termos de colinearidade. Por exemplo, a estimativa do parâmetro para o componente x_3 é maior que a dos outros parâmetros. Menard (2009) alerta



(a) M2

Figura 1 Número ótimo de iterações do algoritmo Boosted Simplex Regression segundo o critério de informação de Akaike (AIC) considerando o modelo M2 (Boosted Simplex Regression do tipo razão).

que a razão para essa discrepância entre as estimativas dos parâmetros do modelo é devida à colinearidade. Conforme menciona Bozdogan (2000), sistemas cujas correlações entre seus componentes são mais evidentes, tendem a inflacionar os erros-padrão das estimativas dos parâmetros do modelo avaliado. Dessa forma, o modelo M1 é preferível e tem maior informação sobre os parâmetros, uma vez que as variâncias dos mesmos são menores do que as do modelo M3.

Além do exposto, Akay e Tez (2011) fizeram uma observação quanto à presença do efeito da sub ou super-dispersão em dados agrupados, como é o caso dos dados Tabela 1, e mencionaram que esse fato deve ser levado em consideração na seleção do modelo. No trabalho de Chen, Li e Jackson (1996) esse fato foi desprezado e Akay e Tez (2011) apresentaram os modelos de regressão logística do tipo razão como uma melhor alternativa para controlar esse efeito. Esse fato corrobora o uso de modelos da classe dos modelos lineares generalizados como

sendo uma metodologia adequada na análise de respostas diferentes da abordagem usual.

No que diz respeito aos dados agrupados, assume-se que das n_i observações em um determinado grupo (ou dieta), todas elas tenham a mesma probabilidade π_i de ter o atributo de estudo (a ocorrência de tumor mamário), então é razoável assumir que a distribuição de Y_i seja binomial com parâmetros π_i e n_i. O caso de dados não agrupados pode ser considerado como sendo um caso especial, onde $n_1 = n_2 = \cdots = n_n = 1$.

A abordagem usual compreende os modelos de regressão linear com os erros sendo distribuídos normalmente, porém se a variável resposta no modelo de regressão é uma proporção ou uma razão, os pressupostos do modelos de regressão clássico podem estar comprometidos. Na tentativa de contornar esse problema, transformações como a do tipo arcseno podem fazer com que a distribuição dos erros sejam aproximadamente normal e variância constante. Contudo, Warton e Hui (2013) alertam para o fato de que, em estudos de simulação, a transformação arcseno perde poder em detectar proporções verdadeiras, o que compromete a predição do modelo. Por outro lado, o modelo de regressão logística apresentou maior poder quando a resposta é proveniente de uma distribuição binomial. Desse modo, considerar o modelo que melhor contemple a natureza da resposta é melhor do que aplicar alguma transformação na mesma.

Em contrapartida, o modelo de regressão logística é apropriado se a resposta é baseada na distribuição binomial, mas esse modelo não admite um parâmetro de dispersão, a menos que ele seja generalizado para acomodar esse parâmetro, como os modelos beta-binomial e os de quasi-verossimilhança. Dentre os modelos que naturalmente apresentam um parâmetro de dispersão, estão os modelos de regressão beta e simplex. O modelo de regressão beta tem se tornado ultimamente

uma alternativa interessante na análise de proporções e respostas limitadas, contudo, a média e o parâmetro de dispersão são dependentes e, dependendo do contexto, a estimação dos parâmetros pode ser comprometida (FERRARI; CRIBARI-NETO, 2004). No presente estudo, o modelo de regressão beta implementado por Cribari-Neto e Zeileis (2010) no software R não se ajustou aos dados da Tabela 1, em que o mesmo acusou inversa singular para a matriz de informação de Fisher.

Geralmente, proporções e razões tem distribuição assimétrica e estão restritas ao intervalo unitário, logo assumir normalidade para os erros pode levar a predições fora do intervalo unitário, ou seja, pode fornecer predições negativas ou maiores do que um. Além disso, nessa abordagem a estimação intervalar e testes de hipóteses assume-se que a resposta é distribuída simetricamente, o que pode levar as conclusões espúrias (SOUZA; CRIBARI-NETO, 2015). Diante dessas considerações, o modelo de regressão simplex se torna uma alternativa viável na modelagem de proporções e ele tem a vantagem de ser menos sujeito a problemas numéricos em relação ao modelo de regressão beta.

Dadas essas considerações, a estimativa para parâmetro de dispersão dos modelos M1 e M3, são dados por $\hat{\phi}_{M1} = 0,7291$ e $\hat{\phi}_{M3} = 0,9434$. Esses valores compõem o vetor escore e a matriz hessiana, influenciando portanto na estimativa e precisão dos parâmetros do modelo. Assim, novamente em comparação ao modelo proposto por Akay e Tez (2011), o modelo M1 forneceu estimativas mais precisas para os parâmetros.

Examinando os pontos experimentais (Figura 2) e referindo-se ao delineamento apresentado na Tabela 1, observou-se que a resposta sofreu uma mudança nos valores próximos do limite inferior do componente x_3 (fibra).

Nos pontos em que o componente x_1 (gordura) tem um valor alto, por exemplo nas dietas 4, 6, 8 e 9, um aumento no número observado de tumores é

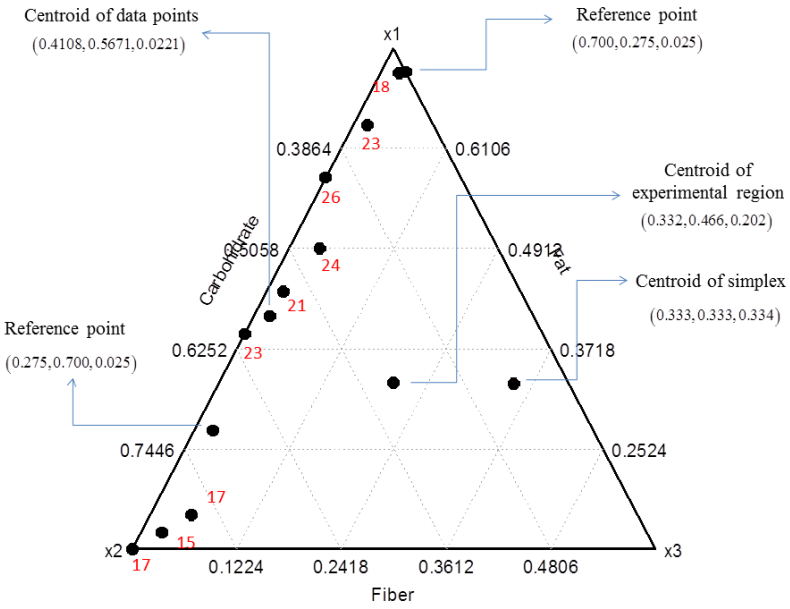


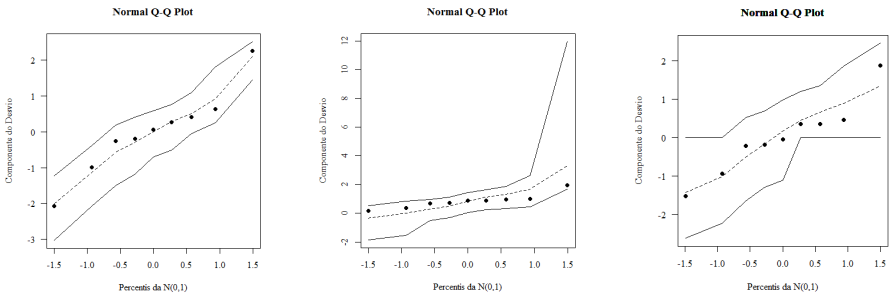
Figura 2 Região simplex restrita dos grupos de ratos tratados com diferentes proporções calóricas de fibra, gordura e carboidrato (componentes originais) e pontos de referência utilizados.

notório. Contudo, nos pontos em que o componente x_1 assumiu valores menores, observou-se um decréscimo no número de tumores. Quando o componente x_2 tem valores maiores, por exemplo nas dietas 1, 2 e 3, o número de tumores é reduzido e quando esse componente é caracterizado por menores valores, existe um aumento na resposta (Figura 2).

Para modelar um comportamento dos componentes desse tipo, utilizar a razão entre os componentes é uma abordagem mais realística, uma vez que modelos com esses tipos de variáveis permitem modelar mudanças quando um dos componentes assume valores menores, como é o caso de x_3 (fibra). Akay e Tez (2011) concluiram que o modelo de regressão logística com variáveis razão utili-

zando a transformação raiz quadrada (M3) foi o melhor dentre os utilizados.

