Modelos Box-Cox simétricos e aplicações a dados nutricionais

Giovana Fumes

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

Programa: Estatística
Orientadora: Profª. Drª. Silvia Lopes de Paula Ferrari

Durante o desenvolvimento deste trabalho a autora recebeu auxílio financeiro da CAPES e do CNPq.

São Paulo, 12 de dezembro de 2014.

Agradecimentos

A Deus, a razão do meu viver e a motivação maior que me levou a realizar este trabalho.

À minha orientadora, Profª Drª Silvia Lopes de Paula Ferrari, por toda paciência, entusiasmo e sabedoria na orientação da pesquisa realizada, pela sua amizade e compreensão.

Aos meus pais, Israel Fumes e Maria Alvacir Gonçalves Fumes, pelo amor e apoio incondicionais.

Às minhas irmãs, Jaqueline Fumes e Juliane Fumes Bazzo, ao meu cunhado, Reginaldo Luiz Bazzo e aos meus sobrinhos Cauê Henrique Fumes Tamburini, Felipe Gabriel Fumes Tamburini, Beatriz Fumes Bazzo e Helena Fumes Bazzo, muito obrigada por compreenderem minha ausência e me motivarem a realizar este desafio.

Ao meu tio, Laudenir Antônio Gonçalves Filho, por sempre me apoiar e incentivar para vida acadêmica.

Ao meu namorado, Youssif Ghantous Filho, pelo carinho e suporte na conclusão deste trabalho.

Ao Prof. Dr. José Eduardo Corrente, por ceder o banco de dados para esta pesquisa e por sua amizade sempre presente.

Aos professores do Instituto de Matemática e Estatística, que contribuíram para minha formação. Aos meus amigos e amigas de doutorado, muito obrigada pelas horas de estudo e partilha.

Aos meus companheiros da E. E. Profº Américo Virgínio dos Santos e amigos das comunidades São João Batista e Menino Deus e Santo Antônio, meus sinceros agradecimentos por todo carinho e preces.

"Tudo posso Naquele que me fortalece."(Fil 4,13)

Dedico este trabalho a minha avó, Helena Paniguel Gonçalves (*in memoriam*), por seu exemplo de sabedoria e força.

Resumo

Na área de nutrição é de grande interesse dos pesquisadores estimar a distribuição usual de consumo e a prevalência de inadequação alimentar de um grupo populacional. Usualmente, transformações de Box-Cox aliadas à distribuição normal são utilizadas na modelagem de dados de consumo alimentar (Dodd et al., 2006; Tooze et al., 2010). No entanto, quando os dados apresentam distribuição altamente assimétrica ou presença de pontos discrepantes, esta abordagem pode conduzir a estimativas não plausíveis. Além disso, tal modelagem não permite a interpretação dos parâmetros em termos de características dos dados originais e requer uma transformação inversa dos dados transformados para a escala original.

Esta tese propõe um enfoque alternativo para estimação da distribuição usual de consumo e da prevalência de inadequação alimentar através do modelo Box-Cox *t* (Rigby & Stasinopoulos, 2006b) com efeitos aleatórios ou mistos. A modelagem proposta é flexível, permitindo modelar dados assimétricos mesmo na presença de observações discrepantes. Ao contrário das abordagem usual, o modelo proposto não requer transformação nos dados e permite a interpretação de relações entre covariáveis e quantis da distribuição de consumo. Em aplicações a um conjunto de dados de consumo de 22 micronutrientes, na maioria dos casos o modelo Box-Cox *t* proporcionou melhor ajuste do que seus competidores. Tem-se ainda, por meio de um estudo de simulação, que o modelo Box-Cox *t* estima a distribuição de consumo usual populacional satisfatoriamente, e este é o preferível em casos de assimetria positiva acentuada, especialmente na presença de cauda direita pesada.

Adicionalmente, esta tese propõe uma nova classe de distribuições para variáveis aleatórias assimétricas positivas, a classe de distribuições Box-Cox simétricas. Um estudo detalhado de propriedades desta nova classe de distribuições é apresentado. Esta classe inclui as distribuições log-simétricas, como por exemplo, a log-normal, as

distribuições simétricas truncadas com suporte nos reais positivos e as distribuições Box-Cox *t* (Rigby & Stasinopoulos, 2006b), Box-Cox Cole-Green (Cole & Green, 1992) e Box-Cox exponencial potência (Rigby & Stasinopoulos, 2004; Voudouris et al., 2012). A nova classe de distribuições permite a interpretação dos parâmetros em termos de quantis (em particular, a mediana), dispersão relativa e assimetria, o que a torna atrativa para modelagem de regressão. Particularmente, a classe de distribuições Box-Cox simétricas é útil para modelagem de dados positivos assimétricos na presença de observações discrepantes, pois inclui distribuições de cauda pesada. Aplicações de modelos Box-Cox simétricos e alternativos para a análise de dados de consumo de 22 micro e 11 macronutrientes são apresentados. Os modelos Box-Cox simétricos forneceram melhor ajuste do que os alternativos.

Palavra chave: consumo alimentar; distribuição Box-Cox *t*; distribuições Box-Cox simétricas; nutrição; transformação de Box-Cox.

Abstract

The issue of estimating usual nutrients intake distributions and the prevalence of food inadequacy is of interest in nutrition studies. Box-Cox transformations coupled with the normal distribution are usually employed for modeling nutrients intake data (Dodd et al., 2006; Tooze et al., 2010). However, when the data present highly asymmetric distribution or include outliers, this approach may lead to implausible estimates. Additionally, it does not allow interpretation of the parameters in terms of characteristics of the original data and requires backtransformation of the transformed data to the original scale.

This thesis proposes an alternative approach for estimating usual nutrients intake and the prevalence of food inadequacy through a Box-Cox t model (Rigby & Stasinopoulos, 2006b) with random or mixed effects. The proposed model is flexible enough for modeling highly asymmetric data even when outliers are present. Unlike the usual approach, the proposed model does not require a transformation of the data and facilitates interpretation of covariates effects. In applications to data sets on intake of 22 micronutrients, the Box-Cox t models provided better fit than its competitors in most of the cases. A simulation study suggests that the Box-Cox t model estimates the usual intake distribution satisfactorily, and that it should be preferable to the usual approach particularly in cases of highly asymmetric heavy tailed data.

Additionally, this thesis proposes a new class of distributions for positive asymmetric random variables, the Box-Cox symmetric class of distributions. A detailed study of properties of the new class of distributions is provided. It includes as special cases not only the log-symmetric distributions, for instance the log-normal distribution, and the truncated symmetric distributions with support on the positive real line, but also the Box-Cox t (Rigby & Stasinopoulos, 2006b), Box-Cox Cole-Green (Cole & Green, 1992) and Box-Cox power exponential (Rigby & Stasinopoulos, 2004; Voudouris et al., 2012)

distributions. The new class of distributions allows interpretation of the parameters in terms of quantiles (in particular, the median), relative dispersion and skewness of the distribution, which makes it attractive for regression modeling. The Box-Cox symmetric class of distributions is particularly useful for modeling positively asymmetric data in the presence of outliers since it includes heavy tailed distributions. Applications of Box-Cox symmetric distributions and alternative models to the analysis of data on the intake of 22 micronutrients and 11 macronutrients are presented. It is noted that the Box-Cox symmetric models provided better fit than the alternative models.
Keywords. Box-Cox symmetric distributions; Box-Cox *t* distribution; Box-Cox transformation; nutrients intake; nutrition.

Índice

CAPÍTULO 1

Introdução

1.1 Motivação

Um dos principais interesses dos pesquisadores da área de nutrição é estimar a distribuição usual de consumo e a inadequação do consumo alimentar de nutrientes (macro ou micronutrientes) de um grupo populacional.

A modelagem estatística para medir o consumo de cada nutriente e determinar a proporção de inadequação alimentar de uma população apresenta um desafio, tanto para nutricionistas quanto para estatísticos, uma vez que indivíduos diferentes tendem a ter hábitos alimentares distintos entre si (variabilidade interpessoal), e o próprio indivíduo não necessariamente apresenta um consumo constante (variabilidade intrapessoal) (Borrelli et al., 1992).

Os métodos estatísticos mais utilizados na análise de dados nutricionais apresentam uma estrutura comum baseada em um instrumento chamado recordatório 24 horas (R24h) para estimação do consumo habitual de um grupo. Através do R24h, o indivíduo relata o seu consumo de alimentos nas últimas 24 horas anteriores ao momento da entrevista (Block, 1982; Slater et al., 2004).

De posse das informações provenientes dos R24h, existem programas específicos que transformam os dados coletados em consumo de micro e macronutrientes. Um desses programas é o NDSR (*Nutrition Data System for Research*) desenvolvido pela Universidade de Minnesota, o qual foi utilizado para gerar os dados tratados nesta tese, e aqui contempla a avaliação com ênfase no consumo de micronutrientes, que são aqueles nutrientes necessários em quantidades menores para o bom funcionamento do organismo, como as vitaminas e os minerais.

Na literatura, é comum se utilizar uma transformação Box & Cox (1964) e analisar os dados de consumo de nutrientes transformados sob a normalidade (Dodd et al., 2006). Assumindo esta suposição, são comuns três alicerces para diversos métodos existentes. O primeiro afirma que o R24h é um instrumento não viesado para estimar o verdadeiro consumo habitual de um indivíduo, isto é, a média de alguns R24h do indivíduo pode representar de maneira não tendenciosa o seu verdadeiro consumo alimentar. O segundo particiona a variância total dos R24h em dois componentes de variância: a variabilidade intra e interpessoal. Para tal, uma análise de variância de um fator é realizada. Por fim, considera-se que a distribuição do consumo de nutrientes é feita de acordo com a "correção" pela variabilidade intrapessoal. O coeficiente de correlação intraclasse de um modelo misto sem covariáveis é usado como um fator atenuante, que pondera a variabilidade dos dados, de modo que, se a variabilidade individual for muito alta, o fator coloca os chamados valores intermediários mais próximos da média do grupo, e se a variabilidade intrapessoal é moderada, este fator considera essa variabilidade individual para o estimador. Métodos como ISU (*Iowa State University*) e BP (*Best Power*) (Dodd et al., 2006) seguem esse pano de fundo.

Todos os métodos supracitados consideram a normalidade nos dados, antes ou após uma transformação, e de modo geral se diferenciam pela transformação inversa a ser aplicada para o retorno aos dados originais, bem como na forma de considerar a ponderação do fator atenuante. É importante ressaltar que tais modelos não consideram a inclusão de covariáveis e a possibilidade de não consumo de um determinado nutriente.

Um modelo também baseado na distribuição normal, que considera a inclusão de covariáveis e leva em conta a proporção de não consumo, é proposto por Tooze et al. (2010) e conhecido como método NCI (*National Cancer Institute*). Este modelo, também exibe os percentis da distribuição usual de consumo, os quais podem ser muito úteis para as intervenções a serem feitas por pesquisadores da área. Todavia, em conjuntos de dados com alta assimetria e/ou presença de *outliers* extremos, a modelagem baseada no método NCI parece não ser a mais adequada, uma vez que as estimativas de prevalência de inadequação alimentar tendem a ser incompatíveis com os dados observados.

Na análise estatística de dados contínuos, a distribuição normal é uma das mais utilizadas devido às suas propriedades, em especial no contexto de modelos lineares. No entanto, a presença de pontos discrepantes (*outliers*) afeta de forma relevante a inferência baseada no modelo normal, incentivando o desenvolvimento de procedimentos robustos, os quais são definidos como aqueles que são menos sensíveis a desvios das suposições sobre as quais se baseiam (Hampel et al., 1986).

Esta tese visa apresentar um procedimento robusto, baseado no modelo Box-Cox t (Rigby & Stasinopoulos, 2006b), no contexto de modelos com efeitos aleatórios e mistos, para a estimação da distribuição usual de consumo e da prevalência de inadequação

alimentar. Além disso, é apresentada uma proposta de generalização para uma classe de distribuições fundamentada na transformação Box-Cox, envolvendo distribuições simétricas truncadas.

1.2 Modelos Aditivos Generalizados para Locação, Escala e Forma

Os Modelos Aditivos Generalizados para Locação, Escala e Forma (*Generalized Additive Models for Location, Scale and Shape (GAMLSS)*) foram introduzidos por Rigby & Stasinopoulos (2001), que os desenvolveram para permitir relaxar algumas suposições exigidas na classe dos modelos lineares generalizados.

A classe dos *GAMLSS* apresenta duas vantagens em relação aos demais modelos lineares generalizados. Primeiramente, permite relaxar a suposição sobre a variável resposta admitindo distribuições mais gerais, e não somente aquelas restritas à família exponencial; e além disso, a parte sistemática do modelo é expandida, e não somente a média (ou locação), mas todos os parâmetros da distribuição condicional da variável resposta são modelados como funções paramétricas ou não paramétricas (*smooth*) envolvendo variáveis explicativas e/ou efeitos aleatórios.

Tal classe de modelos é definida conforme segue. Os p parâmetros $\boldsymbol{\theta}^\top = (\theta_1, \ldots, \theta_p)$ de uma função densidade de probabilidade $f(y|\boldsymbol{\theta})$ são modelados utilizando termos aditivos. Aqui, supõe-se que para $i = 1, \ldots, n$ as y_i observações são independentes e condicionalmente a $\boldsymbol{\theta}^i$, têm função densidade de probabilidade $f(y_i|\boldsymbol{\theta}^i)$, em que $\boldsymbol{\theta}^{i\top} = (\theta_{i1}, \ldots, \theta_{ip})$ é um vetor de p parâmetros relacionados às variáveis explicativas e efeitos aleatórios.

Seja $\boldsymbol{y}^\top = (y_1, \ldots, y_n)$ o vetor de observações da variável resposta. Considere ainda, para $k = 1, \ldots, p$, uma função de ligação monótona $g_k(\cdot)$, a qual relaciona o k-ésimo parâmetro $\boldsymbol{\theta}_k$ às variáveis explicativas e efeitos aleatórios por meio de um modelo aditivo dado por

$$g_k(\boldsymbol{\theta}_k) = \boldsymbol{\eta}_k = \boldsymbol{X}_k\boldsymbol{\beta}_k + \sum_{j=1}^{J_k} \boldsymbol{Z}_{jk}\boldsymbol{\gamma}_{jk}, \tag{1.1}$$

em que $\boldsymbol{\theta}_k$ e $\boldsymbol{\eta}_k$ são vetores de dimensão n x 1, por exemplo, $\boldsymbol{\theta}_k^\top = (\theta_{1k}, \theta_{2k}, \ldots, \theta_{nk})$, $\boldsymbol{\beta}_k^\top = (\beta_{1k}, \ldots, \beta_{J'_k k})$ é um vetor de parâmetros de dimensão J'_k, e $\boldsymbol{X}_k$ e $\boldsymbol{Z}_{jk}$ são matrizes do planejamento (covariáveis) fixas, conhecidas e de ordens n x J'_k e n x q_{jk}, respectivamente, e $\boldsymbol{\gamma}_{jk}$ é uma variável aleatória q_{jk}-dimensional (Rigby & Stasinopoulos, 2005).

No caso em que $J_k = 0$, o modelo (1.1) reduz-se a um modelo linear completamente

paramétrico dado por

$$g_k(\boldsymbol{\theta}_k) = \boldsymbol{\eta}_k = \boldsymbol{X}_k\boldsymbol{\beta}_k. \tag{1.2}$$

Se $\boldsymbol{Z}_{jk} = \boldsymbol{I}_n$, em que $\boldsymbol{I}_n$ é uma matriz identidade de ordem n x n e $\boldsymbol{\gamma}_{jk} = \boldsymbol{h}_{jk} = \boldsymbol{h}_{jk}(\boldsymbol{x}_{jk})$ para todas as combinações de j e k do modelo (1.1), tem-se

$$g_k(\boldsymbol{\theta}_k) = \boldsymbol{\eta}_k = \boldsymbol{X}_k\boldsymbol{\beta}_k + \sum_{j=1}^{J_k} \boldsymbol{h}_{jk}(\boldsymbol{x}_{jk}), \tag{1.3}$$

em que $\boldsymbol{x}_{jk}$ são vetores de tamanho n, para $j = 1, 2, \ldots, J_k$ e $k = 1, 2, \ldots, p$. A função $\boldsymbol{h}_{jk}$ é uma função desconhecida da variável explanatória X_{jk} e $\boldsymbol{h}_{jk} = \boldsymbol{h}_{jk}(\boldsymbol{x}_{jk})$ é um vetor que avalia $\boldsymbol{h}_{jk}$ em $\boldsymbol{x}_{jk}$. As variáveis explicativas que compõem o vetor $\boldsymbol{x}_{jk}$ são assumidas conhecidas. O modelo (1.3) é chamado de *GAMLSS* semiparamétrico, o qual compõe um caso particular do modelo (1.1). Desse modo, a classe dos *GAMLSS* contempla termos paramétricos, semiparamétricos e efeitos aleatórios.