Diante do exposto, pode-se afirmar que o modelo *Boosted Simplex Regression* apresentou melhores indicadores de qualidade de ajuste para a proporção de tumor mamário em ratos fêmeas (Tabela 3). Além disso, os gráficos normais de probabilidade dos resíduos componente do desvio para os modelos avaliados afirmam que a suposição de resposta simplex para a resposta analisada está adequada e o ajuste dos modelos foram satisfatórios (Figura 3).



(a) M1 (b) M2 (c) M3

Figura 3 Envelopes simulados para diagnóstico de qualidade de ajuste dos modelos M1 (Regressão Simplex do tipo razão - (a)), M2 (Boosted Simplex Regression do tipo razão - (b)) e M3 (Regressão Logística do tipo razão - (c)) .

3.2 Discussão dos modelos em relação aos gráficos dos efeitos dos componentes de mistura

Convém ressaltar que o presente trabalho não se ateve às motivações biológicas para a escolha dos referidos pontos, mas sim ao fato de tal escolha ter sido feita estritamente por inspeção da região experimental e aplicabilidade em experimentos de mistura. Para proceder às comparações dos modelos relacionados na

Tabela 3, foi calculado via simulação Monte Carlo as estimativas dos parãmetros dos modelos M1, M2 e M3. Nesse sentido, a Tabela 4 evidencia considerável acurácia entre as estimativas dos parâmetros, uma vez que esses produziram pequeno desvio em relação ao vetor $\boldsymbol{\beta} = (0.2959, 0.1808, -0.0767)$.

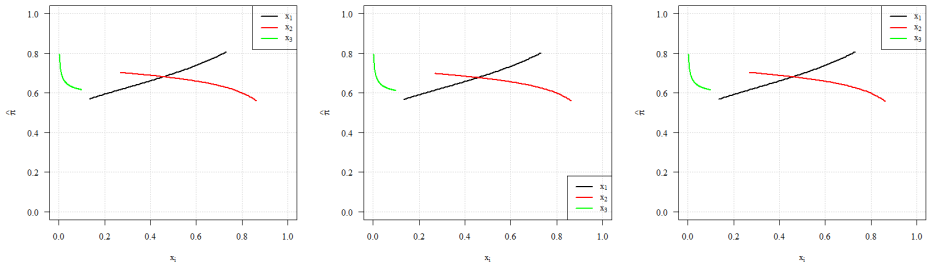
Tabela 4 Resultados de 1000 simulações de Monte Carlo de amostras da distribuição Simplex para as estimativas dos parâmetros dos modelos avaliados, considerando o vetor de parâmetros $\boldsymbol{\beta} = (0.2959, 0.1808, -0.0767)$ e $\sigma^2 = 0.7291$.

Modelo	**Parâmetro**	**Média**	**Desvio Padrão**	**Desvio**
M1	β_0	0,26530	0,19147	0,03060
	β_1	0,17655	0,03818	0,00135
	β_2	-0,07083	0,03377	0,00457
M2	β_0	0,27368	0,21474	0,02221
	β_1	0,17447	0,03705	0,00343
	β_2	-0,07060	0,03637	0,00474
M3	β_0	0,29813	0,22481	0,00223
	β_1	0,17227	0,04012	0,00563
	β_2	-0,06876	0,03766	0,00664

* Parâmetros relacionados ao modelo linear de Razão, com preditor linear dado por $\eta = \beta_0 + \beta_1\sqrt{\frac{x_1}{x_3}} + \beta_2\sqrt{\frac{x_2}{x_3}}$. M1: Regressão Simplex do tipo razão; M2: Boosted Simplex Regression do tipo razão e M3: Regressão Logística do tipo razão.

As correspondentes médias Monte Carlo foram utilizadas para construção dos gráficos de resposta traço e de razão de chances. Na Figura 4 (a), modelo M2, os componentes x_1 (gordura) e x_2 (carboidrato) tem efeito contrário sobre a resposta. Á medida que a proporção de gordura aumenta a proporção esperada de tumor aumenta. Por outro lado, á medida que a proporção de carboidrato aumenta a proporção esperada de tumor diminui. O componente x_3 (fibra) tem mais efeito sobre a resposta do que os outros componentes, uma vez que sucessivos incremen-

tos de fibra na dieta acarretam maior diminuição no número esperado de tumores. De maneira análoga, as mesmas conclusões podem ser feitas para o modelo M3.

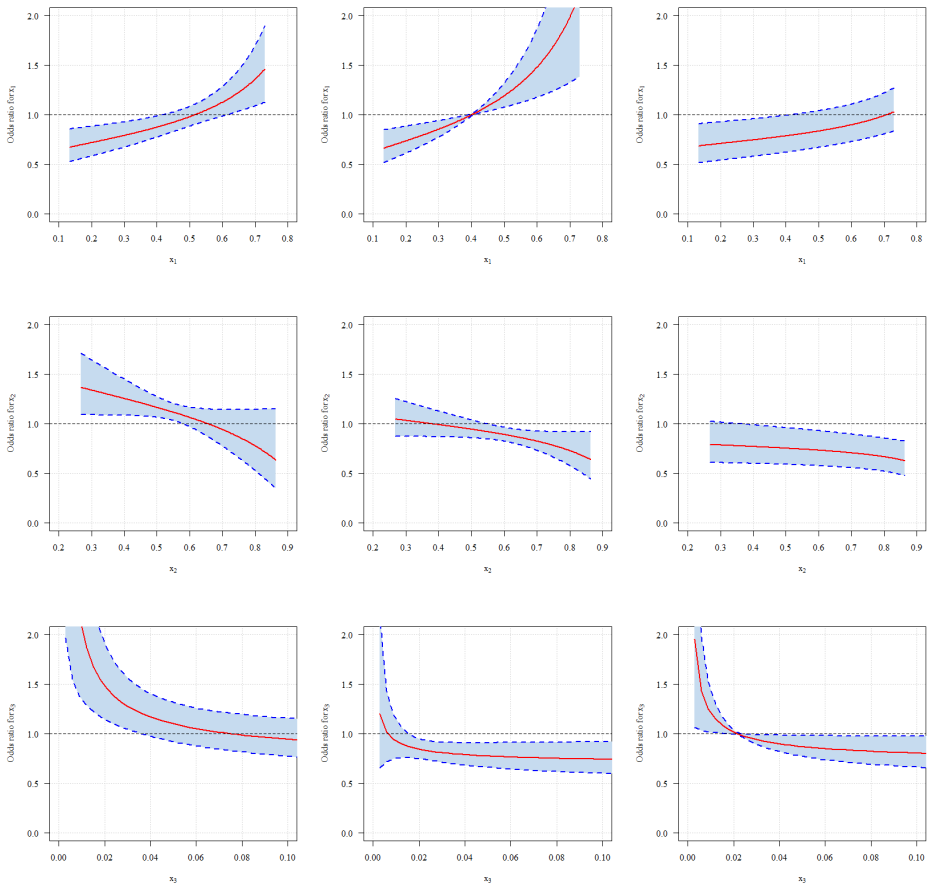


(a) M1 (b) M2 (c) M3

Figura 4 Gráficos de resposta traço dos modelos M2 (Regressão Simplex do tipo razão - (a)), M2 (Boosted Simplex Regression do tipo razão - (b)) e M3 (Regressão Logística do tipo razão - (c)) considerando o ponto $c = (0.4108, 0.5671, 0.0221)$, o centróide dos dados, como ponto de referência.