O modelo (1.1) pode ser estendido para permitir a inclusão de termos não lineares na modelagem dos parâmetros da distribuição, na forma

$$g_k(\boldsymbol{\theta}_k) = \boldsymbol{\eta}_k = h_k(\boldsymbol{X}_k, \boldsymbol{\beta_k}) + \sum_{j=1}^{J_k} \boldsymbol{h}_{jk}(\boldsymbol{x}_{jk}), \tag{1.4}$$

em que h_k para $k = 1, \ldots, p$ são funções não lineares e $\boldsymbol{X}_k$ é uma matriz de covariáveis conhecida de ordem n x $J_k^{''}$. O modelo (1.4) é chamado de *GAMLSS* semiparamétrico não linear. Se $J_k = 0$, então (1.4) se reduz a um modelo *GAMLSS* paramétrico não linear expresso por

$$g_k(\boldsymbol{\theta}_k) = \boldsymbol{\eta}_k = h_k(\boldsymbol{X}_k, \boldsymbol{\beta_k}); \tag{1.5}$$

se $h_k(\boldsymbol{X}_k, \boldsymbol{\beta_k}) = \boldsymbol{X}_k\boldsymbol{\beta}_k$, para $i = 1, \ldots, n$ e $k = 1, \ldots, p$, então, o modelo (1.5) se reduz ao modelo *GAMLSS* paramétrico linear (1.2). Note que alguns termos de $h_k(\boldsymbol{X}_k, \boldsymbol{\beta_k})$ podem ser lineares, o que resulta num modelo com a combinação de termos paramétricos lineares e não lineares.

Em muitas situações práticas são requeridos no máximo quatro parâmetros. Os dois primeiros, $\boldsymbol{\theta}_1$ e $\boldsymbol{\theta}_2$ são usualmente caracterizados como parâmetros de locação e escala, denotados por $\boldsymbol{\mu} = \boldsymbol{\theta}_1$ e $\boldsymbol{\sigma} = \boldsymbol{\theta}_2$, os demais parâmetros, quando presentes, são caracterizados como parâmetros de forma, denotados por $\boldsymbol{\nu} = \boldsymbol{\theta}_3$ e $\boldsymbol{\tau} = \boldsymbol{\theta}_4$. Dessa forma, os *GAMLSS* podem ser escritos resumidamente de forma genérica como:

$$
\begin{aligned}
g_1(\boldsymbol{\theta}_1) &= \boldsymbol{\eta}_1 = \boldsymbol{X}_1\boldsymbol{\beta}_1 + \sum_{j=1}^{J_1} \boldsymbol{Z}_{j1}\boldsymbol{\gamma}_{j1}, \\
g_2(\boldsymbol{\theta}_2) &= \boldsymbol{\eta}_2 = \boldsymbol{X}_2\boldsymbol{\beta}_2 + \sum_{j=1}^{J_2} \boldsymbol{Z}_{j2}\boldsymbol{\gamma}_{j2}, \\
g_3(\boldsymbol{\theta}_3) &= \boldsymbol{\eta}_3 = \boldsymbol{X}_3\boldsymbol{\beta}_3 + \sum_{j=1}^{J_3} \boldsymbol{Z}_{j3}\boldsymbol{\gamma}_{j3}, \\
g_4(\boldsymbol{\theta}_4) &= \boldsymbol{\eta}_4 = \boldsymbol{X}_4\boldsymbol{\beta}_4 + \sum_{j=1}^{J_4} \boldsymbol{Z}_{j4}\boldsymbol{\gamma}_{j4}.
\end{aligned}
$$

Na linguagem de programação *R* (R Core Team, 2008), dois métodos de estimação estão propostos para esta classe de modelos, no contexto utilizado neste estudo (Stasinopoulos et al., 2008). O primeiro é o da maximização de uma função de verossimilhança penalizada, no qual é utilizado um algoritmo de retroajuste (*backfitting*) para estimação dos parâmetros. Este procedimento pode ser encontrado na rotina `gamlss` (Rigby & Stasinopoulos, 2006a). O segundo é o da maximização da verossimilhança marginal aproximada, o qual faz uso do algoritmo EM – *Expectation-Maximization* – para calcular as estimativas dos parâmetros, presente na rotina `gamlss.mx` (Stasinopoulos et al., 2008). Maiores detalhes sobre esta classe de modelos estão relatados no Apêndice A.1.

Neste trabalho, a classe dos *GAMLSS* é utilizada com quatro objetivos. Primeiramente, como ferramenta para seleção da distribuição de probabilidade que melhor se ajusta ao conjunto de dados que motivou o trabalho (detalhes no Capítulo 3), utilizando-se a função *fitDist*. Tal função ajusta distribuições candidatas ao conjunto de dados e retorna o critério de Akaike (Akaike, 1973), através do qual seleciona-se a distribuição que apresenta o menor valor para este critério.

O segundo objetivo do uso da classe dos *GAMLSS* está relacionado com a programação desenvolvida. Na rotina `gamlss` do programa *R* está implementado o ajuste do modelo Box-Cox *t* com efeito aleatório quando este segue uma distribuição normal. Desse modo, a programação feita no presente estudo utiliza a abordagem da rotina `gamlss` como comparação para verificar se a programação realizada está de acordo com o esperado. O programa desenvolvido é baseado no método da maximização da verossimilhança marginal aproximada, utilizando métodos iterativos implementados em *R* para a solução requerida (detalhes na Seção 2.2). O desenvolvimento do programa é feito para flexibilizar a distribuição do efeito aleatório associado ao modelo Box-Cox *t* e resolver alguns problemas de convergência na estimação de parâmetros em modelos

para dados com assimetria acentuada.

Nos estudos de simulação realizados (descritos no Capítulo 4), esta rotina também é usada com modelos sem efeito aleatório para atribuir valores iniciais para o processo iterativo do cálculo das estimativas dos parâmetros. Por fim, no Capítulo 5, essa rotina é utilizada na comparação entre os modelos Box-Cox simétricos e os propostos por Cordeiro & Andrade (2011) e Azzalini (2005).

1.3 Distribuição Box-Cox *t*

A distribuição Box-Cox *t* (Rigby & Stasinopoulos, 2006b) é proposta como alternativa à transformação de Box & Cox (1964). Esta distribuição é útil quando a variável dependente Y apresenta uma distribuição assimétrica e leptocúrtica. Ela é indexada por quatro parâmetros e denotada nesta tese por $BCT(\mu, \sigma, \nu, \tau)$.

Seja Y uma variável aleatória positiva e contínua. A distribuição Box-Cox *t* (BCT) é definida a partir da transformação

$$Z = Z(Y) = \begin{cases} \dfrac{1}{\sigma\nu}\left[\left(\dfrac{Y}{\mu}\right)^{\nu} - 1\right], & \text{se} \quad \nu \neq 0, \\ \dfrac{1}{\sigma}\log\left(\dfrac{Y}{\mu}\right), & \text{se} \quad \nu = 0, \end{cases} \tag{1.6}$$

em que $\mu > 0$, $\sigma > 0$, $-\infty < \nu < \infty$. Assume-se que Z tem distribuição t padrão com τ graus de liberdade ($\tau > 0$) truncada com suporte no intervalo

$$A(\sigma, \nu) = \begin{cases} \left(-\dfrac{1}{\sigma\nu}, \infty\right), & \text{se} \quad \nu > 0, \\ \left(-\infty, -\dfrac{1}{\sigma\nu}\right), & \text{se} \quad \nu < 0, \\ (-\infty, \infty), & \text{se} \quad \nu = 0, \end{cases}$$

e escreve-se $Z \sim t_{\tau, A(\sigma,\nu)}$. Assim, diz-se que Y tem distribuição BCT com parâmetros μ, σ, ν e τ, e escreve-se $Y \sim BCT(\mu, \sigma, \nu, \tau)$, se Z dado em (1.6) é tal que $Z \sim t_{\tau, A(\sigma,\nu)}$. Nota-se que o suporte da distribuição de Y é o conjunto dos números reais positivos.

A função densidade de probabilidade de Y é dada por

$$f_Y(y) = f_Z(z)\left|\frac{dz}{dy}\right| = \frac{y^{\nu-1}}{\mu^{\nu}\sigma} f_Z(z), \tag{1.7}$$

em que $z = Z(y)$, com $Z(\cdot)$ dado em (1.6) e

$$f_Z(z) = \frac{f_T(z)}{F_T\left(\frac{1}{\sigma \mid \nu \mid}\right)},$$

com $T \sim t_\tau$, ou seja, T é uma variável aleatória que segue uma distribuição t padrão com τ graus de liberdade, cuja função densidade de probabilidade é dada por

$$f_T(t) = \frac{\Gamma\left[\frac{\tau+1}{2}\right]}{\Gamma\left(\frac{1}{2}\right)\Gamma\left(\frac{\tau}{2}\right)\tau^{\frac{1}{2}}}\left(1 + \frac{t^2}{\tau}\right)^{-\frac{\tau+1}{2}}, \quad -\infty < t < \infty, \tag{1.8}$$

e $F_T(\cdot)$ é a função distribuição acumulada de T.[1]

A função distribuição acumulada de Y é dada por

$$F_Y(y) = \begin{cases} \dfrac{F_T(z)}{F_T\left(\frac{1}{\sigma \mid \nu \mid}\right)}, & \text{se} \quad \nu \leq 0, \\ \dfrac{F_T(z) - F_T\left(-\frac{1}{\sigma \mid \nu \mid}\right)}{F_T\left(\frac{1}{\sigma \mid \nu \mid}\right)}, & \text{se} \quad \nu > 0. \end{cases} \tag{1.9}$$

A probabilidade da região de truncamento, $\mathbb{R}\backslash A(\sigma, \nu)$, é dada por $F_T(-(\sigma|\nu|)^{-1})$. Se a região de truncamento tem probabilidade negligenciável sob a distribuição t_τ, o que ocorre quando $\sigma|\nu|$ é pequeno, então $F_T(-(\sigma|\nu|)^{-1}) \approx 0$ e, consequentemente, $F_T((\sigma|\nu|)^{-1}) \approx 1$ e, então, $f_Z(z) \approx f_T(z)$ e $F_y(y) \approx F_T(z)$.

Sejam y_α e z_α os quantis de ordem α das distribuições de Y e de Z, respectivamente. Pode-se mostrar que

$$y_\alpha = \begin{cases} \mu(1 + \sigma\nu z_\alpha)^{\frac{1}{\nu}}, & \text{se} \quad \nu \neq 0, \\ \mu \exp(\sigma z_\alpha), & \text{se} \quad \nu = 0, \end{cases}$$

e

$$z_\alpha = \begin{cases} F_T^{-1}\left[\alpha F_T\left(\frac{1}{\sigma \mid \nu \mid}\right)\right], & \text{se} \quad \nu \leq 0, \\ F_T^{-1}\left[1 - (1-\alpha)F_T\left(\frac{1}{\sigma \mid \nu \mid}\right)\right], & \text{se} \quad \nu > 0, \end{cases}$$

com $F_T^{-1}(\cdot)$ representando a inversa da função distribuição acumulada de $T \sim t_\tau$. Se a região de truncamento tem probabilidade negligenciável, então $z_\alpha = t_{\tau,\alpha}$, em que $t_{\tau,\alpha} = F_T^{-1}(\alpha)$, o quantil de ordem α da distribuição t, $T \sim t_\tau$.

[1] Daqui em diante $1/\sigma|\nu|$ quando $\sigma|\nu| = 0$, será interpretado como $\lim_{\sigma\nu \to 0}(1/\sigma|\nu|) = \infty$ e, nesse caso, $F(1/\sigma|\nu|) = 1$.

Em particular, a mediana de Y é dada por

$$y_{1/2} = \begin{cases} \mu(1 + \sigma\nu z_{1/2})^{\frac{1}{\nu}}, & \text{se} \quad \nu \neq 0, \\ \mu \exp(\sigma z_{1/2}), & \text{se} \quad \nu = 0, \end{cases} \tag{1.10}$$

em que

$$z_{1/2} = \begin{cases} F_T^{-1}\left[\frac{1}{2}F_T\left(\frac{1}{\sigma \mid \nu \mid}\right)\right], & \text{se} \quad \nu \leq 0, \\ F_T^{-1}\left[1 - \frac{1}{2}F_T\left(\frac{1}{\sigma \mid \nu \mid}\right)\right], & \text{se} \quad \nu > 0. \end{cases}$$

Note-se que, se a região de truncamento tem probabilidade negligenciável, tem-se $z_{1/2} \approx 0$ e $\mu \approx y_{1/2}$, ou seja, o parâmetro μ é aproximadamente igual à mediana da distribuição $BCT(\mu, \sigma, \nu, \tau)$.

Um coeficiente de variação para uma variável aleatória Y baseado em percentis, denotado por CV_Y, é definido como

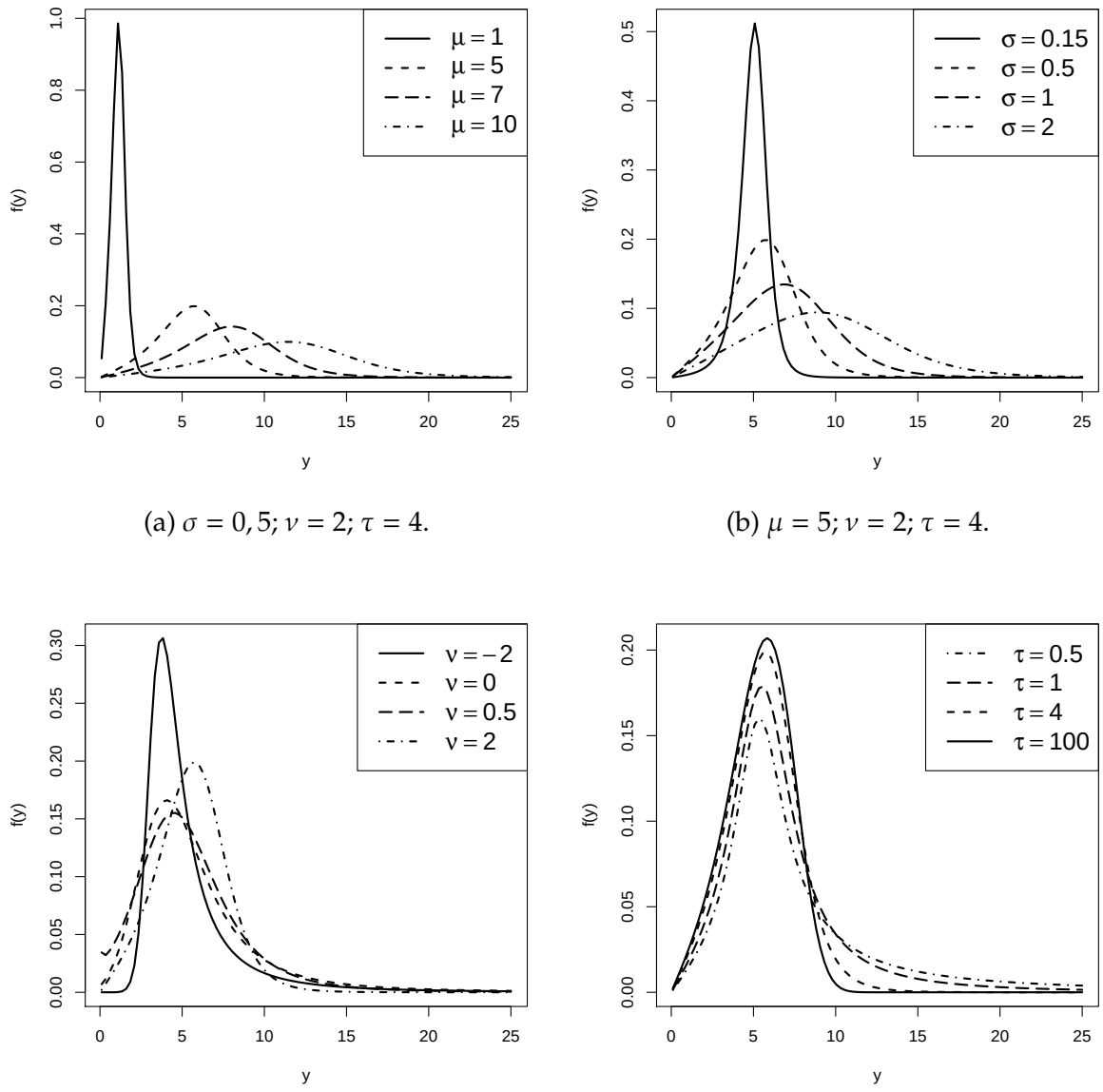
$$CV_Y = \frac{3}{4}\frac{y_{0,75} - y_{0,25}}{y_{0,5}}.$$