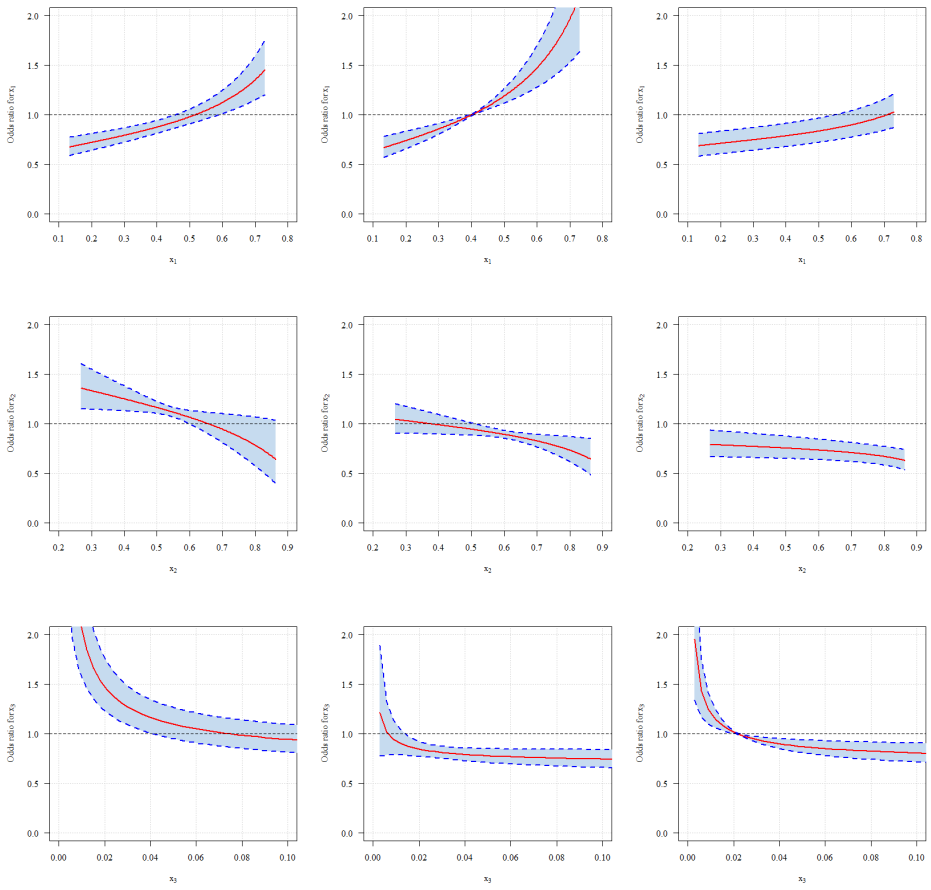
Nesse sentido, as Figuras 5 a 7 apresentam as razões de chances para diferentes pontos de referência em relação ao grupo controle para cada modelo avaliado. Para tal, foram considerados três diferentes pontos de referência $(0.7,\ 0.275,\ 0.025)$, $(0.275,\ 0.7,\ 0.025)$ e $(0.332,\ 0.466,\ 0.202)$. Os dois primeiros pontos de referência estão contidos na região onde os pontos amostrais estão. O terceiro ponto de referência é o centróide da região experimental restrita (Figura 2). Uma observação a ser feita é que deve-se ter cautela quanto às interpretações sobre o centróide da região restrita, uma vez que não foram amostrados dados nessa região. O grupo controle foi dado pelo centróide dos pontos amostrais $\mathbf{c} = (0.4108,\ 0.5671,\ 0.0221)$.

Pode ser observado que para o componente x_1, a chance de ocorrência de tumor mamário em ratos aumenta conforme ocorrem incrementos nesse componente. O respectivo intervalo de confiança de 95% do modelo M3 contém o



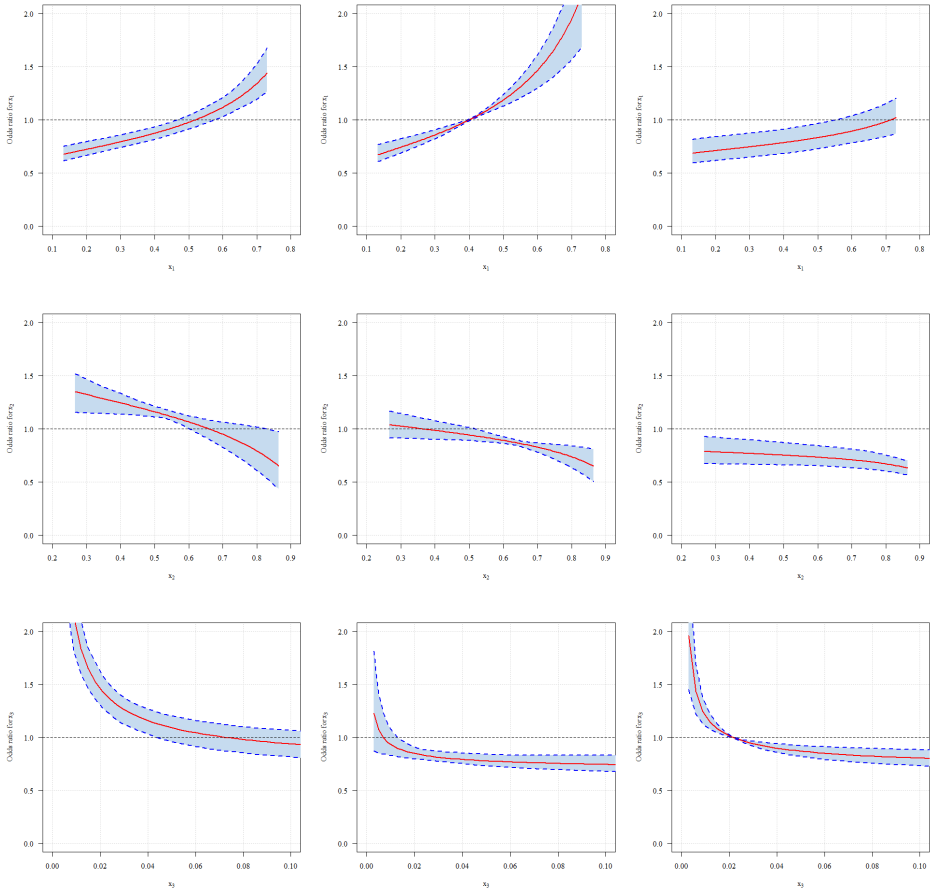
(A) (B) (C)

Figura 5 Razão de Chances (dada pela média Monte Carlo, curva vermelha) e seus respectivos Intervalos de Confiança de 95% de confiança para o modelo M3 (Regressão Logística do tipo razão) em que (A) o ponto de referência é $(0.7, 0.275, 0.025)$ e grupo controle $(0.4108, 0.5671, 0.0221)$, (B) o ponto de referência é $(0.275, 0.7, 0.025)$ e grupo controle $(0.4108, 0.5671, 0.0221)$ e (C) o ponto de referência é $(0.332, 0.466, 0.202)$ e grupo controle $(0.4108, 0.5671, 0.0221)$. Para visualização desses pontos veja Figura 2.



(A) (B) (C)

Figura 6 Razão de Chances (dada pela média Monte Carlo, curva vermelha) e seus respectivos Intervalos de Confiança de 95% de confiança para o modelo M1 (Regressão Simplex do tipo razão) em que (A) o ponto de referência é $(0.7, 0.275, 0.025)$ e grupo controle $(0.4108, 0.5671, 0.0221)$, (B) o ponto de referência é $(0.275, 0.7, 0.025)$ e grupo controle $(0.4108, 0.5671, 0.0221)$ e (C) o ponto de referência é $(0.332, 0.466, 0.202)$ e grupo controle $(0.4108, 0.5671, 0.0221)$. Para visualização desses pontos veja Figura 2.



(A) (B) (C)

Figura 7 Razão de Chances (dada pela média Monte Carlo, curva vermelha) e seus respectivos Intervalos de Confiança de 95% de confiança para o modelo M2 (Boosted Simplex Regression do tipo razão) em que (A) o ponto de referência é $(0.7, 0.275, 0.025)$ e grupo controle $(0.4108, 0.5671, 0.0221)$, (B) o ponto de referência é $(0.275, 0.7, 0.025)$ e grupo controle $(0.4108, 0.5671, 0.0221)$ e (C) o ponto de referência é $(0.332, 0.466, 0.202)$ e grupo controle $(0.4108, 0.5671, 0.0221)$. Para visualização desses pontos veja Figura 2.

valor 1, utilizado para comparar as razões de chances, nas quantidades de 0,4 a 0,6 aproximadamente. Logo pode-se dizer que embora a chance aumenta para as quantidades de 0,4 a 0,6, aproximadamente, de x_1, ela não é significativa, no sentido de que na população o componente x_1 (gordura) não influencia de maneira significativa na ocorrência de tumor mamário em ratos (Figura 5 (A)).

Em se tratando de estimativa pontual, as mesmas conclusões podem ser obtidas para os modelos M1 e M2 e considerando o componente x_1, no entanto, observa-se que o seu respectivo intervalo de confiança de 95% para a razão de chances não contém o valor unitário em algumas quantidades de x_1 (Figuras 6(A) e 7(A)), o que indica que esse componente influencia de maneira significativa na ocorrência de tumor mamário em ratos em quantidades diferentes daquelas explicadas pelo modelo M3. Esse fato é evidenciado pela amplitude do intervalo de confiança para a razão de chances desse componente, o qual é mais estreito nos modelos M1 e M2. Logo, os modelo M1 e M2 forneceram estimativas de razão de chances do componente x_1 mais precisas do que o modelo M3 (Figuras 6 e 7).