Ignorando a região de truncamento, sendo $\nu \neq 0$ e considerando σ pequeno, tem-se que $CV_Y \approx 0,75\{[1 + \sigma\nu t_{\tau;0,75}]^{1/\nu} - [1 - \sigma\nu t_{\tau;0,75}]^{1/\nu}\}$, em que $t_{\tau;0,75}$ é o terceiro quartil de uma variável aleatória T que segue uma distribuição t padrão com τ graus de liberdade. Segundo Johnson et al. (1982, p.375), uma aproximação para os quantis da distribuição t_τ é dada por $t_{\tau;\alpha} \approx u_\alpha + u_\alpha(u_\alpha^2 + 1)/(4\tau)$, em que u_α é o percentil de ordem α de uma variável aleatória normal padrão. Assim, utilizando esta aproximação e considerando $u_{0,75} \approx 0,67 \approx 2/3$, tem-se que $CV_Y \approx \sigma[1 + 0,36/\tau]$. Para $\nu \approx 0$ tem-se que $CV_Y \approx 1,5$ senh$(\sigma t_{\tau;0,75})$, uma função crescente em σ (vale a igualdade quando $\nu = 0$), em que senh$(\cdot)$ é a função seno hiperbólico.

Assim, tem-se que os parâmetros μ, σ, ν e τ podem ser interpretados como escala (relacionado à mediana), dispersão relativa (associado ao coeficiente de variação aproximado baseado nos percentis), assimetria (transformação potência para simetria) e curtose (graus de liberdade), respectivamente.

Detalhes sobre o logaritmo da função de verossimilhança, as derivadas de primeira ordem e o valor esperado das derivadas de segunda ordem e das derivadas cruzadas da distribuição Box-Cox *t* são dados no Apêndice A.2.

A Figura 1.1 ilustra a função de densidade de probabilidade da distribuição Box-Cox *t* para várias combinações de valores dos parâmetros. Note que τ controla o peso da cauda da distribuição.



(a) $\sigma = 0,5$; $\nu = 2$; $\tau = 4$.

(b) $\mu = 5$; $\nu = 2$; $\tau = 4$.

(c) $\mu = 5$; $\sigma = 0,5$; $\tau = 4$.

(d) $\mu = 5$; $\sigma = 0,5$; $\nu = 2$.

Figura 1.1: Gráficos de função densidade probabilidade da distribuição $BCT(\mu, \sigma, \nu, \tau)$.

1.4 Distribuição Box-Cox Cole-Green

Sabe-se que, se uma variável aleatória unidimensional segue uma distribuição t e seu número de graus de liberdade tende a infinito, esta variável tende à distribuição normal (Johnson et al., 1982, p.363). Analogamente, a distribuição Box-Cox t, quando o parâmetro referente ao número de graus de liberdade tende a infinito, converge para a chamada distribuição Box-Cox Cole-Green (BCCG), baseada no modelo normal. Por este motivo, tal distribuição é apresentada neste trabalho como comparação ao modelo

proposto, em especial para os casos nos quais não há presença de *outliers* extremos.

A distribuição BCCG surge a partir do método LMS (*Lambda – Mu – Sigma*) proposto por Cole & Green (1992). Este método foi criado para traçar curvas de percentis que mostram a evolução de uma medida em função de uma covariável de interesse; em geral são usadas como curvas para medir crescimento em função da idade. No método LMS três parâmetros estão envolvidos: o parâmetro λ da transformação potência Box-Cox, a média, denotada por μ, e o coeficiente de variação, denotado por σ; as letras iniciais dos nomes atribuídos aos parâmetros originaram o nome ao método.

O trabalho de Cole & Green (1992) apresenta a transformação dada em (1.6), tendo a variável transformada Z seguindo uma distribuição normal padrão. Esta distribuição foi implementada na rotina `gamlss` presente no *software R* e chamada de distribuição Box-Cox Cole-Green (BCCG). É importante ressaltar que na implementação feita por Stasinopoulos et al. (2008) é considerada a região de truncamento para a variável transformada (análoga à explicitada na Seção 1.3), uma vez que a variável aleatória que segue uma distribuição BCCG é estritamente positiva.

As distribuições Box-Cox *t* (BCT) e Box-Cox exponencial potência (BCPE) (Rigby & Stasinopoulos, 2004; Voudouris et al., 2012) surgem como uma extensão do método LMS com a inclusão de um parâmetro que flexibiliza a curtose da distribuição.

1.5 Objetivos e estrutura da tese

O objetivo desta tese é propor uma nova metodologia para a estimação da distribuição usual de consumo e da prevalência de inadequação alimentar de uma população baseada no consumo mediano dos indivíduos, para casos nos quais os dados são altamente assimétricos e/ou possuem observações discrepantes.

Para tal finalidade, a metodologia baseia-se na distribuição Box-Cox *t*. A partir do estudo desta distribuição, propõe-se também uma generalização para uma família de distribuições baseada na transformação dada em (1.6) e nas distribuições simétricas; tal generalização é chamada nesta tese de distribuições Box-Cox simétricas.

Este trabalho está organizado em seis capítulos. No Capítulo 2, é apresentada uma revisão sobre o método NCI (Tooze et al., 2010), já consolidado na literatura, com ênfase no seu caso particular dado pela quantidade de consumo, bem como a proposta da nova metodologia baseada na distribuição Box-Cox *t* (Rigby & Stasinopoulos, 2006b). No Capítulo 3, é realizada uma aplicação a um banco de dados real, na qual a nova metodologia e o modelo já existente na literatura são utilizados. No Capítulo 4 é feito um estudo de simulação a fim de corroborar o modelo proposto e compará-lo com o modelo já estabelecido. No Capítulo 5, apresenta-se a definição da classe de distribuições Box-Cox simétricas (BCS) e algumas de suas propriedades, bem como

uma comparação dessa classe com os modelos simétricos transformados (Cordeiro & Andrade, 2011) e os assimétricos propostos por Azzalini (2005). Por fim, no Capítulo 6 são apresentadas as principais contribuições dessa tese e propostas para estudos futuros.

1.6 Suporte computacional

A implementação computacional para a estimação do modelo descrito na Seção 2.2 foi desenvolvida na linguagem de programação *R*, versão 3.0.1 (R Core Team, 2008); detalhes em *http://www.R-project.org*.

As macros MIXTRAN e DISTRIB do programa SAS (*Statistical Analysis System*)(versão 9.2) foram usadas para estimação dos parâmetros do método NCI definido na Seção 2.1 (SAS, 1990); para consulta acessar *http://www.sas.com/*.

Para o estudo de simulação com a implementação feita no programa *R* utilizou-se o *cluster* Puma, administrado pelo Laboratório de Computação Científica Avançada (LCCA) da Universidade de São Paulo, que pode ser acessado em *http://www.usp.br/lcca/index.html*.

Para cálculos feitos no Capítulo 5 o *software* Maple 13 foi utilizado; detalhes em *http://www.maplesoft.com*.

Para a escrita desta tese foi utilizado o sistema tipográfico LATEX, que consiste em uma série de macros e rotinas baseadas no sistema TEX; detalhes podem ser encontrados em *http://www.latex-project.org/*.

CAPÍTULO 2

Modelo Box-Cox t com efeitos aleatórios e aplicações a dados de consumo alimentar

O modelo Box-Cox t com efeito aleatório pode ser aplicado em diversas áreas do conhecimento, todavia nesta tese tem como objetivo específico compor uma nova proposta para a estimação da distribuição usual de consumo e da prevalência de inadequação alimentar. Por isso seu desenvolvimento é apresentado fazendo um paralelo ao método NCI, especificamente num caso particular desse método, o qual considera somente a quantidade de consumo de micronutrientes ingerida.

2.1 Método NCI

O método NCI foi proposto para estimar a distribuição do consumo usual diário de alimentos ou nutrientes (Tooze et al., 2010). A peça central dessa metodologia é um modelo de medidas repetidas composto por duas partes com efeitos aleatórios correlacionados. Os dois componentes consistem na probabilidade e na quantidade diária de consumo de um determinado alimento ou nutriente. Entretanto, quando a probabilidade de consumo do alimento ou nutriente é próxima ou igual a 1, somente a parte da quantidade de consumo é requerida, tratando-se assim de um caso especial do modelo de duas partes, que é o considerado ao longo desta tese.

O método NCI tem como instrumento de coleta de dados os chamados recordatórios 24 horas (R24h). Conforme visto anteriormente, esta ferramenta é um registro aberto, no qual o indivíduo relata seu consumo alimentar nas últimas 24 horas anteriores ao

momento da entrevista.

O caso particular que considera somente a quantidade de consumo do método NCI é descrito a seguir (Tooze et al., 2010). Seja T_{ij} o verdadeiro consumo de um nutriente para o indivíduo i ($i = 1, \ldots, N$) no dia j ($j = 1, \ldots, n_i$). O consumo usual para o indivíduo i é dado pela esperança condicional de seu consumo, ou seja, $T_i = E(T_{ij}|i)$. O consumo medido a partir de um recordatório 24 horas (R24h) será denotado por R_{ij}.

Assim, supõe-se que o R24h é um instrumento não viesado para o consumo usual diário, ou seja, $E(R_{ij}|i) = T_i$. Entretanto, devido à assimetria da distribuição de consumo, uma transformação nos dados provenientes do R24h é aplicada. O método NCI usa a transformação Box-Cox, que inclui a transformação logarítmica como caso limite. Logo, define-se $R^*_{ij} = g(R_{ij}, \nu)$, sendo $g(r, \nu) = (r^\nu - 1)\nu^{-1}$. Quando $\nu = 0$, a transformação do logaritmo natural é utilizada (Box & Cox, 1964). Sob a escala transformada, supõe-se que

$$\begin{aligned} R^*_{ij} &= \mu^*_i + e_{ij}, \\ \mu^*_i &= \mu_i + \gamma_i, \end{aligned}$$

em que os erros intrapessoais, e_{ij}, são independentes, com $e_{ij} \sim N(0, \sigma^2_e)$, $\sigma^2_e > 0$. Pode ser incorporado ao modelo um vetor de covariáveis e, neste caso, assume-se que $\mu_i = \boldsymbol{X}_i\boldsymbol{\beta}$, em que $\boldsymbol{\beta}' = (\beta_0, \beta_1, \beta_2, \ldots, \beta_p)$ é o vetor de parâmetros p de dimensão $p+1$ associado a p covariáveis, $\boldsymbol{X}$ é a matriz de ordem N x $(p+1)$, com i-ésima linha $\boldsymbol{X}_i = (1, X_{i1}, X_{i2}, \ldots, X_{ip})$ correspondente às variáveis explicativas do i-ésimo indivíduo; e os erros interpessoais, γ_i, são independentes, com $\gamma_i \sim N(0, \lambda^2)$, $\lambda^2 > 0$. Combinando as equações anteriores, tem-se o modelo de efeito aleatório misto não linear para o R24h,

$$g(R_{ij}, \nu) = R^*_{ij} = \boldsymbol{X}_i\boldsymbol{\beta} + \gamma_i + e_{ij},$$

com $\gamma_i \sim N(0, \lambda^2)$ e $e_{ij} \sim N(0, \sigma^2_e)$ variáveis aleatórias independentes. Desse modo, tem-se que $R^*_{ij}|\gamma_i \sim N(\boldsymbol{X}_i\boldsymbol{\beta} + \gamma_i, \sigma^2_e)$ e são independentes.

É importante ressaltar que a variabilidade interpessoal é considerada através do efeito aleatório γ_i e a variabilidade intrapessoal é absorvida juntamente com as outras fontes de variação casuais, através da variável aleatória e_{ij}.

Desse modo, a contribuição do indivíduo i, $i = 1, \ldots, N$, para a verossimilhança é dada pela função densidade marginal de $(R_{i1}, \cdots, R_{in_i})$, ou seja

$$L_i(\nu, \boldsymbol{\beta}, \lambda, \sigma_e; r_{i1}, \ldots, r_{in_i}) = \int_{-\infty}^{\infty} \prod_{j=1}^{n_i} p(r_{ij}|\nu, \boldsymbol{\beta}, \gamma_i, \sigma_e) h(\gamma_i|\lambda) \mathrm{d}\gamma_\mathrm{i},$$

em que

$$p(r_{ij}|\nu, \boldsymbol{\beta}, \gamma_i, \sigma_e) = \frac{1}{\sqrt{2\pi\sigma_e^2}} \exp\left\{ \frac{-\left(\frac{r_{ij}^{\nu}-1}{\nu} - (\boldsymbol{X}_i\boldsymbol{\beta} + \gamma_i)\right)^2}{2\sigma_e^2} \right\} r_{ij}^{\nu-1},$$

a densidade de $R_{ij}^*|\gamma_i$, e $h(\gamma_i|\lambda)$ é a função de densidade de uma variável aleatória normal com média zero e variância λ^2.

A função de verossimilhança do modelo, considerando os N indivíduos é dada por

$$L(\nu, \boldsymbol{\beta}, \sigma_e, \lambda) = \prod_{i=1}^{N} L_i(\nu, \boldsymbol{\beta}, \lambda, \sigma_e; r_{i1}, \ldots, r_{in_i}) = \prod_{i=1}^{N} \int_{-\infty}^{\infty} \prod_{j=1}^{n_i} p(r_{ij}|\nu, \boldsymbol{\beta}, \gamma_i, \sigma_e) h(\gamma_i|\lambda) \mathrm{d}\gamma_i.$$

Assim, o logaritmo da função de verossimilhança pode ser escrito como

$$\ell(\nu, \boldsymbol{\beta}, \lambda, \sigma_e) = \sum_{i=1}^{N} \log \left[\int_{-\infty}^{\infty} \prod_{j=1}^{n_i} p(r_{ij}|\nu, \boldsymbol{\beta}, \gamma_i, \sigma_e) h(\gamma_i|\lambda) \mathrm{d}\gamma_i \right].$$

A integral envolvida no modelo não tem solução explícita e utiliza-se uma aproximação pela quadratura de Gauss adaptativa (Pinheiro & Bates, 1995) para sua resolução. A verossimilhança é maximizada com a otimização feita pelo procedimento de quasi-Newton, implementado no PROC NLMIXED do *software* SAS.

Nesta tese, este modelo para a quantidade de consumo do método NCI será chamado simplesmente de método NCI ou modelo padrão.