Esse fato pode ser estendido para os outros dois componentes e os outros pontos de referência utilizados (Figuras 7(B) e 7(C)). Portanto, o modelo *Boosted Simplex Regression* do tipo razão forneceu estimativas mais acuradas e precisas para a razão de chances.

Uma informação importante que pode ser fornecida por um modelo de mistura é a mistura que proporciona a máxima ou mínima incidência de tumor, respeitando-se as restrições de cada componente. Reportando o modelo que apresentou os melhores indicadores de qualidade de ajuste, o modelo M2, este afirma que a máxima incidência de tumor esperada é de $90,82\%$ e a mistura que proporciona esse valor é formulada por $72,38\%$ de gordura, $27,32\%$ de carboidrato e $0,30\%$ de fibra. Por outro lado, a mínima incidência de tumor esperada é de $56,19\%$ e a mistura que proporciona esse valor é formulada por $13,33\%$ de gordura, $86,10\%$ de carboidrato e $0,57\%$ de fibra (Tabela 5). Ou seja, as maiores diferença entre os componentes que maximizam e minimizam a resposta ficou caracterizado pela proporção de gordura e carboidrato na mistura. Além disso, pode-se dizer que o modelo M3 superestima a resposta máxima esperada e subestima a mínima esperada.

Tabela 5 Misturas dos componentes x_1 (gordura), x_2 (carboidrato) e x_3 (fibra) que proporcionam a máxima e mínima incidência de tumor esperada estimada ($\hat{y}$), considerando a média Monte Carlo das estimativas dos parâmetros dos modelos ajustados.

Modelo *	**Máximo**				**Mínimo**			
	x_1	x_2	x_3	$\hat{y}$	x_1	x_2	x_3	$\hat{y}$
M1	0,7186	0,2783	0,0030	0,9107	0,1331	0,8621	0,0047	0,5604
M2	0,7238	0,2732	0,0030	0,9082	0,1333	0,8610	0,0057	0,5619
M3	0,7294	0,2674	0,0031	0,9115	0,1332	0,8621	0,0047	0,5577

* M1 (Regressão Simplex do tipo razão), M2 (Boosted Simplex Regression do tipo razão) M3 (Regressão Logística do tipo razão).

4 CONCLUSÕES

O modelo de regressão simplex apresentou ajuste satisfatório na análise do experimento de mistura que avaliou a incidência de tumor mamário em ratos

fêmeas, sendo uma opção viável na análise de situações em que a resposta é limitada. O uso desse modelo também contempla a sub ou super-dispersão presente em dados agrupados.

Os gráficos de envelope simulado para o modelo *Boosted Simplex Regression* foram aplicados com sucesso para confirmar o ajuste desses modelos bem como corroborar a suposição de distribuição simplex no algoritmo boosting.

Os intervalos de confiança para a razão de chances evidenciaram que diferentes escolhas dos pontos de referência afetam a razão de chances e o respectivo intervalo de confiança. O modelo *Boosted Simplex Regression* forneceu estimativas acuradas para a razão de chances e intervalos de confiança, construídos por simulação Monte Carlo, mais precisos do que a análise pela regressão logística, independente do ponto de referência escolhido.

Os intervalos de confiança das razões de chances obtidas pelo modelo *Boosted Simplex Regression* foram ligeiramente mais precisos do que os obtidos pelo modelo de regressão simplex em sua abordagem por máxima verossimilhança.

REFERÊNCIAS

AKAY, K. U.; TEZ, M. Analyzing mixture experiments via generalized linear models, **International Journal of Pure** & **Applied Mathematics**, v. 36, n. 3, p. 373 - 390, 2007.

AKAY, K. U.; TEZ, M. Alternative modeling techniques for the quantal response data in mixture experiments, **Journal of Applied Statistics**, v. 38, n. 11, p. 2597 - 2616, 2011.

AITCHISON, J.; BACON-SHONE, J. Log contrast models for experiments with mixtures, **Biometrika**, v. 71, s. 2, p. 323 - 330, 1984.

BANDORFF-NIELSEN, O. E.; JORGENSEN, B. Some parametric models on

the simplex. **Journal of Multivariate Analysis**, v. 39, p. 106 - 116, 1991.

BOX, G. E. P; DRAPER, N. E. **Response surfaces, mixtures, and ridge analyses**, Wiley Series in Probability an Statistics, 2º ed., 874 p., 2007.

BOZDOGAN, H. Akaike's Information Criterion and recent developments in Information Complexity, **Journal of Mathematical Psychology**, v. 44, p. 62 - 91, 2000. doi:10.1006/jmps.1999.1277

BUHLMANN, P.; HOTHORN, T. Boosting Algorithms: Regularization, Prediction and Model Fitting, **Statistical Science**, v. 22, n. 4, p. 477-505, 2007.

CAO, D. S. et al. The Boosting: A new idea of building models, **Chemometrics and Intelligent Laboratory Systems**, v. 100, p. 1-11, 2010.

CHEN, J. J.; LI, L. A.; JACKSON, C. D. Analysis of quantal response data from mixture expreriments, **Environmetrics**, v. 7, n. 5, p. 503-512, 1996.

CORNELL, J. A. **Experiments with Mixtures: Designs, Models, and the Analysis of Mixture Data**, 3º ed., John Wiley Sonc Inc., USA, 2002.

CORNELL, J. A.; GORMAN, J. W. Two new mixture models: living with collinearity but removing its influence, **Journal Quality Technology**, v. 35, n. 1, p. 78-88, 2003.

CRIBARI-NETO, F.; ZEILEIS, A. Beta Regression in R, **Journal of Statistical Software**, v. 34, n. 2, p. 1-24, 2010.

CUMMINGS, P. Methods for estimating adjusted risk ratios, **The Stata Journal**, v. 9, n. 2, p. 175-196, 2009.

DRAPER, N. R.; ST JOHN, R. C. A mixture model with inverse terms. **Technometrics**, v. 19, n. 1, p. 37 - 46, 1977.

ESPINHEIRA, P. L.; FERRARI, S. L. P.; CRIBARI-NETO, F. On beta regression residuals. **Journal of Applied Statistics**, Routledge, v. 35, n. 4, p. 407 - 419, 2008.

FERRARI, S. L. P.; CRIBARI-NETO, F. Beta regression for modelling rates and proportions. **Journal of Applied Statistics**, Taylor and Francis Group, v. 31, n. 7, p. 799 - 815, 2004.

FREUND, Y.; SCHAPIRE, R. E. Experiments with a new Boosting algorithm, **In: International Conference on Machine Learning.**, p. 148-156, 1996.

FRIEDMAN, J. Greedy function aproximation: A gradient boosting machine, **The Annals of Statistics**, v. 29, p. 1189 - 1232, 2001.

FRIEDMAN, J. H.; HASTIE, T. J.; TIBSHIRANI, R. J. Additive logistic regression: A statistical view of Boosting (with discussion), **The Annals of Statistics**, v. 28, p. 337 - 407, 2000.

HOFNER, B.; MAYR, A.; ROBINZONOV, N.; SCHMID, M. Model-based boosting in R: a hands-on tutorial using the R package mboost, **Computation Statistics**, v. 29, p. 3 - 35, 2014. doi 10.1007/s00180-012-0382-5

HOSMER, D. W.; LEMESHOW, S.; STURDIVANT, R. X. **Applied Logistic Regression**, 3º ed., John Wiley, New York, 528p., 2013.

JORGENSEN, B. **The theory of dispersion models**, Chapman and Hall, London, 1997.

LISKA, G. R; MENEZES, F. S.; CIRILLO, M. A.; CORTEZ, R. M.; RIBEIRO, D. E. Evaluation os sensory panels of consumers of specialty coffee beverage using the boosting method in discriminant analysis, **Semina: Ciências Agrárias**, v. 36, n. 6, p. 3671-3680, 2015.

MENARD, S. **Logistic Regression: From Introductory to Advanced Concepts and Applications**, Sage Publications, Inc., USA, 2009.

NATEKIN, A.; KNOLL, A. Gradient boosting machines, a tutorial, **Frontier in Neurorobotics**, v. 7, p. 1-21, 2013.