Estimação da prevalência de inadequação alimentar

No método NCI, percentis da distribuição usual de consumo dos nutrientes de uma população são estimados usando um procedimento de Monte Carlo. Para gerar a distribuição dos valores que refletem o padrão das covariáveis na população, utilizam-se as covariáveis para os indivíduos amostrados, com as combinações entre os valores estimados do vetor de parâmetros associados às covariáveis ($\boldsymbol{\beta}$) e de seus respectivos componentes de variância (λ^2 e σ_e^2).

Para que o padrão das covariáveis seja compatível com o da amostra, calcula-se primeiramente $\boldsymbol{X}_i\widehat{\boldsymbol{\beta}}$ para cada indivíduo amostrado i, em que $\widehat{\boldsymbol{\beta}}$ é a estimativa do vetor de parâmetros $\boldsymbol{\beta}$. Em seguida, simulam-se k realizações do efeito aleatório, como uma variável aleatória normal $N(0, \widehat{\lambda^2})$, em que $\widehat{\lambda}$ é a estimativa de λ e é calculado $\mu_l^* = \boldsymbol{X}_l^*\widehat{\beta} + \gamma_l$, em que $l = 1, \ldots, kN$, e $\boldsymbol{X_i} = \boldsymbol{X_I}^*$, para $l = k(i-1)+1, \ldots, ki$. Usualmente simulam-se $k = 100$ valores por indivíduo.

Desse modo, a distribuição dos kN valores simulados ($\mu_l^*, l = 1, \cdots, kN$) reflete uma amostra do consumo usual transformado da população. Uma expansão em série de Taylor pode ser utilizada para aproximar $T_l = E(R_l|\boldsymbol{X}_l, \gamma_l) = E(g^{-1}(R_l^*, \nu)|\boldsymbol{X}_l^*, \gamma_l; \boldsymbol{\beta})$. Desta maneira, para a transformação Box-Cox, o procedimento inverso é dado por

$$T_l \simeq g^{-1}(\mu_l^*, \nu) + \frac{1}{2}\sigma_e^2 \frac{\partial^2\{g^{-1}(\mu_l^*, \nu)\}}{\partial \mu_l^{*2}} = (\mu_l^*\nu + 1)^{\frac{1}{\nu}} + \frac{1}{2}\sigma_e^2(1-\nu)(\mu_l^*\nu + 1)^{\frac{1}{\nu}-2}.$$

As estimativas de ν e σ_e^2 são utilizadas para se obter $\widetilde{T}_l$, os consumos usuais simulados. Os quantis amostrais dos kN valores $\widetilde{T}_l$ são os quantis populacionais estimados. Para o estudo em questão, os percentis de 5%, 10%, 25%, 50%, 75%, 90% e 95% da distribuição usual de consumo são considerados. Tais percentis serão utilizados no estudo de simulação descrito no Capítulo 4 desta tese.

Na área de nutrição, é comum o interesse dos pesquisadores em saber se determinado grupo populacional apresenta ou não adequabilidade no consumo habitual alimentar. A avaliação para medir o consumo é feita seguindo um padrão estabelecido. Desse modo, a partir de um ponto de corte especificado, define-se a chamada prevalência de inadequação alimentar, que é a proporção de indivíduos cujo consumo habitual se encontra abaixo desse valor instituído (Morimoto et al., 2012; Suitor & Gleason, 2002).

Em geral, no método NCI utiliza-se a necessidade média estimada, denotada por EAR (*Estimated Average Requeriments*) como ponto de corte para o cálculo da prevalência de inadequação. Essa medida compõe um sistema internacional de recomendações nutricionais conhecido como DRI (*Dietary Reference Intake*) proveniente do Instituto de Medicina da Academia Nacional de Ciências dos Estados Unidos (*Institute of Medicine (IOM) of the U.S. National Academy of Sciences*). A EAR é o valor médio estimado da ingestão de um nutriente para cobrir as necessidades de 50% dos indivíduos saudáveis de determinada faixa etária, estado civil e sexo (INS, 2003).

A prevalência de inadequação alimentar é estimada pelo método NCI como sendo

$$\widehat{\rho} = \frac{\#\widetilde{T}_l < EAR}{kN}, \tag{2.1}$$

em que EAR é o ponto de corte especificado, k é o número de simulações por indivíduo e N é o número de indivíduos, ou seja, $\widehat{\rho}$ é a proporção de $\widetilde{T}_l's$ menores que a EAR. Assim, a estimativa da prevalência de inadequação será dada pela proporção dos consumos usuais que está abaixo do padrão estabelecido pela EAR para o nutriente (Carriquiry, 1998), calculada sobre os kN valores gerados.

2.2 Modelo Box-Cox t com efeito aleatório ou misto

O método NCI é amplamente aplicado na literatura para determinar a prevalência de inadequação alimentar. Entretanto, para casos nos quais os dados são altamente assimétricos, nota-se que a transformação realizada nos dados não necessariamente conduz à normalidade, e, consequentemente, as estimativas dos consumos simuladas para os indivíduos não refletem os consumos usuais, os quais por sua vez, afetam as estimativas das prevalências de inadequação alimentar, que parecem bastante distantes do que se observa nos dados brutos em tais situações.

Para acomodar melhor tais casos de alta assimetria, propõe-se nesta tese o modelo Box-Cox t com efeitos aleatórios ou misto. Tal modelagem, além de proporcionar uma acomodação para assimetria, também apresenta como vantagem a robustez da estimação envolvida, sendo menos sensível a observações discrepantes. Além disso, o modelo evita a necessidade de uma transformação inversa na variável resposta.

Outro aspecto interessante da abordagem é que se utiliza a distribuição das medianas de consumo para o cálculo da prevalência de inadequação alimentar. Esta medida, por ser mais robusta, gera estimativas de prevalência de inadequação alimentar mais próximas da realidade dos dados brutos, em especial para os casos de dados com alta assimetria; maiores detalhes serão discutidos no Capítulo 4. Adicionalmente, o ponto de corte para o cálculo da prevalência de inadequação é igual ao utilizado no método NCI, a EAR, o qual parece ser mais coerente neste contexto, uma vez que o mesmo é valor médio de ingestão do nutriente baseado na mediana da necessidade populacional.

O modelo Box-Cox t é definido conforme segue. Seja R_{ij} o consumo do nutriente em estudo relatado no R24h do indivíduo i no dia j, para $i = 1, \ldots, N$ e $j = 1, \ldots, n_i$. Assume-se que $R_{ij}|\gamma_i \sim BCT(\mu_i, \sigma, \nu, \tau)$, sendo

$$\begin{aligned} \log(\mu_i) &= \boldsymbol{X}_i\boldsymbol{\beta} + \gamma_i, \\ \log(\sigma) &= \delta, \\ \nu &= \eta, \\ \log(\tau) &= \zeta, \end{aligned} \tag{2.2}$$

em que $\boldsymbol{X}_i = (1, X_{i1}, X_{i2}, \ldots, X_{ip})$ é a i-ésima linha da matriz de variáveis explicativas $\boldsymbol{X}$, correspondente ao i-ésimo indivíduo e $\boldsymbol{\beta}' = (\beta_0, \beta_1, \ldots, \beta_p)$ é o vetor de $p+1$ parâmetros associado às covariáveis do i-ésimo indivíduo, δ, η e ζ são parâmetros desconhecidos, associados ao coeficiente de variação, a assimetria e a curtose, respectivamente, e γ_i é o efeito aleatório associado ao $i-$ésimo indivíduo, que modela as medidas repetidas em diferentes momentos. Aqui, supõe-se que os γ_i's são variáveis aleatórias independentes

e identicamente distribuídas, para $i = 1, \ldots, N$.

Duas diferentes distribuições para o efeito aleatório γ_i são propostas: a distribuição normal, com média zero e variância λ^2 e a distribuição t central, com κ graus de liberdade e parâmetro de dispersão λ^2, cuja função densidade de probabilidade é da forma

$$h(\gamma_i|\lambda,\kappa) = \frac{1}{\lambda}\frac{\Gamma\left[\frac{\kappa+1}{2}\right]}{\Gamma\left(\frac{1}{2}\right)\Gamma\left(\frac{\kappa}{2}\right)\kappa^{\frac{1}{2}}}\left(1+\frac{\gamma_i^2}{\kappa\lambda^2}\right)^{-\frac{\kappa+1}{2}}, \gamma_i \in \mathbb{R}. \tag{2.3}$$

Nota-se que a variabilidade interpessoal é incorporada através do efeito aleatório presente no preditor linear do parâmetro μ do modelo Box-Cox t. Já a variabilidade intrapessoal junto com as outras fontes de variação casuais é modelada pela distribuição Box-Cox t.

Seja R_i o vetor de observações da variável resposta referente ao consumo do nutriente relatado no R24h do indivíduo i, $i = 1, \ldots, N$, o qual contém diferentes medidas de consumo relatadas em n_i diferentes dias da semana, ou seja, $R_i^\top = (R_{i1}, \ldots, R_{in_i})$. Seja $R = (R_1^\top, R_2^\top, \ldots, R_N^\top)^\top$ o vetor das observações da variável resposta referente ao consumo relatado no R24h dos N indivíduos. A função de verossimilhança, obtida da função densidade marginal de R é dada por

$$\begin{aligned} L(\sigma,\nu,\tau,\boldsymbol{\beta},\lambda,\kappa) &= \prod_{i=1}^{N} L_i(\sigma,\nu,\tau,\boldsymbol{\beta},\lambda,\kappa; r_{i1},\ldots,r_{in_i}) = \\ &= \prod_{i=1}^{N}\int_{-\infty}^{\infty}\prod_{j=1}^{n_i} f(r_{ij}|\sigma,\nu,\tau,\boldsymbol{\beta},\gamma_i)h(\gamma_i|\lambda,\kappa)\mathrm{d}\gamma_\mathrm{i}, \end{aligned}$$

em que h é a função densidade dos efeitos aleatórios, que segue uma normal com média zero e variância λ^2, ou t conforme definida em (2.3), e

$$f(r_{ij}|\sigma,\nu,\tau,\boldsymbol{\beta},\gamma_i) = \frac{r_{ij}^{\nu-1}}{[\exp(\boldsymbol{X}_i\boldsymbol{\beta}+\gamma_i)]^\nu\sigma}\frac{f_T(z_{ij})}{F_T\left(\frac{1}{\sigma|\nu|}\right)},$$

sendo $f_T(\cdot)$ e $F_T(\cdot)$ definidos, respectivamente, em (1.8) e (1.9).

O logaritmo da função de verossimilhança é, portanto, dado por

$$\ell(\sigma,\nu,\tau,\boldsymbol{\beta},\lambda,\kappa) = \sum_{i=1}^{N}\log\left[\int_{-\infty}^{\infty}\prod_{j=1}^{n_i} f(r_{ij}|\sigma,\nu,\tau,\boldsymbol{\beta},\gamma_i)h(\gamma_i|\lambda,\kappa)\mathrm{d}\gamma_\mathrm{i}\right]. \tag{2.4}$$

A integral em (2.4) não tem solução explícita e se propõe aproximá-la pelo método de quadratura de Gauss-Hermite (Pinheiro & Bates, 1995) (detalhes podem ser vistos no Apêndice A.3). As estimativas de máxima verossimilhança marginal aproximada

são obtidas através de métodos iterativos utilizando o pacote `bbmle` na linguagem de programação *R*. Detalhes sobre a implementação computacional são apresentados na Seção 2.3; ver também Apêndices A.1 e A.4.

Uma análise de perfil do logaritmo da função de verossimilhança é proposta para o modelo Box-Cox t com efeito aleatório quando este segue uma distribuição t. O perfil é estudado com a finalidade de verificar o comportamento de tal função e assim averiguar se as estimativas encontradas pelo método de máxima verossimilhança marginal aproximada refletem o máximo global.

Estimação da prevalência de inadequação alimentar

O procedimento para a estimação da prevalência de inadequação alimentar é análogo ao realizado pelo método NCI. Todavia a distribuição usual do consumo será baseada nos consumos medianos.

Desse modo, novamente os percentis da distribuição usual de consumo dos nutrientes da população serão estimados. Para gerar a distribuição dos valores que refletem o padrão das covariáveis na população, utilizam-se as covariáveis para os indivíduos amostrados, com as combinações entre os valores estimados do vetor de parâmetros associados às covariáveis ($\boldsymbol{\beta}$) e da variância do efeito aleatório (λ^2).

Primeiramente, calcula-se $\boldsymbol{X}_i\widehat{\boldsymbol{\beta}}$ para cada indivíduo amostrado i, em que $\widehat{\boldsymbol{\beta}}$ é a estimativa do vetor de parâmetros $\boldsymbol{\beta}$. Em seguida, simulam-se k realizações do efeito aleatório, como uma variável aleatória $N(0, \widehat{\lambda^2})$ em que $\widehat{\lambda^2}$ é a estimativa de λ, e é calculado $\mu_l^* = \exp\{\boldsymbol{X}_l^*\widehat{\boldsymbol{\beta}} + \gamma_l\}$, em que $l = 1, \ldots, kN$, e $\boldsymbol{X}_i = \boldsymbol{X}_l^*$, para $l = k(i-1)+1, \ldots, ki$. Adota-se $k = 100$ valores por indivíduo. Assim, a distribuição dos kN valores de μ_l^* reflete uma amostra do consumo usual da população.

Neste ponto, o método difere do NCI, pois não existe uma transformação inversa a ser realizada. Aqui, com base nos kN pseudo valores simulados, calculam-se os consumos medianos estimados a partir de (1.10) e denotados aqui por $\widetilde{T}_l$. Os quantis amostrais dos kN valores $\widetilde{T}_l$ são os quantis medianos populacionais estimados. Para o estudo em questão, percentis de 5%, 10%, 25%, 50%, 75%, 90% e 95% dos valores medianos estimados são considerados para realização do estudo de simulação (Capítulo 4).

Analogamente ao método NCI, porém usando os consumos medianos estimados, calcula-se a prevalência de inadequação alimentar. Assim, a proporção de consumos medianos dos indivíduos que se encontra abaixo da EAR dada em (2.1) estima essa prevalência.

O modelo Box-Cox Cole-Green (BCCG) com efeito aleatório é definido analogamente ao explicitado anteriormente, sendo um caso limite quando $\tau \longrightarrow \infty$ do modelo descrito nesta seção, conforme discutido na Seção 1.4. Os modelos BCT e BCCG

com efeito aleatório serão chamados ao longo desta tese de modelos medianos, por utilizarem os consumos medianos dos indivíduos para o cálculo da prevalência de inadequação alimentar.

Análise de resíduos

Propõe-se o uso dos resíduos quantílicos (Dunn & Smyth, 1996) para verificar a qualidade do ajuste do modelo.

O valor predito é da forma $\widehat{\mu_i} = \exp\{\boldsymbol{X}_i\widehat{\boldsymbol{\beta}} + \widehat{\gamma_i}\}$, em que $\widehat{\mu_i}$ é uma medida associada ao consumo mediano estimado para o i-ésimo indivíduo, $\boldsymbol{X}_i$ é a i-ésima linha da matriz de variáveis explicativas $\boldsymbol{X}$, correspondente ao i-ésimo indivíduo e $\widehat{\boldsymbol{\beta}}' = (\widehat{\beta_0}, \widehat{\beta_1}, \ldots, \widehat{\beta_p})$ é a estimativa do vetor de $p+1$ parâmetros associado às covariáveis do i-ésimo indivíduo e, $\widehat{\gamma_i}$ é o valor predito do efeito aleatório referente ao indivíduo i. A predição do efeito aleatório pelo valor predito de Bayes empírico é definida como (Usuga Manco, 2013, p.25)

$$\widehat{\gamma_i} = E[\gamma_i|R_{ij};\widehat{\boldsymbol{\beta}},\widehat{\sigma},\widehat{\nu},\widehat{\tau},\widehat{\lambda},\widehat{\kappa}] = \frac{\displaystyle\int_{-\infty}^{\infty}\gamma_i f(r_{ij}|\widehat{\boldsymbol{\beta}},\widehat{\sigma},\widehat{\nu},\widehat{\tau},\gamma_i)h(\gamma_i|\widehat{\lambda},\widehat{\kappa})\mathrm{d}\gamma_\mathrm{i}}{\displaystyle\int_{-\infty}^{\infty}f_i(r_{ij}|\widehat{\boldsymbol{\beta}},\widehat{\sigma},\widehat{\nu},\widehat{\tau},\gamma_i)h(\gamma_i|\widehat{\lambda},\widehat{\kappa})\mathrm{d}\gamma_\mathrm{i}}.$$

Para o cálculo do valor predito de Bayes empírico foram usadas as estimativas dos parâmetros encontradas por máxima verossimilhança marginal aproximada. Para o cálculo da integral foi utilizada a função `integrate` presente no programa *R*.

Seja $F(\cdot;\mu,\sigma,\nu,\tau)$ a função distribuição acumulada de $Y \sim BCT(\mu,\sigma,\nu,\tau)$ definida em (1.9); $F(Y;\mu,\sigma,\nu,\tau)$ é uma função contínua uniformemente distribuída sobre um intervalo unitário. Os resíduos quantílicos são definidos por

$$d_{ij} = \Phi^{-1}(F(R_{ij};\widehat{\mu_i},\widehat{\sigma},\widehat{\nu},\widehat{\tau})),$$

em que $\Phi^{-1}(\cdot)$ é a inversa da função distribuição acumulada de uma variável com distribuição normal padrão. Espera-se que os d_{ij} sigam aproximadamente o comportamento de uma distribuição normal padrão se o modelo produzir um bom ajuste.

2.3 Implementação computacional

A programação para o método de estimação por máxima verossimilhança marginal e perfilada [1], ambos procedimentos para se encontrar as estimativas dos parâmetros dos modelos Box-Cox *t* e Box-Cox Cole-Green com efeito aleatório, bem como a implementação dos resíduos quantílicos, foi elaborada no *software R*.

A rotina `gamlss` presente no *software R* apresenta uma implementação dos modelos medianos para o caso em que o efeito aleatório segue uma distribuição normal. Entretanto, a proposta da tese é trabalhar com modelos que apresentam procedimentos robustos, tanto para a variável resposta quanto para o efeito aleatório. Assim, uma implementação foi realizada a fim de possibilitar que a distribuição *t* também fosse proposta para os efeitos aleatórios, o que proporciona uma flexibilidade maior para modelar a variabilidade entre os indivíduos. Além disso, a implementação possibilitou a resolução de alguns problemas de convergência encontrados ao se utilizar a rotina `gamlss` do *R*, em relação ao ajuste de modelos que envolvem a distribuição Box-Cox Cole-Green em dados com presença de alta assimetria.

A rotina `gamlss` foi utilizada para verificação da programação. Adicionalmente, a rotina foi usada para a seleção dos modelos, para atribuir valores iniciais no processo de simulação, bem como para verificar a necessidade da inclusão de um efeito aleatório adicional para a variabilidade intrapessoal.

[1] Um estudo sobre a verossimilhança perfilada do modelo Box-Cox *t* é realizado no Capítulo 3 (vide Apêndice A.1).

CAPÍTULO 3

Aplicação

3.1 Análise descritiva dos dados

Os dados analisados nesta tese são provenientes de uma amostra de 368 indivíduos pertencentes a um estudo epidemiológico que teve como objetivo avaliar a adequação do consumo alimentar em idosos do município de Botucatu, interior do estado de São Paulo, Brasil.

A amostra foi coletada no ano de 2011 por meio da aplicação de recordatório 24 horas (R24h). Para cada indivíduo, no máximo três R24h foram obtidos em dias da semana não consecutivos, sendo um deles necessariamente no final de semana.

Os dados coletados pelo R24h foram transformados em dados de consumo de micronutrientes utilizando-se o programa NDSR (*Nutrition Data System for Research*). O detalhamento encontra-se no processo FAPESP 2008/10261-8 referente ao projeto "Avaliação da Adequação Nutricional na Terceira Idade", de José Eduardo Corrente.

Os consumos de nutrientes e suas recomendações são distintos para homens e mulheres; desse modo, as análises neste capítulo são feitas segundo o gênero dos indivíduos amostrados. A Tabela 3.1 mostra a média, o desvio padrão (DP), os valores mínimo e máximo e a mediana de consumo de 22 micronutrientes referentes a 136 homens (354 observações) e 232 mulheres (594 observações); a Figura 3.1 apresenta os *box-plot* ajustados (Hubert & Vandervieren, 2008).

Tabela 3.1: Medidas descritivas do consumo de micronutrientes em idosos. Botucatu, 2011.

Micronutriente (m)	Homens					Mulheres				
	Média	DP	Mínimo	Máximo	Mediana	Média	DP	Mínimo	Máximo	Mediana
Vitamina A (mcg)	1249,71	3163,41	1,86	41372,02	713,22	1135,33	1712,59	0,00	21011,75	756,86
Vitamina D (mcg)	4,33	3,42	0,03	27,35	3,64	3,94	2,83	0,00	25,36	3,40
Vitamina E (mg)	7,55	12,61	0,72	227,05	6,13	6,13	3,57	0,90	37,00	5,48
Vitamina K (mcg)	241,11	1656,86	3,65	31035,98	103,37	196,00	1145,12	6,59	23211,77	101,77
Vitamina C (mg)	209,41	856,62	0,00	10566,01	54,72	104,98	269,25	0,00	3931,45	54,08
Vitamina B1 (mg)	1,78	1,23	0,39	16,90	1,53	1,48	0,74	0,19	9,46	1,34
Vitamina B2 (mg)	1,69	0,91	0,29	10,80	1,55	1,52	0,72	0,23	8,90	1,41
Vitamina B3 (mg)	24,23	18,22	1,84	258,77	20,60	19,83	11,37	3,41	83,80	17,50
Vitamina B6 (mg)	2,08	2,09	0,31	34,00	1,77	1,67	0,82	0,21	6,54	1,52
Vitamina B12 (mcg)	6,25	10,73	0,09	87,74	3,47	7,54	18,20	0,01	191,48	2,93
Ácido pantotênico (mg)	5,37	3,13	1,00	42,47	5,00	4,72	1,98	0,74	18,92	4,44
Folato (mcg)	462,89	427,55	76,10	6518,08	399,81	367,44	204,26	49,75	2685,53	336,36
Cálcio (mg)	738,38	555,73	59,28	6488,50	637,66	645,45	357,54	39,28	2408,83	559,98
Fósforo (mg)	1295,06	836,94	251,72	11916,16	1174,11	1071,86	478,87	184,01	3742,67	1011,75
Magnésio (mg)	310,60	292,43	51,47	4763,40	276,55	247,22	116,88	45,96	942,24	227,18
Ferro (mg)	17,06	12,38	1,44	132,25	15,14	13,89	8,70	2,12	97,33	12,02
Zinco (mg)	13,05	7,40	2,31	85,20	11,99	10,41	5,11	1,82	44,76	9,60
Cobre (mg)	1,33	1,49	0,22	25,04	1,18	1,23	1,57	0,18	30,04	1,01
Selênio (mcg)	141,86	77,90	24,49	602,37	126,00	126,79	179,84	18,94	3858,30	102,59
Sódio (mg)	6243,03	19880,43	215,29	237803,69	3604,27	4174,16	10186,93	457,64	156500,00	3004,92
Potássio (mg)	3249,45	5444,21	389,31	99448,13	2763,79	2493,09	1079,42	424,04	9373,87	2344,87
Manganês (mg)	5,59	15,14	0,65	187,81	3,12	9,81	34,62	0,63	554,68	2,57

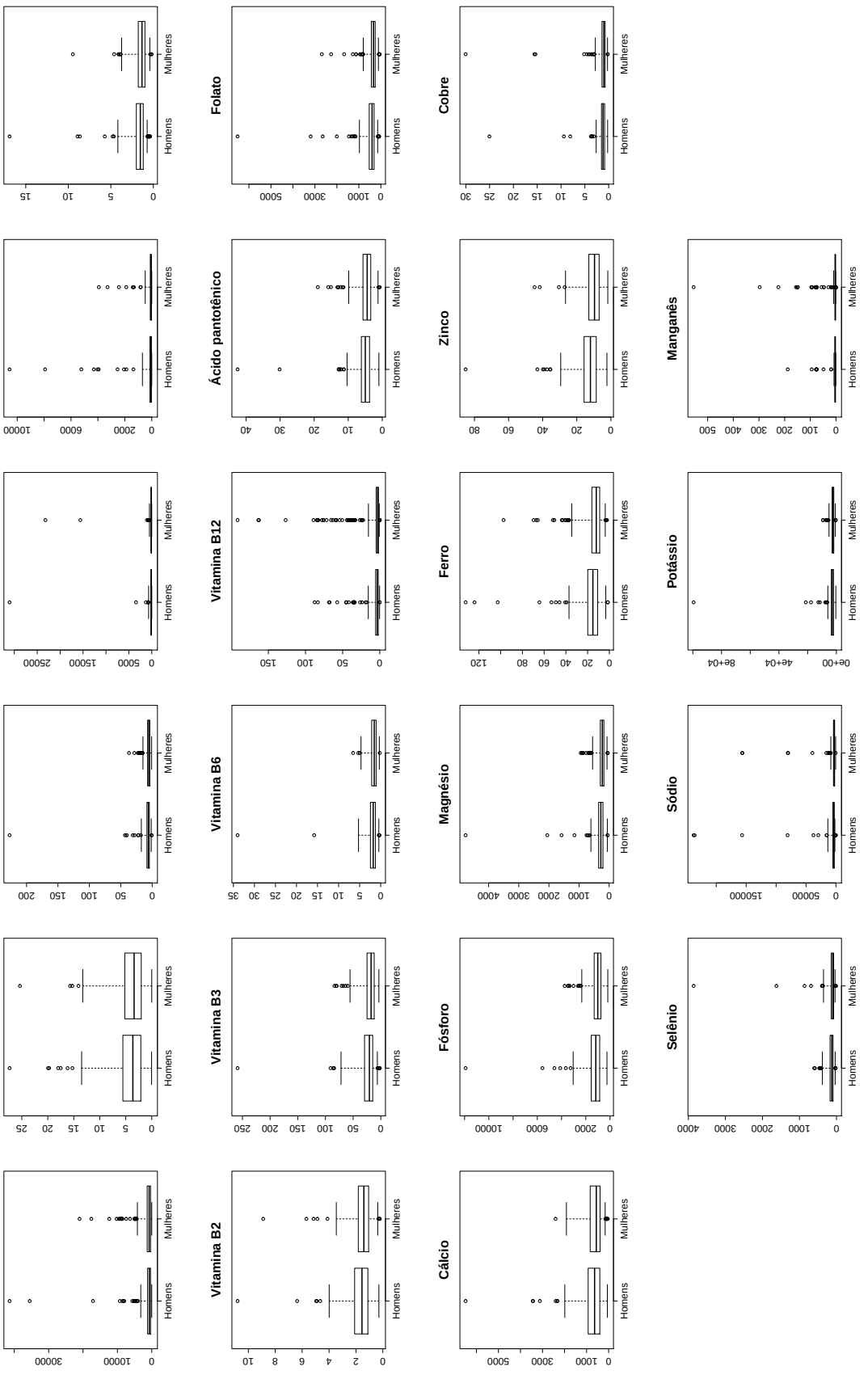


Figura 3.1: *Box – plot* ajustados do consumo de micronutrientes em idosos. Botucatu, 2011.

Os *box-plot* ajustados exibidos na Figura 3.1 evidenciam a assimetria das distribuições envolvidas (Hubert & Vandervieren, 2008). Ressaltam-se assimetrias acentuadas à direita nas distribuições de consumo dos micronutrientes, bem como a presença de pontos aberrantes extremos como, por exemplo, para o consumo de vitamina A em homens, para o qual o maior valor de consumo é de 41372,02 mcg, enquanto a mediana de consumos é de apenas 713,22 mcg.

Na Tabela 3.1, observa-se que os desvios padrões (DP) para alguns nutrientes, tais como as vitaminas A, K e C, foram muito altos indicando uma alta dispersão em torno da média. Observam-se alguns valores mínimos iguais a zero (vitamina C para homens e mulheres, e vitaminas A e D para mulheres), mas foram muito poucas observações, as quais foram consideradas com um pequeno acréscimo para efeito da análise estatística, pois o modelo proposto para este conjunto de dados contempla somente valores positivos.

Desse modo, as medidas descritivas relatadas sugerem uma proposta de modelagem estatística que contemple assimetria e estimação robusta.

3.2 Modelo Box-Cox *t* com efeito aleatório

Por meio da função *fitDist* da rotina `gamlss` do *software R*, verificou-se qual a distribuição de probabilidade que melhor se ajusta ao conjunto de dados brutos, sendo testadas para o consumo bruto de micronutrientes as seguintes distribuições: Box-Cox Cole-Green, Box-Cox exponencial potência, Box-Cox normal, Box-Cox *t*, exponencial, exponencial potência, família *t*, gama, gama generalizada, inversa gaussiana, inversa gaussiana generalizada, log-normal, *skew t*, valor extremo e Weibull (Rigby & Stasinopoulos, 2006a).

A Tabela 3.2 mostra o resumo da seleção das distribuições pelo critério de Akaike (AIC). Nota-se que a distribuição Box-Cox *t* é a mais frequente, seguida da distribuição *skew t*. Este fato é plausível, uma vez que a distribuição *t* é uma extensão paramétrica robusta do modelo normal para casos nos quais a presença de valores discrepantes é identificada (Arellano-Valle, 1994, p.58), algo muito comum em dados de consumo alimentar. Vale a pena ressaltar que nos casos para os quais a distribuição Box-Cox *t* não apresentou o menor valor para o critério de seleção, as diferenças dos valores encontrados entre a "melhor" distribuição e a Box-Cox *t* foram mínimas. O critério bayesiano de Schwarz (SBC) também foi aplicado e os resultados foram semelhantes ao AIC.

Tabela 3.2: Distribuições selecionadas pelo critério de Akaike para o consumo bruto de micronutrientes segundo gênero (Masculino e Feminino) a partir da ferramenta *fitDist* do pacote `gamlss`.

Micronutriente (m)	Distribuição selecionada	
	Masculino	Feminino
Vitamina A (mcg)	Box-Cox *t*	Box-Cox *t*
Vitamina D (mcg)	Box-Cox Exponencial Potência	Box-Cox Exponencial Potência
Vitamina E (mg)	Box-Cox *t*	Box-Cox *t*
Vitamina K (mcg)	*skew t*	*skew t*
Vitamina C (mg)	Box-Cox *t*	Box-Cox *t*
Vitamina B1 (mg)	Box-Cox *t*	*skew t*
Vitamina B2 (mg)	Box-Cox *t*	Box-Cox *t*
Vitamina B3 (mg)	Box-Cox *t*	Log normal
Vitamina B6 (mg)	Box-Cox *t*	Gama
Vitamina B12 (mcg)	Box-Cox *t*	Box-Cox *t*
Ácido pantotênico (mg)	Box-Cox *t*	Box-Cox *t*
Folato (mcg)	Box-Cox *t*	Box-Cox *t*
Cálcio (mg)	Box-Cox *t*	Box-Cox Exponencial Potência
Fósforo (mg)	Box-Cox *t*	Box-Cox *t*
Magnésio (mg)	*skew t*	Box-Cox *t*
Ferro (mg)	Box-Cox *t*	Box-Cox *t*
Zinco (mg)	Box-Cox *t*	Box-Cox Cole-Green
Cobre (mg)	*skew t*	Box-Cox *t*
Selênio (mcg)	Box-Cox *t*	Box-Cox *t*
Sódio (mg)	Box-Cox *t*	Box-Cox *t*
Potássio (mg)	*skew t*	Box-Cox *t*
Manganês (mg)	Box-Cox *t*	*skew t*

Para o cálculo das prevalências de inadequação alimentar, utilizou-se o método NCI e os modelos Box-Cox *t* e Box-Cox Cole-Green com intercepto aleatório, descritos no Capítulo 2. Para a modelagem, considerou-se o consumo dos micronutrientes como variável resposta e não houve inclusão de covariáveis, assim os modelos foram compostos apenas por um valor médio ou mediano geral e o intercepto aleatório. Para que uma comparação com os dados brutos pudesse ser feita, as prevalências empíricas das médias e medianas individuais [1] foram calculadas. A prevalência empírica é a proporção do consumo médio e mediano dos indivíduos que estão abaixo do ponto de corte estabelecido pela EAR.