R CORE TEAM (2015). **R: A language and environment for statistical computing**, R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/.

RIZZO, M. L. **Statistical Computing with R**. Chapman & Hall/CRC, 416 p., 2007.

SCHEFFÉ, H. Experiments with mixtures, **Journals of the Royal Statistical Society**, v. 20, p. 344 - 360, 1958.

SCHMID, M.; WICKLER, F.; MALONEY, K. O.; MITCHELL, R.; MAYR, A. Boosted Beta regression, **PLOS ONE**, v. 8, n. 4, p. 1-15, 2013.

SOUZA, T. C.; CRIBARI-NETO, F. Intelligence, religiosity and homosexuality non-acceptance: Empirical evidence, **Intelligence**, v. 52, p. 63-70, 2015.

SMITH, W. F. **Experimental Design for Formulation**, ASA-SIAM Series on Statistics and Applied Probability, SIAM, Philadelphia, ASA, Alexandria, VA, 2005.

SNEE, S. H. Techniques for the analysis of mixture data, **Technometrics**, v. 15, s. 3, p. 517 - 528, 1973.

WARTON, D. I.; HUI, F. K. C. The arcsine is asinine: The analysis of proportions in ecology,**Ecology**, v. 92, p. 3-10, 2011.

WILLIAMS, D. A. Generalized linear model diagnostic using the deviance and single case deletion, **Applied Statistics**, v. 36, p. 181-191, 1987.

CONSIDERAÇÕES FINAIS

O presente estudo apresentou novas abordagens na análise de experimentos de mistura e viabilizou novas perspectivas no que diz respeito à precisão de uma informação importante em experimentos do tipo quando a resposta analisada é caracterizada por proporções, a razão de chances. Em particular, a distribuição simplex e seu respectivo modelo de regressão foi proposta em comparação ao modelo de regressão logística, bastante usual em análises de respostas com a pressuposição de distribuição binomial. Diante dessa comparação feita no primeiro artigo, foi possível verificar que, com um mesmo conjunto de dados provenientes de um mesmo experimento, é possível aumentar a precisão da razão de chances.

A precisão da razão de chances é afetada pelo tamanho da amostra, mas em situações como os de experimentos de mistura, a exigência de grandes tamanhos amostrais é comprometida. O presente estudo mostrou que a precisão também é afetada pela distribuição de probabilidade assumida para a resposta e a escolha do componente sistemático de um MLG. Portanto, pesquisas que proporcionem aumento na precisão dos resultados desses experimentos são de grande importância e nesse sentido novos estudos com outras distribuições de probabilidade e componente sistemático podem ser feitos.

Os algoritmos boosting têm se mostrado como uma opção bastante promissora em termos de acurácia na predição e o estudo realizado no segundo artido evidenciou que o algoritmo boosting proposto, o *Boosted Simplex Regression*, foi uma opção viável na análise de experimentos de mistura. Por se tratar de uma abordagem relativamente recente, estudos sobre esses algoritmos têm recebido grande atenção pela comunidade estatística, ciência da computação e aprendizado de máquinas. Alguns autores como Hofner et al. (2014) e Schmid et al. (2013) atribuem

o termo "*blackbox*" a esses algoritmos. Uma das justificativas para essa denominação está no fato de o algoritmo GBF não fornecer a matriz hessiana das estimativas dos parâmetros do modelo ajustado. Para contornar esse problema, os intervalos de confiança para a razão de chances do algoritmo *Boosted Simplex Regression* foram construídos via simulação Monte Carlo.

Dessa forma, novos trabalhos podem ser propostos no sentido de resolver questões teóricas sobre o algoritmo boosting para modelos de regressão e novos algoritmos podem ser construídos para resolver problemas práticos.

Para que a pesquisa possa ser reproduzida e aplicada em outros problemas, foi desenvolvido o pacote estatístico em R intitulado por *SimplexMixmodel*. No apêndice se encontram os principais comandos para ajustar modelos de regressão simplex, diagnóstico de ajuste, construção dos gráficos de razões de chances e *traceplots*. Convém ressaltar que esse pacote se encontra em uma versão preliminar e novas versões serão disponibilizadas no CRAN do programa R.

APÊNDICE

APÊNDICE A - Ilustração Didática do Algoritmo Gradiente Boosting de Friedman

APÊNDICE B - Rotina do pacote SimplexMixmodel para ajuste do modelo de Regressão Simplex e Logística via Máxima Verossimilhança, critérios e gráficos de diagnóstico e gráfico de razão de chances

APÊNDICE A - Ilustração Didática do Algoritmo Gradiente Boosting de Friedman

Para exemplificar a ideia do algoritmo Gradiente Boosting de Friedman, considere o seguinte exemplo com variável resposta assumida como cotínua, três variáveis preditoras $\boldsymbol{x}_1$, $\boldsymbol{x}_2$, $\boldsymbol{x}_3$, três bases aprendizes lineares com coeficientes $\hat{\beta}_j^{(m)}$, $j = 1, 2, 3$. Considere o seguinte conjunto de dados:

Y_i	X_{1i}	X_{2i}	X_{3i}
8	2	1	4
10	-1	2	1
9	1	-3	4
6	2	1	2
12	1	4	6

1. Como primeiro passo do algoritmo, um valor inicial $\hat{f}^{(0)}(\cdot)$ considerando a função perda erro quadrático $\rho(y,\, g) = \frac{1}{2}(y - g)^2$ é a média da resposta Y. Essa derivação será feita depois desse exemplo. Logo

$$\hat{f}^{(0)} = \bar{Y} = 9$$

2. Aumentamos m em 1 e calculamos o vetor gradiente negativo referente a

perda $\rho\left(y,\, g\right) = \frac{1}{2}\left(y - g\right)^2$. Assim,

$$
\begin{aligned}
z_i = -\frac{\partial \rho\left(Y_i,\, f\right)}{\partial f} = Y_i - \hat{f}^{(0)} &\Rightarrow z_1 = 8 - 9 = -1 \\
&\Rightarrow z_2 = 10 - 9 = 1 \\
&\Rightarrow z_3 = 9 - 9 = 0 \\
&\Rightarrow z_4 = 6 - 9 = -3 \\
&\Rightarrow z_5 = 12 - 9 = 3
\end{aligned}
$$

3. No caso de ajuste de modelos lineares generalizados, o procedimento base adequado é o da equação 2.26 com parâmetros estimados por 2.27. Logo

$$
\begin{aligned}
\hat{\beta}_{(j=1)} &= \frac{2 \times (-1) + (-1) \times 1 + 1 \times 0 + 2 \times (-3) + 1 \times 3}{2^2 + (-1)^2 + 1^2 + 2^2 + 1^2} = -0,5454 \\
\hat{\beta}_{(j=2)} &= \frac{1 \times (-1) + 2 \times 1 + (-3) \times 0 + 1 \times (-3) + 4 \times 3}{1^2 + 2^2 + (-3)^2 + 1^2 + 4^2} = 0,3226 \\
\hat{\beta}_{(j=3)} &= \frac{4 \times (-1) + 1 \times 1 + 4 \times 0 + 2 \times (-3) + 6 \times 3}{4^2 + 1^2 + 4^2 + 2^2 + 6^2} = 0,1233
\end{aligned}
$$

queremos o $\hat{\beta}$ que retorna a menor soma de quadrados do resíduo, que é

dado resolvendo-se a expressão 2.28, daí

$$\begin{aligned}
\hat{\lambda}^{(j=1)} =& [(-1) - (-0,5454) \times 2]^2 + [1 - (-0,5454) \times (-1)]^2 \\
&+ [0 - (-0,5454) \times 1]^2 + [-3 - (-0,5454) \times 2]^2 \\
&+ [3 - (-0,5454) \times 1]^2 = 16,7273 \\
\hat{\lambda}^{(j=2)} =& [(-1) - 0,3226 \times 1]^2 + [1 - 0,3226 \times 2]^2 + [0 - 0,3226 \times (-3)]^2 \\
&+ [-3 - 0,3226 \times 1]^2 + [3 - 0,3226 \times 4]^2 = 16,7742 \\
\hat{\lambda}^{(j=3)} =& [(-1) - 0,1233 \times 4]^2 + [1 - 0,1233 \times 1]^2 + [0 - 0,1233 \times 4]^2 \\
&+ [-3 - 0,1233 \times 2]^2 + [3 - 0,1233 \times 6]^2 = 18,8904
\end{aligned}$$

Portanto, a variável escolhida nessa iteração é X_1, uma vez que produziu menor valor para $\hat{\lambda}$.