É importante realçar ainda que, utilizando a rotina `gamlss` do *software R*, um teste foi realizado para verificar se havia necessidade da inclusão de um efeito aleatório adicional que pudesse melhor explicitar a variabilidade intrapessoal. Este efeito aleatório estaria relacionado com o parâmetro de dispersão da Box-Cox *t* (coeficiente de variação, σ). O resultado do teste foi que o efeito não era necessário.

[1] A média e a mediana do consumo de cada indivíduo foi calculada considerando os três recordatórios.

A Tabela 3.3 mostra os critérios de Akaike para o método NCI, e os modelos Box-Cox *t* (BCT) e Box-Cox Cole-Green (BCCG) todos com efeito aleatório normal. No consumo dos homens, os modelos medianos apresentaram os menores valores do critério de Akaike para todos os nutrientes, sendo o modelo BCT o mais frequente, apenas para a vitamina D e selênio o critério privilegia ligeiramente o modelo BCCG. No consumo das mulheres, o modelo BCT novamente foi o que apresentou o menor valor na maioria dos casos, e o modelo padrão apresentou menor AIC para 6 micronutrientes entre os 22 estudados, a saber vitaminas D, B3 e B6, cálcio, fósforo e zinco.

Tabela 3.3: Critérios de Akaike para os modelos ajustados para o cálculo da estimativa de prevalência de inadequação alimentar.

Micronutriente (m)	Homens			Mulheres		
	NCI	BCCG	BCT	NCI	BCCG	BCT
Vitamina A (mcg)	5641,90	5633,84	5568,76	9403,40	9503,87	9388,39
Vitamina D (mcg)	1637,30	1623,18	1625,60	2630,90	2651,28	2647,79
Vitamina E (mg)	1861,90	1835,63	1812,65	2879,90	2879,87	2870,92
Vitamina K (mcg)	4244,20	4230,81	4217,56	6951,60	6926,84	6914,30
Vitamina C (mg)	4011,20	4014,59	3957,72	6535,30	6525,54	6462,68
Vitamina B1 (mg)	783,70	761,64	760,82	1062,40	1061,02	1051,73
Vitamina B2 (mg)	719,70	718,33	712,06	1042,20	1042,78	1031,16
Vitamina B3 (mg)	2747,50	2749,21	2727,70	4330,40	4330,68	4332,68
Vitamina B6 (mg)	967,00	957,92	942,44	1310,40	1314,42	1316,42
Vitamina B12 (mcg)	1871,00	1858,63	1805,91	3209,00	3173,66	3032,52
Ácido pantotênico (mg)	1504,80	1499,26	1475,05	2282,70	2282,30	2267,06
Folato (mcg)	4732,50	4713,53	4682,75	7597,80	7592,62	7552,44
Cálcio (mg)	5141,20	5141,79	5139,70	8420,10	8428,97	8430,42
Fósforo (mg)	5428,80	5426,64	5414,93	8857,90	8862,24	8862,71
Magnésio (mg)	4424,30	4389,04	4376,73	7076,00	7069,71	7061,83
Ferro (mg)	2415,60	2402,23	2379,38	3808,10	3782,59	3771,09
Zinco (mg)	2247,40	2247,08	2241,42	3458,30	3461,80	3463,22
Cobre (mg)	586,40	541,07	532,37	931,00	871,34	861,89
Selênio (mcg)	3925,70	3925,05	3925,95	6481,80	6446,21	6420,50
Sódio (mg)	6576,40	6448,58	6417,07	10564,00	10427,98	10418,17
Potássio (mg)	6102,80	6048,25	6032,53	9824,90	9827,47	9817,50
Manganês (mg)	1585,60	1474,66	1435,00	2982,00	2574,49	2547,28

A Tabela 3.4 mostra os valores de recomendação da EAR e as estimativas de prevalências de inadequação calculadas pela forma empírica, pelo método NCI, e pelo modelo BCT e seu caso limite, dado pela distribuição BCCG, para homens e mulheres, respectivamente.

Tabela 3.4: EAR (em unidade de massa/dia) e estimativas de prevalência de inadequação alimentar (em porcentagem).

Micronutriente (m)	Homens						Mulheres					
	EAR	Empíricas		NCI	BCCG	BCT	EAR	Empíricas		NCI	BCCG	BCT
		Médias	Medianas					Médias	Medianas			
Vitamina A (mcg)	625	39,7	45,6	18,1	43,1	46,0	500	26,3	33,2	9,2	26,8	27,4
Vitamina D (mcg)	10	96,3	95,6	98,5	97,0	97,4	10	98,7	97,8	99,6	99,3	99,3
Vitamina E (mg)	12	94,1	97,1	97,4	98,9	99,7	12	96,6	96,6	99,7	99,9	99,8
Vitamina K (mcg)	120	47,8	57,4	39,8	67,4	66,4	90	38,8	43,1	20,6	48,9	44,8
Vitamina C (mg)	75	51,5	59,6	35,3	59,9	65,0	60	43,1	52,6	28,5	52,7	57,4
Vitamina B1 (mg)	1,0	6,6	8,1	2,1	7,9	6,8	0,9	9,5	11,6	0,5	2,5	2,0
Vitamina B2 (mg)	1,1	12,5	19,9	8,1	11,1	11,4	0,9	9,9	11,2	2,5	5,4	3,7
Vitamina B3 (mg)	12	6,6	8,1	0,6	2,4	0,2	11	10,3	14,2	0,6	2,0	2,2
Vitamina B6 (mg)	1,4	19,9	22,8	9,8	26,1	19,9	1,3	28,0	34,1	16,5	25,0	24,8
Vitamina B12 (mcg)	2	11,0	18,4	1,0	13,8	11,4	2	16,4	23,3	0,0	36,4	14,3
Ácido Pantotênico (mg)	5	50,7	50,7	41,6	59,3	53,8	5	64,2	66,4	64,3	72,6	72,5
Folato (mcg)	320	21,3	22,1	17,0	29,8	23,0	320	41,8	44,8	29,7	47,6	43,4
Cálcio (mg)	1000	80,9	81,6	81,0	85,7	85,2	1000	85,8	87,9	91,5	92,4	92,6
Fósforo (mg)	580	2,9	2,2	0,5	1,0	1,2	580	6,5	8,2	1,0	1,5	1,5
Magnésio (mg)	350	78,7	80,9	74,0	83,3	81,8	265	65,5	69,8	66,8	77,1	76,2
Ferro (mg)	6	1,5	2,2	0,3	1,5	0,3	5	1,3	2,2	0,0	0,0	0,0
Zinco (mg)	9,4	22,1	25,7	13,3	21,4	19,4	6,8	14,7	18,5	4,0	7,7	7,8
Cobre (mg)	0,7	9,6	11,0	3,3	9,7	7,5	0,7	13,8	17,2	3,6	13,5	9,8
Selênio (mcg)	45	1,5	0,7	0,0	0,0	0,0	45	1,7	2,6	0,0	0,0	0,0
Sódio (mg)	1300	2,2	2,2	0,0	0,8	0,0	1300	2,2	5,2	0,1	0,8	0,4
Potássio (mg)	4700	95,6	96,3	94,1	95,3	97,7	4700	100,0	99,1	99,9	100,0	99,9
Manganês (mg)	1,8	9,6	14,0	6,7	13,9	8,1	1,8	15,1	18,1	12,3	11,1	9,2

A partir da Tabela 3.4, pode-se fazer uma comparação entre as prevalências estimadas pelos métodos NCI e BCT e BCCG com efeito aleatório e o comportamento dos dados. Desse modo, nota-se que os nutrientes que possuem estimativas de prevalência de inadequação próximas de zero ou um foram estimadas de forma análoga pelo procedimento padrão (NCI) e pelos métodos medianos propostos (vide estimativas referentes aos nutrientes vitaminas D, E e B3, ferro, selênio, sódio e potássio, tanto para homens quanto para mulheres). Já nas estimativas para os nutrientes ácido pantotênico, cálcio e magnésio para ambos os sexos e vitaminas C e K para homens, os modelos medianos tenderam a produzir estimativas superiores às encontradas nos dados brutos. Para os demais nutrientes, as estimativas encontradas pelos modelos medianos apresentaram-se mais plausíveis com o observado nos dados.

É importante destacar que para o ajuste do modelo BCT aos nutrientes vitaminas A e K, ferro, zinco e potássio no consumo para homens, o parâmetro referente ao número de graus de liberdade foi prefixado antes da estimação dos demais parâmetros do modelo. Para o ajuste do modelo BCCG para vitamina A em mulheres, o parâmetro referente à transformação dos dados também foi fixado anteriormente aos outros. A fixação desses parâmetros foi necessária devido a não convergência no processo de estimação ou por produzir estimativas não realistas. Para a fixação dos valores do parâmetro referente ao número de graus de liberdade no modelo BCT, a rotina `gamlss` foi usada através do modelo Box-Cox *t* com efeito aleatório normal, já para o modelo BCCG, o valor da transformação dado no método NCI foi o utilizado. Para compreender melhor as questões evidenciadas na aplicação, um estudo de simulação foi conduzido e os detalhes são apresentados no Capítulo 4.

A fim de explicitar uma análise mais detalhada, escolheu-se o consumo da vitamina C para homens. A escolha da distribuição desse nutriente é para ilustrar um caso de alta assimetria e pontos discrepantes, conforme pode-se notar nas medidas descritivas presentes na Tabela 3.1 e Figura 3.1. Além disso, as estimativas de prevalência de inadequação alimentar foram bem distintas nos diversos métodos aplicados: segundo a Tabela 3.4, a estimativa da prevalência de inadequação alimentar pelo método padrão (NCI) é de $35,3\%$, bem diferente da estimativa obtida pela distribuição empírica dos dados, $51,5\%$.

A Figura 3.2 descreve o perfil do consumo de vitamina C relatado pelos idosos do sexo masculino em três R24h. Observa-se que existe uma variabilidade notável nos registros individuais de alguns indivíduos. A variabilidade entre os indivíduos não parece alta.

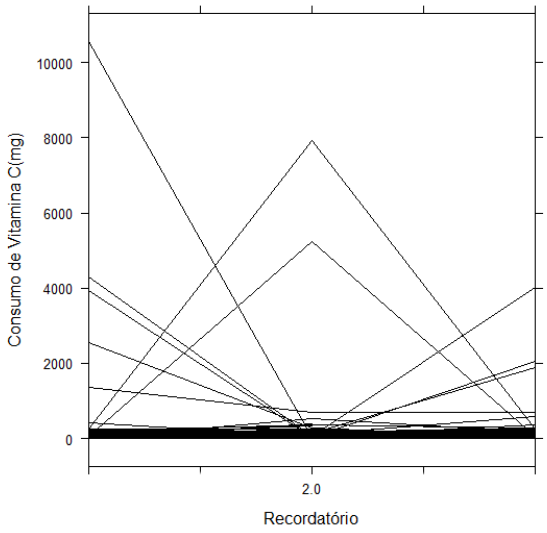


Figura 3.2: Análise de perfil para consumo de vitamina C em homens.

Na Tabela 3.5 são apresentadas as estimativas dos parâmetros do modelo Box-Cox t com efeito aleatório encontradas pelo método de máxima verossimilhança marginal aproximada. O modelo adotado é

$$\begin{aligned} \log(\mu_i) &= \beta_0 + \gamma_i, \\ \log(\sigma) &= \delta, \\ \nu &= \eta, \\ \log(\tau) &= \zeta. \end{aligned}$$

Nota-se que a estimativa do parâmetro relativo ao consumo mediano do grupo (β_0) apresentou pouca variação nas distintas abordagens utilizadas. As estimativas dos parâmetros referentes ao modelo Box-Cox t com efeito aleatório, quando a distribuição deste efeito é normal, obtidas pela rotina `gamlss` e pela programada, foram semelhantes. Quando a suposição sobre a distribuição do efeito aleatório mudou de normal para t, as estimativas dos parâmetros referentes à transformação potência (η) e ao número de graus da liberdade (ζ) da distribuição Box-Cox t variaram pouco e o parâmetro associado ao coeficiente de variação (δ) sofreu maiores modificações. Todavia, para as distintas abordagens, as prevalências de inadequação alimentar (ρ) estimadas pelo modelo Box-Cox t com efeito aleatório normal e t, foram mais próxima da estimativa encontrada na distribuição empírica dos dados (59, 6%), quando comparada ao modelo padrão (vide Tabela 3.4). Os valores do critério de Akaike foram semelhantes, sendo o menor obtido no modelo Box-Cox t com efeito aleatório seguindo uma distribuição t.

Tabela 3.5: Estimativas dos parâmetros dos modelos Box-Cox *t* com efeito aleatório; consumo de vitamina C para homens.

Efeito aleatório	Normal *gamlss*		Normal		*t*	
Parâmetros	Estimativa	EP	Estimativa	EP	Estimativa	EP
β_0	4,006	0,056	4,007	0,093	4,071	0,091
δ	$-0,210$	0,064	$-0,218$	0,149	$-0,306$	0,150
η	0,276	0,050	0,275	0,069	0,310	0,087
ζ	0,473	0,097	0,462	0,203	0,420	0,203
λ	0,710	0,055	0,713	0,144	0,460	0,777
κ					1,806	0,508
$\rho(\%)$	64,5		65,0		64,7	
AIC	3957,7		3957,7		3933,76	

A Tabela 3.6 mostra os percentis de 5%, 10%, 25%, 50%, 75%, 90% e 95% da distribuição de consumo usual populacional estimado pelos método NCI e modelos Box-Cox *t* com efeito aleatório normal e *t*, respectivamente. Observa-se que os percentis da distribuição usual de consumo estimados pelo modelo baseado na média (NCI) apresentam valores maiores em relação aos percentis estimados pelos modelos medianos (BCT com efeito aleatório normal e *t*). Nota-se ainda que no modelo BCT, a mudança do efeito aleatório normal para *t* proporcionou uma pequena alteração nos percentis estimados das distribuições de consumo mediano, sendo esta diferença mais acentuada no percentil de 95%, para o qual o modelo com efeito aleatório *t* produziu estimativas de consumos medianos maiores do que o observado no modelo com efeito aleatório normal.

Tabela 3.6: Percentis das distribuições de consumo usual estimados pelo método NCI e modelos BCT com efeito aleatório normal e *t*.

Percentis (%)	5	10	25	50	75	90	95
NCI (mg)	23,38	33,13	57,55	104,86	180,23	294,00	381,92
BCT (normal)(mg)	17,49	22,76	35,21	57,34	93,12	141,66	182,68
BCT (t) (mg)	13,97	23,85	41,45	61,24	90,76	151,77	258,79

A Figura 3.3 apresenta o gráfico de perfil do logaritmo da função de verossimilhança [2], com a finalidade de estudar o comportamento de tal função no caso do modelo Box-Cox *t*, com efeito aleatório que segue uma distribuição *t*. Para tal, os parâmetros referentes aos graus de liberdade da distribuição Box-Cox *t* e da distribuição *t* foram fixados. A Figura baseia-se num *grid* na amplitude de 1 a 15, variando a cada 1 unidade para o número de graus de liberdade de cada distribuição *t* proposta e, em seguida, uma nova figura é apresentada, a qual fundamenta-se num *grid* mais específico em torno do ponto encontrado na primeira figura. Observou-se que o valor que minimiza

[2]O Apêndice A.1 apresenta um resumo do método da máxima verossimilhança perfilada.

o critério de Akaike (*AIC*) encontra-se próximo ao ponto (2,2) da malha traçada. O gráfico de contorno da superfície de verossimilhança vista de cima também foi traçado para melhor explicitar o ponto de máximo encontrado.

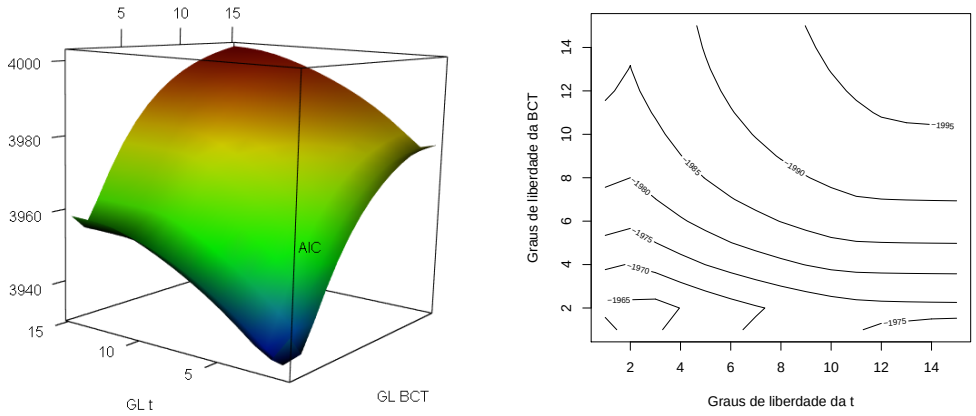


Figura 3.3: À esquerda, critério de Akaike (AIC) baseado no perfil do logaritmo da verossimilhança e, à direita, gráfico de contorno do modelo Box-Cox t com efeito aleatório quando este efeito segue uma distribuição t.

A Figura 3.4 apresenta o *grid* mais específico feito em torno da malha de dimensões $[1,3]$ por $[1,3]$, variando numa sequência a cada 0,2. Assim, os números de graus de liberdade associados ao ponto de máximo da função do modelo Box-Cox t com efeito aleatório que segue a distribuição t foram $1,6$ e $1,8$, para Box-Cox t e t, respectivamente, o que corrobora os resultados estimados pelo método da máxima verossimilhança marginal aproximada, estimados em $\widehat{\tau} = \exp(\widehat{\zeta}) = \exp(0,420) \approx 1,587$ e $\widehat{\kappa} = 1,806$ para Box-Cox t e t, respectivamente. O gráfico de contorno é novamente apresentado. Portanto, o valor encontrado é o ponto de máximo da função de verossimilhança.

É importante ressaltar que um estudo do perfil da verossimilhança para o modelo Box-Cox t quando o efeito aleatório segue uma distribuição normal também foi realizado, e os resultados tiveram um comportamento análogo ao explicitado no caso do efeito aleatório t, sendo por este motivo omitidos.

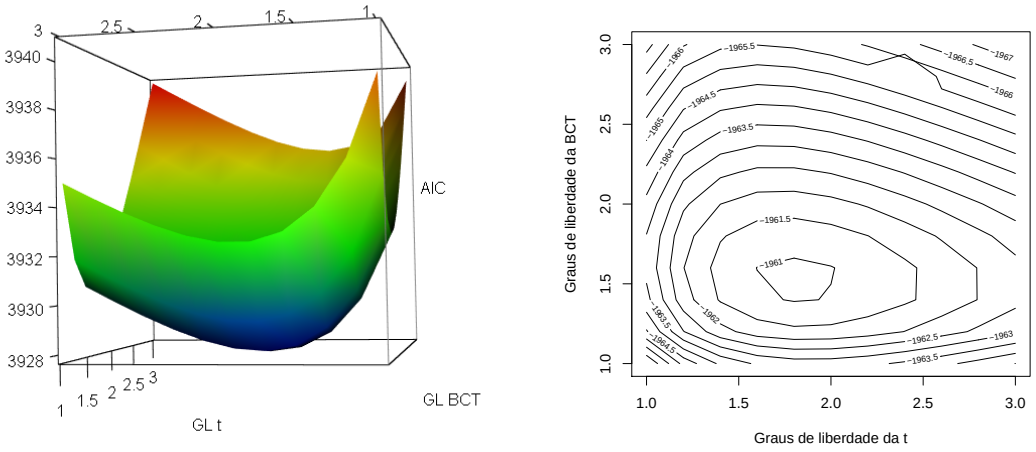


Figura 3.4: À esquerda, critério de Akaike (AIC) baseado no perfil do logaritmo da verossimilhança e, à direita, gráfico de contorno do modelo Box-Cox t com efeito aleatório quando este efeito segue uma distribuição t.

As Figuras 3.5 e 3.6 apresentam gráficos de resíduos quantílicos para os modelos BCT com efeito aleatório normal e t. Cada figura apresenta quatro gráficos, os quais mostram: os valores preditos versus os resíduos quantílicos, o índice e a densidade dos resíduos quantílicos e a densidade da distribuição $N(0,1)$ (valores teóricos) versus os resíduos quantílicos (valores amostrais).

Nos gráficos referentes aos resíduos quantílicos versus valores preditos, notam-se valores preditos extremos maiores no modelo Box-Cox t quando o efeito aleatório segue uma distribuição t. Destacam-se dois pontos que apresentam um valor predito maior em relação aos demais, os quais correspondem às observações de números 340 e 341, referentes a recordatórios de um mesmo indivíduo, cujo consumo mediano amostral é de $691,75$ mg. O modelo BCT com efeito aleatório t apresenta um valor predito mais próximo do valor mediano amostral, dado por $612,17$ mg, enquanto o modelo BCT com efeito aleatório normal apresenta um valor predito de $293,52$ mg.

Os gráficos de índice e densidade apresentaram um bom comportamento. O gráfico do resíduos quantílicos versus quantil teórico (densidade $N(0,1)$) também mereceu destaque; nota-se na Figura 3.5 que tal gráfico apresenta uma pequena fuga de pontos nas caudas da distribuição, já na Figura 3.6, observa-se que os resíduos tiveram um comportamento ligeiramente melhor em relação à normalidade. Desse modo, os resíduos quantílicos referentes ao modelo Box-Cox t quando o efeito aleatório segue uma distribuição t, resultaram um comportamento um pouco melhor do que o modelo no qual o efeito aleatório é normalmente distribuído.

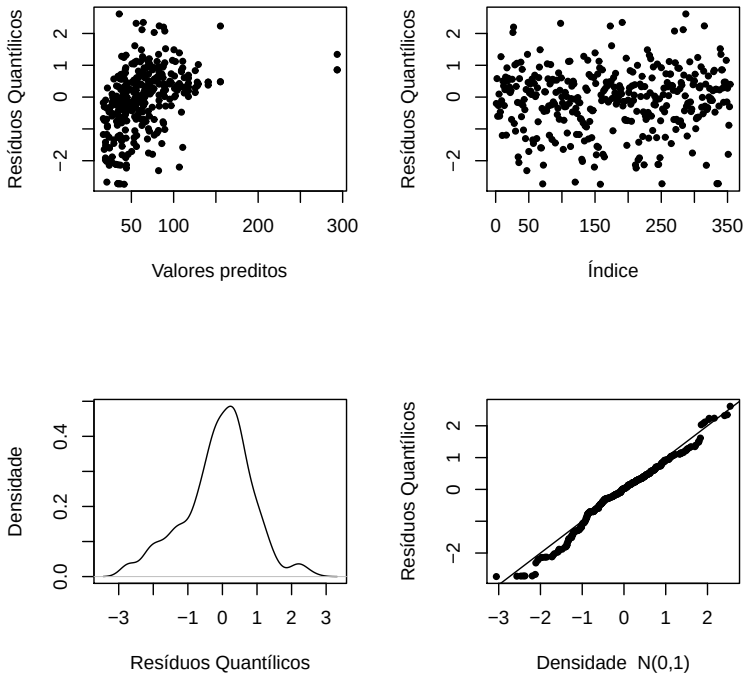


Figura 3.5: Resíduos quantílicos para o modelo Box-Cox *t* quando o efeito aleatório segue uma distribuição normal.

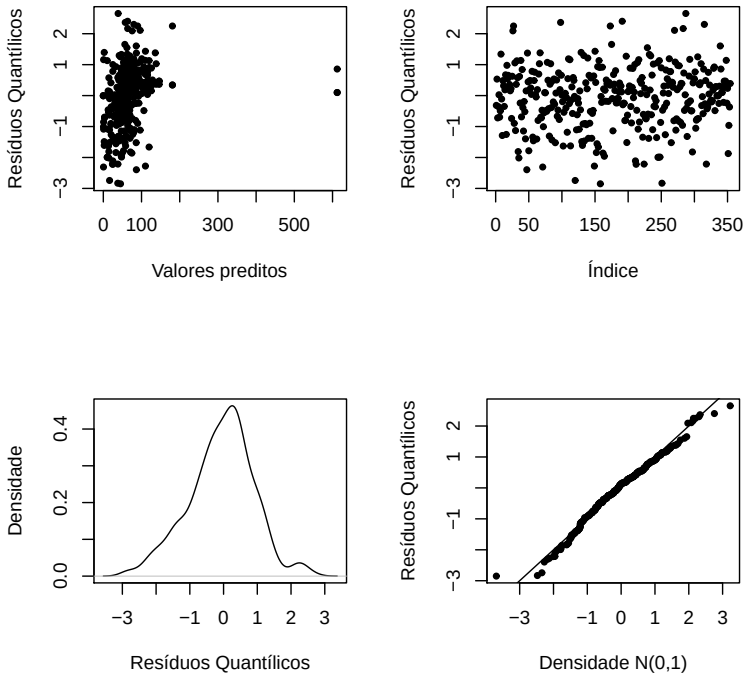


Figura 3.6: Resíduos quantílicos para o modelo Box-Cox *t* quando o efeito aleatório segue uma distribuição *t*.

3.3 Modelo Box-Cox *t* misto

Com o objetivo de verificar se a inclusão de covariáveis melhora a precisão para estimar a prevalência de inadequação alimentar, um estudo foi feito utilizando como variável resposta o consumo de vitamina C.

As covariáveis testadas foram: estado civil (não casado, casado), escolaridade (até primeiro grau, acima do primeiro grau), hipertensão (sim, não), diabetes (sim, não), atividades de vida diária (AVD) (tarefas que a pessoa precisa realizar para cuidar de si, tais como tomar banho, vestir-se etc; os indivíduos foram classificados como independentes, dependentes, e foi estabelecida uma categoria para os dados faltantes), atividades instrumentais de vida diária (AIVD) (habilidades para administrar o ambiente em que vive e inclui as seguintes ações: preparar refeições, fazer tarefas domésticas etc; indivíduos foram classificados como na covariável AVD) (Marra et al., 2007).

A Tabela 3.7 apresenta as frequências das covariáveis categóricas utilizadas. A única covariável contínua do banco de dados escolhida foi idade. A seguir, as medidas resumo para homens e mulheres, respectivamente: Média - 71,20/71,79 anos; DP - 7,11/7,21 anos; Mínimo - 60/52 anos; Máximo - 90/92 anos e Mediana - 71/72 anos.

Tabela 3.7: Distribuição de homens e mulheres segundo as covariáveis categóricas.

Covariável	Categoria	Homens (%)	Mulheres(%)
Estado Civil	Não casado	18,1	54,0
	Casado	81,9	46,0
Escolaridade	Até 1º grau	66,1	81,6
	Acima 1º grau	33,9	18,4
Hipertensão	Sim	51,7	58,9
	Não	48,3	41,1
Diabetes	Sim	29,1	27,8
	Não	70,9	72,2
AVD	Independência	86,7	80,3
	Dependência	6,8	10,3
	Faltante	6,5	9,4
AIVD	Independência	70,0	59,1
	Dependência	23,5	31,5
	Faltante	6,5	9,4

Cada uma destas covariáveis foram testadas individualmente para determinar a ordem a serem incluídas no modelo, sendo usado o procedimento *forward*.

A Tabela 3.8 apresenta as estimativas de prevalência de inadequação alimentar obtidas pelo modelo Box-Cox *t* com efeito aleatório normal e método NCI, sem e com covariáveis significantes.

Tabela 3.8: Prevalência de inadequação alimentar (dada em porcentagem) para o consumo de vitamina C, calculadas a partir do modelo Box-Cox *t* com efeito aleatório normal e método NCI, sem e com covariáveis significantes.

Modelo	Homens		Mulheres	
	BCT	NCI	BCT	NCI
Sem covariáveis	65,00	35,30	56,78	28,75
Com covariável escolaridade	66,51		57,79	29,21
Com covariável AIVD	65,01	37,80	57,76	28,46
Com covariáveis escolaridade e AIVD			58,03	28,78
Com covariáveis escolaridade e idade		37,38		
Com covariáveis escolaridade, AIVD e idade		37,44		

As estimativas das prevalências de inadequação alimentar não sofreram grandes alterações na ausência e presença das covariáveis, isso significa dizer que embora algumas tenham sido significantes, as mesmas não se apresentaram relevantes para determinar uma alteração significativa na estimação da prevalência.

Para o conjunto dos homens, o modelo BCT que apresentou o menor critério de Akaike foi o associado à covariável AIVD. O modelo é dado por

$$\begin{aligned} \log(\mu_i) &= \boldsymbol{X}_i\boldsymbol{\beta} + \gamma_i, \\ \log(\sigma) &= \delta, \\ \nu &= \eta, \\ \log(\tau) &= \zeta, \end{aligned} \tag{3.1}$$

em que $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)$ sendo β_0 o intercepto, e β_1 e β_2 sendo duas variáveis *dummies* criadas para compor a covariável AIVD e $\boldsymbol{X}_i = (1, X_{i1}, X_{i2})$ a i-ésima linha da matriz de covariáveis associadas ao indivíduo i, os demais componentes de (3.1) são os mesmos já explicitados em (2.2).

Para o modelo em questão, as estimativas dos parâmetros relacionados ao coeficiente de variação (σ), à transformação (ν) e ao número de graus de liberdade (τ) não sofreram muita alteração em relação ao modelo sem covariáveis, tendo resultados bem próximos aos relatados na Tabela 3.5. As estimativas dos parâmetros associados à mediana e seus erros padrões foram dados, respectivamente, por $\widehat{\beta_0} = 4,11(0,10)$, $\widehat{\beta_1} = -0,59(0,21)$ e $\widehat{\beta_2} = 0,34(0,31)$. Para a covariável AIVD, a categoria de base utilizada foi o fato dos idosos serem independentes, o coeficiente β_1 refere-se à categoria de dependência e β_2, à de observações faltantes. Logo, idosos independentes de outras pessoas para executarem suas atividades fora de casa apresentaram quase o dobro do consumo mediano de vitamina C quando comparados àqueles que executam tais tarefas de forma dependente ($\exp(\beta_1) = 0,55$).

Para as mulheres, o modelo que apresentou o melhor ajuste foi o composto pelas

covariáveis escolaridade e AIVD. Desse modo, com a mesma estrutura explicitada em (3.1), o vetor de parâmetros foi composto por $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \beta_3)$, em que β_0, β_1, β_2 e β_3 são associados, respectivamente, ao intercepto, à covariável escolaridade e à covariável AIVD e a linha i referente a matriz de delineamento com as característica do indivíduo i, é dada por $\boldsymbol{X}_i = (1, X_{i1}, X_{i2}, X_{i3})$. As estimativas (com erros padrões) finais foram: $\widehat{\beta_0} = 3,95(0,08)$; $\widehat{\beta_1} = 0,44(0,14)$; $\widehat{\beta_2} = -0,28(0,12)$; $\widehat{\beta_3} = 0,14(0,18)$; $\widehat{\delta} = -0,43(0,08)$; $\widehat{\eta} = 0,39(0,24)$; $\widehat{\zeta} = 0,75(0,34)$ e $\widehat{\lambda^2} = 0,31(0,01)$. Portanto, para o conjunto das mulheres, a mediana de consumo de vitamina C foi 55% maior naquelas que apresentaram um nível de escolaridade acima do 1º grau, em relação àquelas que estão classificadas como tendo escolaridade até o 1ºgrau ($\exp(\beta_1) = 1,55$), considerando as demais covariáveis fixas. Quanto à covariável AIVD, analogamente aos homens, mulheres idosas que apresentaram independência de outras pessoas para executarem atividades fora de casa, tiveram um consumo mediano de 24% a mais de vitamina C quando comparadas àquelas que não conseguem executar tais tarefas sozinhas ($\exp(\beta_2) = 0,76$), considerando as demais covariáveis fixadas.