4. Então a atualização é dada por

$$\begin{aligned}
\hat{f}^{(1)}(x) =& \hat{f}^{(0)} + v \times \hat{g}^{(1)}(x) \\
=& \hat{f}^{(0)} + v \times \hat{\beta}^{(\hat{\lambda}_1)} x^{(\hat{\lambda}_1)} \\
=& 9 + 0,1 \times (-0,5454) \times X_1 \\
\hat{f}^{(1)}(x) =& 9 - 0,0545 \times X_1
\end{aligned}$$

Agora procedemos à segunda iteração. Retornemos ao passo dois do algoritmo.

1. o vetor gradiente negativo é dado por

$$z_i = Y_i - \underbrace{(9 - 0,0545 \times X_1)}_{\hat{f}^{(1)}} \Rightarrow z_1 = 8 - 9 + 0,0545 \times 2 = -0,8909$$

$$\Rightarrow z_2 = 10 - 9 + 0,0545 \times (-1) = 0,9454$$

$$\Rightarrow z_3 = 9 - 9 + 0,0545 \times 1 = 0,0545$$

$$\Rightarrow z_4 = 6 - 9 + 0,0545 \times 2 = -2,8909$$

$$\Rightarrow z_5 = 12 - 9 + 0,0545 \times 1 = 3,0545$$

2. Ajustando o vetor gradiente ao procedimento base, a fim de obter $\hat{g}^{(2)}(x)$, daí

$$\hat{\beta}_{(j=1)} = \frac{2 \times z_1 + (-1) \times z_2 + 1 \times z_3 + 2 \times z_4 + 1 \times z_5}{2^2 + (-1)^2 + 1^2 + 2^2 + 1^2} = -0,4909$$

$$\hat{\beta}_{(j=2)} = \frac{1 \times z_1 + 2 \times z_2 + (-3) \times z_3 + 1 \times z_4 + 4 \times z_5}{1^2 + 2^2 + (-3)^2 + 1^2 + 4^2} = 0,3279$$

$$\hat{\beta}_{(j=3)} = \frac{4 \times z_1 + 1 \times z_2 + 4 \times z_3 + 2 \times z_4 + 6 \times z_5}{4^2 + 1^2 + 4^2 + 2^2 + 6^2} = 0,1390$$

queremos o $\hat{\beta}$ que retorna menor $\hat{\lambda}$

$$\begin{aligned}
\hat{\lambda}^{(j=1)} =& [-0,8909 + 0,4909 \times 2]^2 + [0,9454 + 0,4909 \times (-1)]^2 \\
&+ [0,0545 + 0,4909 \times 1]^2 + [-2,8909 + 0,4909 \times 2]^2 \\
&+ [3,0545 + 0,4909 \times 1]^2 = 16,7273 \\
\hat{\lambda}^{(j=2)} =& [-0,8909 - 0,3279 \times 1]^2 + [0,9454 - 0,3279 \times 2]^2 \\
&+ [0,0545 - 0,3279 \times (-3)]^2 + [-2,8909 - 0,3279 \times 1]^2 \\
&+ [3,0545 - 0,3279 \times 4]^2 = 16,0460 \\
\hat{\lambda}^{(j=3)} =& [-0,8909 - 0,1390 \times 4]^2 + [0,9454 - 0,1390 \times 1]^2 \\
&+ [0,0545 - 0,1390 \times 4]^2 + [-2,8909 - 0,1390 \times 2]^2 \\
&+ [3,0545 - 0,1390 \times 6]^2 = 17,9682
\end{aligned}$$

3. A atualização de $\hat{f}^{(2)}(x)$ é dada por

$$\begin{aligned}
\hat{f}^{(2)}(x) =& \hat{f}^{(1)}(x) + v \times \hat{g}^{(2)}(x) \\
=& \hat{f}^{(1)}(x) + v \times \hat{\beta}^{(\hat{\lambda}_2)} x^{(\hat{\lambda}_2)} \\
=& \underbrace{9 - 0,0545 \times X_1}_{\hat{f}^{(1)}(x)} + 0,1 \times 0,3279 \times X_2 \\
\hat{f}^{(2)}(x) =& 9 - 0,0545 \times X_1 + 0,0328 \times X_2
\end{aligned}$$

A terceira iteração é feita de forma análoga ás iterações anteriores.

1. voltando para o passo 2 do algoritmo, calculamos o vetor gradiente negativo

$$
\begin{aligned}
z_1 &= -\,0,9237 \\
z_2 &= 0,8799 \\
z_3 &= 0,1529 \\
z_4 &= -\,2,9237 \\
z_5 &= 2,9234
\end{aligned}
$$

2. Obtendo $\hat{g}^{(3)}(x)$

$$
\begin{aligned}
\hat{\beta}^{(j=1)} &= -\,0,4998 \qquad \hat{\lambda}^{(j=1)} = 15,9967 \\
\hat{\beta}^{(j=2)} &= 0,2951 \qquad \hat{\lambda}^{(j=1)} = 15,9967 \\
\hat{\beta}^{(j=3)} &= 0,1299 \qquad \hat{\lambda}^{(j=3)} = 19,5007
\end{aligned}
$$

3. Logo, a atualização é dada por

$$
\begin{aligned}
\hat{f}^{(3)}(x) &= \hat{f}^{(2)}(x) + v \times \hat{g}^{(3)}(x) \\
&= \hat{f}^{(2)}(x) + v \times \hat{\beta}^{(\hat{\lambda}_3^1)} x^{(\hat{\lambda}_3^1)} \\
&= \underbrace{9 - 0,0545 \times X_1 + 0,0328 \times X_2}_{\hat{f}^{(2)}(x)} + 0,1 \times (-0,4998) \times X_1 \\
&= 9 - (0,0545 + 0,0500) \times X_1 + 0,0328 \times X_2 \\
\hat{f}^{(3)}(x) &= 9 - 0,1045 \times X_1 + 0,0328 \times X_2
\end{aligned}
$$

O algoritmo poderia continuar a ser executado por várias iterações e até mesmo por um número muito grande de iterações. Como dito anteriormente, executar o algoritmo de forma indefinida pode acarretar problemas no modelo, como por exemplo, forçar a escolha de uma variável não significativa ao modelo. Uma forma de determinar o número ideal de iterações é plotar em um gráfico o AIC resultante do modelo a cada iteração e quando o AIC atingir seu valor mínimo, este representará o número ótimo de iterações do algoritmo.

Como visto no exemplo, uma mesma variável pode ser escolhida não apenas em uma iteração, mas em várias iterações, aumentando sua contribuição individual no modelo final. Desse processo pode ocorrer também de alguma variável não estar no modelo final, caracterizando portanto, o sistema de seleção de variáveis, que está embutido no algoritmo.

APÊNDICE B - Rotina do pacote SimplexMixmodel para ajuste do modelo de Regressão Simplex e Logística via Máxima Verossimilhança, critérios e gráficos de diagnóstico e gráfico de razão de chances

```
rm(list=ls(all=TRUE))
##############################################################
##############################################################
#######                Regressão Simplex             #########
##############################################################
##############################################################
library(SimplexMixmodel)
Tumor
modelo<-tumor~-1+fat+carb+fiber+fat*carb
reg<-Simplexreg(modelo, Tumor)
reg
summary(reg)
?Simplexreg

##############################################################
#######          envelope simulado                ##########
##############################################################

modelo<-tumor~-1+fat+carb+fiber+fat*carb
simplex.envel(modelo, Tumor, reg)

##############################################################
#####  funcao para criar um ponto experimental    #######
#####          considerando restricoes            #######
##############################################################

X<-Gmix(L1=0.133,U1=0.730,L2=0.267,U2=0.864,
      L3=0.003,U3=0.600,c=3,nobs=100000)

X
apply(X, 1,sum) # see that mixture constraint is satisfyed

##############################################################
######  funcao para calcular os valores preditos  ######
######              na regiao simplex              ######
##############################################################
colnames(X)<-c("fat","carb","fiber")
modelo<-~-1+fat+carb+fiber+fat*carb

preditos <- Predict(formula=modelo, reg, newdata= X)
preditos
```

```
summary(preditos)

##########################################################
## encontrando o maximo da resposta dada as restricoes #
##           de mistura e o respectiva mistura          ##
##########################################################
des<-data.frame(X, y=preditos)
head(des)
dim(des)
opt.mix(des)


####  regiao simplex com os valores preditos  ##########
####  para ver os valores simulados na regiao restrita##
# OBS.: atencao para colocar os rotulos dos eixos!

x1<-X$fat
x2<-X$carb
x3<-X$fiber
y<-preditos
des<-data.frame(x1,x2,x3,y)

MixturePlot(des=des,x1lab="Fat",x2lab="Carbohidrate",
            x3lab="Fiber",
            lims=c(0.133,0.730,0.267,0.864,0.003,0.600),
            constrts=FALSE,
            corner.labs=c("x3","x2","x1"),mod=2,
            n.breaks=1,cols=FALSE, pseudo=FALSE)

#######  regiao simplex com os valores preditos  #########
mymodSI=function(grid){
  x1=grid$x1 #fat
  x2=grid$x2 #carb
  x3=grid$x3 #fiber
  y=exp(-0.8241*x1-0.8896*x2-7.9549*x3+9.5299*x1*x2)/
    (1+exp(-0.8241*x1-0.8896*x2-7.9549*x3+9.5299*x1*x2))
  return(y)}
# Create vectors of design points for each variable
x1<-X$fat
x2<-X$carb
x3<-X$fiber

obs=as.data.frame(cbind(x1,x2,x3))
y<-mymodSI(obs)

## pode-se definir um modelo de mistura qualquer e plotar
## o grafico simplex. para isso usa-se a funcao ModelPlot,
```

```
## que tem como requisito o arqumento "model" deve receber
## um objeto que tenha a funcao predict.

ModelPlot(obs, user.func=mymodSI, dimensions =
            list(x1="x1",x2="x2",x3="x3"),
          constraints=TRUE,contour=TRUE,
          lims=c(0.133,0.730,0.267,0.864,0.003,0.600),
          cornerlabs=c("x1","x2","x3"),
          fill=TRUE,color.palette = heat.colors,
          axislabs=c("Fat", "Carbohidrate", "Fiber"),
          pseudo=FALSE)


############################################################
############     indice de Bozdogan's     ##################
############################################################

library(fBasics)

ICOMP(reg)

############################################################
##### traceplots function for all variables  ##############
############################################################

cent<-c(0.4108,0.5671,0.0221)
modelo<-tumor~-1+fat+carb+fiber+fat*carb
traceplot(modelo,reg,cent,Tumor,n=50,
  L1=0.133,U1=0.730,L2=0.267,U2=0.864,L3=0.003,U3=0.600)

############################################################
####### ORplots function for each variable  ###############
############################################################

##### ORplots functions for x1 variable  ##############
cent<-c(0.4108,0.5671,0.0221)
pref<-c(0.7,0.275,0.025)
modelo<-tumor~-1+fat+carb+fiber+fat*carb
ORplotx1(modelo,reg,cent,pref,Tumor,n=50,
         L1=0.133,U1=0.730, alpha=0.1)
#?ORplotx1
#ponto 2
cent<-c(0.4108,0.5671,0.0221)
pref<-c(0.275,0.7,0.025)
modelo<-tumor~-1+fat+carb+fiber+fat*carb
ORplotx1(modelo,reg,cent,pref,Tumor,n=20,
         L1=0.133,U1=0.730, alpha=0.05)
```

```
#ponto 3
cent<-c(0.4108,0.5671,0.0221)
pref<-c(0.332,0.466,0.202)
modelo<-tumor~-1+fat+carb+fiber+fat*carb
ORplotx1(modelo,reg,cent,pref,Tumor,n=20,
         L1=0.133,U1=0.730, alpha=0.05)

##### ORplots functions for x2 variable  ##############
cent<-c(0.4108,0.5671,0.0221)
pref<-c(0.7,0.275,0.025)
modelo<-tumor~-1+fat+carb+fiber+fat*carb
ORplotx2(modelo,reg,cent,pref,Tumor,n=50,
           L2=0.267, U2=0.864, alpha=0.05)

#ponto 2
cent<-c(0.4108,0.5671,0.0221)
pref<-c(0.275,0.7,0.025)
modelo<-tumor~-1+fat+carb+fiber+fat*carb
ORplotx2(modelo,reg,cent,pref,Tumor,n=50,
         L2=0.267, U2=0.864, alpha=0.05)

#ponto 3
cent<-c(0.4108,0.5671,0.0221)
pref<-c(0.332,0.466,0.202)
modelo<-tumor~-1+fat+carb+fiber+fat*carb
ORplotx2(modelo,reg,cent,pref,Tumor,n=50,
         L2=0.267, U2=0.864, alpha=0.05)


##### ORplots functions for x3 variable  ##############
cent<-c(0.4108,0.5671,0.0221)
pref<-c(0.7,0.275,0.025)
modelo<-tumor~-1+fat+carb+fiber+fat*carb
ORplotx3(modelo,reg,cent,pref,Tumor,n=50,
         L3=0.003, U3=0.600, alpha=0.05)

#ponto 2
cent<-c(0.4108,0.5671,0.0221)
pref<-c(0.275,0.7,0.025)
modelo<-tumor~-1+fat+carb+fiber+fat*carb
ORplotx3(modelo,reg,cent,pref,Tumor,n=50,
         L3=0.003, U3=0.600, alpha=0.05)

#ponto3
pref<-c(0.332, 0.466, 0.202)
control<-c(0.4108, 0.5671, 0.0221)
ORplotx3(modelo,reg,cent,pref,Tumor,n=50,
         L3=0.003, U3=0.600, alpha=0.05)
```

```
############################################################
##########  modelo com termos inversos do    #############
##########          tipo raiz quadrada       #############
############################################################
library(SimplexMixmodel)
Tumor
modelo<-tumor~I(sqrt(fat/fiber))+I(sqrt(carb/fiber))
reg<-Simplexreg.iv(modelo, Tumor)
reg
summary(reg)


##########################################################
#######          envelope simulado              ##########
##########################################################

modelo<-tumor~I(sqrt(fat/fiber))+I(sqrt(carb/fiber))
simplex.envel.iv(modelo, Tumor, reg)

##########################################################
#####  funcao para criar um ponto experimental   #######
#####          considerando restricoes           #######
##########################################################

X<-Gmix(L1=0.133,U1=0.730,L2=0.267,U2=0.864,
        L3=0.003,U3=0.600,c=3,nobs=100000)

X
apply(X, 1,sum) # see that mixture constraint is satisfyed

##########################################################
######  funcao para calcular os valores preditos  ######
######              na regiao simplex             ######
##########################################################
colnames(X)<-c("fat","carb","fiber")
modelo<-~-1+fat+carb+fiber+fat*carb

preditos <- Predict(formula=modelo, reg, newdata= X)
preditos

summary(preditos)

############################################################
## encontrando o maximo da resposta dada as restricoes #
##          de mistura e o respectiva mistura         ##
############################################################
des<-data.frame(X, y=preditos)
```

```
head(des)
dim(des)
opt.mix(des)


####  regiao simplex com os valores preditos  ##########
####  para ver os valores simulados na regiao restrita##
# OBS.: atencao para colocar os rotulos dos eixos!

x1<-X$fat
x2<-X$carb
x3<-X$fiber
y<-preditos
des<-data.frame(x1,x2,x3,y)

MixturePlot(des=des,x1lab="Fat",x2lab="Carbohidrate",
            x3lab="Fiber",
            lims=c(0.133,0.730,0.267,0.864,0.003,0.600),
            constrts=FALSE,
            corner.labs=c("x3","x2","x1"),mod=2,
            n.breaks=1,cols=FALSE, pseudo=FALSE)

#######  regiao simplex com os valores preditos  #########
mymodSI=function(grid){
  x1=grid$x1 #fat
  x2=grid$x2 #carb
  x3=grid$x3 #fiber
  y=exp(-0.8241*x1-0.8896*x2-7.9549*x3+9.5299*x1*x2)/
    (1+exp(-0.8241*x1-0.8896*x2-7.9549*x3+9.5299*x1*x2))
  return(y)}
# Create vectors of design points for each variable
x1<-X$fat
x2<-X$carb
x3<-X$fiber

obs=as.data.frame(cbind(x1,x2,x3))
y<-mymodSI(obs)

## pode-se definir um modelo de mistura qualquer e plotar
## o grafico simplex. para isso usa-se a funcao ModelPlot,
## que tem como requisito o arqumento "model" deve receber
## um objeto que tenha a funcao predict.

ModelPlot(obs, user.func=mymodSI, dimensions =
            list(x1="x1",x2="x2",x3="x3"),
          constraints=TRUE,contour=TRUE,
          lims=c(0.133,0.730,0.267,0.864,0.003,0.600),
          cornerlabs=c("x1","x2","x3"),
```

```
        fill=TRUE, color.palette = heat.colors,
        axislabs=c("Fat", "Carbohidrate", "Fiber"),
        pseudo=FALSE)


############################################################
############     indice de Bozdogan's     ##################
############################################################

library(fBasics)

ICOMP(reg)

############################################################
#####  traceplots function for all variables  #############
############################################################

cent<-c(0.4108,0.5671,0.0221)
modelo<-tumor~I(sqrt(fat/fiber))+I(sqrt(carb/fiber))
traceplot.iv(modelo,reg,cent,Tumor,n=50,
        L1=0.133,U1=0.730,L2=0.267,U2=0.864,L3=0.003,U3=0.600)

############################################################
#######  ORplots function for each variable  ############
############################################################

#####  ORplots functions for x1 variable  #############
cent<-c(0.4108,0.5671,0.0221)
pref<-c(0.7,0.275,0.025)
modelo<-tumor~I(sqrt(fat/fiber))+I(sqrt(carb/fiber))
ORplotx1.iv(modelo,reg,cent,pref,Tumor,n=50,
        L1=0.133,U1=0.730, alpha=0.05)
#?ORplotx1
#ponto 2
cent<-c(0.4108,0.5671,0.0221)
pref<-c(0.275,0.7,0.025)
modelo<-tumor~I(sqrt(fat/fiber))+I(sqrt(carb/fiber))
ORplotx1.iv(modelo,reg,cent,pref,Tumor,n=20,
        L1=0.133,U1=0.730, alpha=0.05)

#ponto 3
cent<-c(0.4108,0.5671,0.0221)
pref<-c(0.332,0.466,0.202)
modelo<-tumor~I(sqrt(fat/fiber))+I(sqrt(carb/fiber))
ORplotx1.iv(modelo,reg,cent,pref,Tumor,n=20,
        L1=0.133,U1=0.730, alpha=0.05)

#####  ORplots functions for x2 variable  #############
```

```
cent<-c(0.4108,0.5671,0.0221)
pref<-c(0.7,0.275,0.025)
modelo<-tumor~-1+fat+carb+fiber+fat*carb
ORplotx2.iv(modelo,reg,cent,pref,Tumor,n=50,
          L2=0.267, U2=0.864, alpha=0.05)

#ponto 2
cent<-c(0.4108,0.5671,0.0221)
pref<-c(0.275,0.7,0.025)
modelo<-tumor~I(sqrt(fat/fiber))+I(sqrt(carb/fiber))
ORplotx2.iv(modelo,reg,cent,pref,Tumor,n=50,
          L2=0.267, U2=0.864, alpha=0.05)

#ponto 3
cent<-c(0.4108,0.5671,0.0221)
pref<-c(0.332,0.466,0.202)
modelo<-tumor~I(sqrt(fat/fiber))+I(sqrt(carb/fiber))
ORplotx2.iv(modelo,reg,cent,pref,Tumor,n=50,
          L2=0.267, U2=0.864, alpha=0.05)


##### ORplots functions for x3 variable  #############
cent<-c(0.4108,0.5671,0.0221)
pref<-c(0.7,0.275,0.025)
modelo<-tumor~I(sqrt(fat/fiber))+I(sqrt(carb/fiber))
ORplotx3.iv(modelo,reg,cent,pref,Tumor,n=50,
          L3=0.003, U3=0.600, alpha=0.05)

#ponto 2
cent<-c(0.4108,0.5671,0.0221)
pref<-c(0.275,0.7,0.025)
modelo<-tumor~I(sqrt(fat/fiber))+I(sqrt(carb/fiber))
ORplotx3.iv(modelo,reg,cent,pref,Tumor,n=50,
          L3=0.003, U3=0.600, alpha=0.05)

#ponto3
pref<-c(0.332, 0.466, 0.202)
control<-c(0.4108, 0.5671, 0.0221)
ORplotx3.iv(modelo,reg,cent,pref,Tumor,n=50,
          L3=0.003, U3=0.600, alpha=0.05)


#################################################################
#######         Regressao Logistica          ###################
#################################################################
setwd("E:/Documents/Ufla/Tese")
dados.glm=read.table("dados_artigo_mistura_mod3.txt", h=T)
str(dados.glm)
```

```
attach(dados.glm)
stumor<-30-tumor
Resp<-cbind(tumor, stumor)
Resp

reg <- glm(Resp ~ -1 + fat + carb + fiber + fat*carb,
           family=binomial, data=dados.glm)
summary(reg)

############################################################
############     indice de Bozdogan's     ##################
############################################################

library(fBasics)

ICOMP(reg)
AIC(reg)
BIC(reg)
############################################################
##### traceplots function for all variables  ##############
############################################################

cent<-c(0.4108,0.5671,0.0221)
modelo<-tumor~-1+fat+carb+fiber+fat*carb
traceplot(modelo,reg,cent,Tumor,n=50,
    L1=0.133,U1=0.730,L2=0.267,U2=0.864,L3=0.003,U3=0.600)

############################################################
####### ORplots function for each variable  ############
############################################################

##### ORplots functions for x1 variable  ##############
cent<-c(0.4108,0.5671,0.0221)
pref<-c(0.7,0.275,0.025)
modelo<-tumor~-1+fat+carb+fiber+fat*carb
ORplotx1(modelo,reg,cent,pref,Tumor,n=50,
         L1=0.133,U1=0.730, alpha=0.05)
#?ORplotx1
#ponto 2
cent<-c(0.4108,0.5671,0.0221)
pref<-c(0.275,0.7,0.025)
modelo<-tumor~-1+fat+carb+fiber+fat*carb
ORplotx1(modelo,reg,cent,pref,Tumor,n=20,
         L1=0.133,U1=0.730, alpha=0.05)

#ponto 3
cent<-c(0.4108,0.5671,0.0221)
pref<-c(0.332,0.466,0.202)
```

```
modelo<-tumor~-1+fat+carb+fiber+fat*carb
ORplotx1(modelo,reg,cent,pref,Tumor,n=20,
         L1=0.133,U1=0.730, alpha=0.05)

##### ORplots functions for x2 variable  #############
cent<-c(0.4108,0.5671,0.0221)
pref<-c(0.7,0.275,0.025)
modelo<-tumor~-1+fat+carb+fiber+fat*carb
ORplotx2(modelo,reg,cent,pref,Tumor,n=50,
         L2=0.267, U2=0.864, alpha=0.05)

#ponto 2
cent<-c(0.4108,0.5671,0.0221)
pref<-c(0.275,0.7,0.025)
modelo<-tumor~-1+fat+carb+fiber+fat*carb
ORplotx2(modelo,reg,cent,pref,Tumor,n=50,
         L2=0.267, U2=0.864, alpha=0.05)

#ponto 3
cent<-c(0.4108,0.5671,0.0221)
pref<-c(0.332,0.466,0.202)
modelo<-tumor~-1+fat+carb+fiber+fat*carb
ORplotx2(modelo,reg,cent,pref,Tumor,n=50,
         L2=0.267, U2=0.864, alpha=0.05)


##### ORplots functions for x3 variable  #############
cent<-c(0.4108,0.5671,0.0221)
pref<-c(0.7,0.275,0.025)
modelo<-tumor~-1+fat+carb+fiber+fat*carb
ORplotx3(modelo,reg,cent,pref,Tumor,n=50,
         L3=0.003, U3=0.600, alpha=0.05)

#ponto 2
cent<-c(0.4108,0.5671,0.0221)
pref<-c(0.275,0.7,0.025)
modelo<-tumor~-1+fat+carb+fiber+fat*carb
ORplotx3(modelo,reg,cent,pref,Tumor,n=50,
         L3=0.003, U3=0.600, alpha=0.05)

#ponto3
pref<-c(0.332, 0.466, 0.202)
control<-c(0.4108, 0.5671, 0.0221)
ORplotx3(modelo,reg,cent,pref,Tumor,n=50,
         L3=0.003, U3=0.600, alpha=0.05)


## Mesmos comandos usados para regressao simplex
```

```
## com termos inversos nos graficos dos
## intervalos de confianca para razao
## de chances podem ser usados em regressao
## logistica.
## O modelo Boosted Simplex Regression sera
## disponibilizado em breve no pacote SimplexMixmodel
```