CAPÍTULO 4

Simulação

4.1 Estrutura geral

O objetivo do estudo de simulação é verificar se a metodologia baseada na mediana de consumo alimentar é adequada para estimar os percentis da distribuição usual de consumo populacional, e compará-la com o método padrão (Tooze et al., 2010). Este estudo é fundamental para estimar de forma adequada a prevalência de inadequação alimentar, uma vez que esta medida é determinada a partir da distribuição estimada do consumo usual populacional, conforme detalhado no Capítulo 2.

Para tal objetivo, percentis das chamadas distribuições "verdadeiras" das médias e medianas foram gerados, calculados a partir de 500 amostras. Cada amostra foi composta por 500 indivíduos simulados e para cada indivíduo foram gerados 365 recordatórios 24 horas. Assim, tem-se 365 dias de observação do consumo alimentar simulado de cada indivíduo, ou seja, o seu consumo anual, através do qual pode-se calcular o seu "verdadeiro" consumo médio e mediano. Desse modo, para cada pessoa foi calculado o seu consumo médio e mediano e, em cada amostra, foram encontrados os percentis das distribuições empíricas das médias e medianas de consumo. Por fim, as médias de cada um dos percentis de 5%, 10%, 25%, 50%, 75%, 90% e 95% das 500 amostras foram calculadas, e estas compuseram os percentis das chamadas distribuições "verdadeiras" das médias e medianas populacionais.

Modelos sem covariáveis com intercepto aleatório normal foram ajustados, baseados nas distribuições normal, Box-Cox Cole-Green (BCCG) e Box-Cox t (BCT), esta última com o parâmetro referente ao número de graus de liberdade estimado e fixado. Para o modelo normal, o método NCI descrito na Seção 2.1 foi aplicado, a partir do

qual a distribuição de consumo usual é calculada baseada na média; os demais modelos baseiam-se no consumo mediano, através do modelo Box-Cox t com efeito aleatório e seu caso limite, dado pela distribuição Box-Cox Cole-Green, descrito na Seção 2.2. A partir das estimativas dos parâmetros dos modelos ajustados, percentis da distribuição de consumo baseado na média (método NCI) e medianas (modelos BCT e BCCG) foram simulados em cada amostra. Em seguida, as médias de cada um dos percentis de 5%, 10%, 25%, 50%, 75%, 90% e 95% de 500 amostras foram calculadas para cada modelo, e estas formaram os percentis estimados das distribuições usuais de cada um dos modelos supracitados.

Os modelos baseados nas distribuições BCCG e BCT requerem estimativas iniciais para o processo de otimização. Para tal, as estimativas foram encontradas através dos modelos disponíveis no pacote `gamlss` sem efeito aleatório. A estimativa inicial para o parâmetro associado à variância do efeito aleatório foi calculada a partir de um modelo linear com intercepto aleatório, porém, se esta variância era muito grande, um valor inicial arbitrário lhe foi atribuído.

Os dados foram simulados conforme segue. Foram geradas $R = 500$ amostras, cada uma composta por 500 indivíduos com 3 recordatórios 24 horas cada um. Desse modo, o recordatório 24 horas é a variável resposta. Esta foi gerada aleatoriamente a partir das distribuições Box-Cox t, Box-Cox Cole-Green, "normal transformada" e gama. Sob a distribuição "normal transformada" os dados são gerados segundo uma distribuição normal e uma transformação inversa Box-Cox é aplicada. Nota-se que o método NCI baseia-se na suposição de que os dados seguem essa distribuição. Consideram-se ainda cenários em que os dados são contaminados por algumas observações discrepantes. Em todos os cenários, o efeito aleatório foi simulado a partir de uma distribuição normal.

O viés relativo em porcentagem e a raiz quadrada do erro quadrático médio relativo foram calculados. Assim, seja ψ o verdadeiro valor do percentil de interesse da distribuição populacional e $\widehat{\psi}_m$ sua estimativa na amostra m. Tem-se que

$$\begin{aligned}
\overline{\psi} &= \frac{1}{R}\sum_{m=1}^{R}\widehat{\psi}_m, \\
\text{Vies}(\widehat{\psi}) &= \left(\frac{\overline{\psi}-\psi}{\psi}\right)\times 100, \\
\sqrt{EQMR(\widehat{\psi})} &= \sqrt{\frac{1}{R}\sum_{m=1}^{R}\left(\frac{\widehat{\psi}_m-\psi}{\psi}\right)^2},
\end{aligned}$$

são, respectivamente, as estimativas da média, do viés relativo em porcentagem e da raiz do erro quadrático médio relativo dos percentis estimados.

4.2 Cenário 1: Dados gerados a partir das distribuições BCT e BCCG

Neste cenário os dados são gerados sob o seguinte modelo:

$$\begin{aligned} Y_{ij}|\gamma_i &\overset{ind}{\sim} BCT(\mu_i, \sigma, \nu, \tau), \\ \log(\mu_i) &= \beta_0 + \gamma_i, \\ \gamma_i &\overset{ind}{\sim} N(0, \lambda^2), \end{aligned} \tag{4.1}$$

$i = 1, \ldots, 500$, $j = 1, 2, 3$. Adicionalmente, considera-se o modelo Box-Cox Cole-Green, em que (4.1) é substituído por

$$Y_{ij}|\gamma_i \overset{ind}{\sim} BCCG(\mu_i, \sigma, \nu).$$

Os valores dos parâmetros são fixados em $\beta_0 = 5$, $\sigma = 0,5$, $\nu = 0,5$, $\tau = 4$ e $\lambda^2 = 0,5$.

A Tabela 4.1 mostra os percentis medianos e médios, os vieses relativos em porcentagem, a raiz quadrada do erro quadrático médio relativo calculados a partir dos modelos Box-Cox *t* e Box-Cox Cole-Green com intercepto aleatório e método NCI, para os dados simulados seguindo a distribuição Box-Cox *t*. Nota-se que os menores valores de vieses relativos foram encontrados nos modelos Box-Cox *t* com τ estimado e com τ fixado nos valores 4, 3 e 2, sendo os vieses relativos menores ou iguais a 1% em todos os percentis. Para o modelo BCT com $\tau = 1,5$, os vieses relativos foram maiores, o que era de se esperar pela natureza dos dados gerados. O modelo BCCG apresentou vieses relativos maiores nos percentis de ordem superior. O modelo padrão NCI apresentou vieses relativos maiores em quase todos os percentis quando comparados aos modelos baseados na mediana de consumo. Nota-se ainda que o método NCI subestima consideravelmente os percentis de ordem elevada. Por outro lado, os modelos BCT mantêm os vieses pequenos para todos os percentis.

É importante ressaltar que no modelo BCT, quando o parâmetro relativo ao número de graus de liberdade foi estimado simultaneamente com os demais, a estimativa média deste parâmetro foi de 4,55, próxima do valor utilizado para geração dos dados.

A porcentagem de amostras descartadas por falta de convergência na estimação dos parâmetros foi baixa: no modelo BCCG, 0,2%, no BCT com τ estimado, 6,2%, no BCT com $\tau = 1,5$ fixo, 0,8%; nos demais cenários não houve amostras desconsideradas.

Tabela 4.1: Cenário 1: Percentis medianos e médios, vieses relativos em porcentagem e raiz quadrada do erro quadrático médio relativo, dados gerados a partir da distribuição BCT com $\tau = 4$.

Percentil		5	10	25	50	75	90	95
Verdadeiro (mediano)		46,58	60,30	92,79	149,32	240,65	370,94	480,62
Modelo								
BCCG	$\overline{\widehat{\psi}}$	46,71	60,77	94,38	153,91	251,17	390,20	508,20
	Viés $(\widehat{\psi})$	0,28	0,77	1,71	3,07	4,37	5,19	5,74
	$\sqrt{EQMR(\widehat{\psi})}$	0,06	0,05	0,04	0,04	0,06	0,08	0,09
BCT	$\overline{\widehat{\psi}}$	46,86	60,54	92,91	149,54	240,78	369,82	478,28
	Viés $(\widehat{\psi})$	0,60	0,40	0,12	0,15	0,05	$-0,30$	$-0,49$
	$\sqrt{EQMR(\widehat{\psi})}$	0,06	0,05	0,04	0,03	0,04	0,05	0,05
BCT $\tau = 4$	$\overline{\widehat{\psi}}$	46,68	60,31	92,54	148,97	239,79	368,17	475,82
	Viés $(\widehat{\psi})$	0,22	0,02	$-0,27$	$-0,24$	$-0,36$	$-0,75$	$-1,00$
	$\sqrt{EQMR(\widehat{\psi})}$	0,06	0,05	0,04	0,04	0,04	0,05	0,06
BCT $\tau = 3$	$\overline{\widehat{\psi}}$	46,76	60,42	92,74	149,35	240,64	369,76	478,21
	Viés $(\widehat{\psi})$	0,38	0,20	$-0,06$	0,02	0,00	$-0,32$	$-0,50$
	$\sqrt{EQMR(\widehat{\psi})}$	0,06	0,05	0,04	0,03	0,04	0,05	0,06
BCT $\tau = 2$	$\overline{\widehat{\psi}}$	46,79	60,46	92,78	149,42	240,68	369,81	478,32
	Viés $(\widehat{\psi})$	0,45	0,26	$-0,01$	0,07	0,01	$-0,31$	$-0,48$
	$\sqrt{EQMR(\widehat{\psi})}$	0,06	0,05	0,04	0,03	0,04	0,05	0,06
BCT $\tau = 1,5$	$\overline{\widehat{\psi}}$	46,30	59,80	91,81	147,84	238,10	365,67	472,75
	Viés $(\widehat{\psi})$	$-0,61$	$-0,84$	$-1,06$	$-0,99$	$-1,06$	$-1,42$	$-1,64$
	$\sqrt{EQMR(\widehat{\psi})}$	0,06	0,05	0,04	0,04	0,04	0,05	0,06
Verdadeiro (médio)		52,06	67,31	103,72	167,26	269,80	415,04	536,42
NCI	$\overline{\widehat{\psi}}$	52,20	71,03	113,12	180,47	274,43	388,79	472,96
	Viés $(\widehat{\psi})$	0,27	5,52	9,06	7,90	1,71	$-6,33$	$-11,83$
	$\sqrt{EQMR(\widehat{\psi})}$	0,08	0,09	0,10	0,09	0,05	0,08	0,13

A Tabela 4.2 apresenta os resultados obtidos quando os dados simulados seguem uma distribuição Box-Cox Cole-Green. O modelo que apresentou o melhor ajuste em quase todos os percentis foi o BCCG, seguido pelo modelo BCT com τ estimado e com $\tau = 2$. É importante ressaltar que o ajuste para os demais modelos BCT foram muito bons, sendo os vieses relativos inferiores a 2% para todos os percentis. O modelo padrão (NCI) apresentou novamente vieses relativos altos quando comparados com os dos modelos medianos.

Tabela 4.2: Cenário 1: Percentis medianos e médios, vieses relativos em porcentagem e raiz quadrada do erro quadrático médio relativo, dados gerados a partir da distribuição BCCG.

Percentil		5	10	25	50	75	90	95
Verdadeiro (mediano)		46,20	59,81	92,16	148,31	238,51	366,33	472,67
Modelo								
BCCG	$\overline{\widehat{\psi}}$	46,55	60,13	92,20	148,32	238,73	366,42	473,68
	Viés $(\widehat{\psi})$	0,74	0,54	0,05	0,01	0,10	0,03	0,21
	$\sqrt{EQMR(\widehat{\psi})}$	0,06	0,05	0,04	0,04	0,04	0,05	0,06
BCT	$\overline{\widehat{\psi}}$	46,57	60,09	92,05	147,93	237,81	364,66	471,11
	Viés $(\widehat{\psi})$	0,79	0,48	−0,12	−0,25	−0,29	−0,46	−0,33
	$\sqrt{EQMR(\widehat{\psi})}$	0,06	0,05	0,04	0,03	0,04	0,05	0,05
BCT $\tau = 4$	$\overline{\widehat{\psi}}$	45,87	59,28	91,02	146,57	236,16	362,83	469,13
	Viés $(\widehat{\psi})$	−0,71	−0,87	−1,24	−1,17	−0,98	−0,96	−0,75
	$\sqrt{EQMR(\widehat{\psi})}$	0,06	0,05	0,04	0,04	0,04	0,05	0,06
BCT $\tau = 3$	$\overline{\widehat{\psi}}$	46,78	60,45	92,87	149,67	241,33	371,08	479,98
	Viés $(\widehat{\psi})$	1,24	1,08	0,77	0,92	1,18	1,30	1,55
	$\sqrt{EQMR(\widehat{\psi})}$	0,06	0,05	0,04	0,04	0,04	0,05	0,06
BCT $\tau = 2$	$\overline{\widehat{\psi}}$	46,52	60,10	92,24	148,41	239,00	367,12	474,43
	Viés $(\widehat{\psi})$	0,69	0,49	0,09	0,07	0,21	0,21	0,37
	$\sqrt{EQMR(\widehat{\psi})}$	0,06	0,05	0,04	0,03	0,04	0,05	0,05
BCT $\tau = 1,5$	$\overline{\widehat{\psi}}$	45,65	58,98	90,47	145,58	234,36	359,75	464,96
	Viés $(\widehat{\psi})$	−1,21	−1,38	−1,83	−1,84	−1,74	−1,80	−1,63
	$\sqrt{EQMR(\widehat{\psi})}$	0,05	0,05	0,04	0,04	0,04	0,05	0,06
Verdadeiro (médio)		49,18	63,59	97,91	157,84	254,64	390,95	505,34
NCI	$\overline{\widehat{\psi}}$	48,04	64,50	102,18	165,17	258,16	378,53	471,65
	Viés $(\widehat{\psi})$	−2,33	1,42	4,36	4,65	1,38	−3,18	−6,67
	$\sqrt{EQMR(\widehat{\psi})}$	0,07	0,06	0,06	0,06	0,04	0,06	0,09

No caso de dados gerados a partir de uma distribuição BCCG, tanto para o modelo BCCG quanto para o modelo BCT com τ estimado a porcentagem de amostras descartadas foi de $0,4\%$, e para os modelos BCT com $\tau = 4$ e $\tau = 3$, foram de $0,6\%$ e 1%, respectivamente. Para os demais ajustes não houve amostras descartadas.

Em suma, os modelos Box-Cox *t* e Box-Cox Cole-Green mostraram-se mais adequados à estimação do consumo usual populacional do que o método NCI.

4.3 Cenário 2: Dados gerados a partir da distribuição "normal transformada"

Neste cenário os dados são gerados sob o modelo assumido pelo método NCI, ou seja,

$$\begin{aligned} Y^*_{ij} &= \beta_0 + \gamma_i + \epsilon_{ij}, \\ \gamma_i &\overset{ind}{\sim} N(0, \sigma^2_\gamma), \\ \epsilon_{ij} &\overset{ind}{\sim} N(0, \sigma^2_\epsilon), \end{aligned}$$

com γ_i e ϵ_i independentes, para $i = 1, \ldots, 500$ e $j = 1, 2, 3$; após a geração dos dados, uma transformação inversa Box-Cox é aplicada

$$Y_{ij} = (\nu Y^*_{ij} + 1)^{\frac{1}{\nu}}.$$

Neste cenário foram atribuídos os seguintes valores para os parâmetros:

(I) $\beta_0 = 10$, $\sigma^2_\gamma = 0,5$, $\sigma^2_\epsilon = 1$ e $\nu = 0,3$ (assimetria à direita leve);

(II) $\beta_0 = 7,4056$, $\sigma^2_\gamma = 0,3897$, $\sigma^2_\epsilon = 1,2186$ e $\nu = 0,0371$ (valores estimados pelo método NCI para o ajuste do consumo de vitamina A em mulheres descrito no Capítulo 3; assimetria à direita acentuada);

(III) a fim de avaliar a influência no comportamento do método na presença de *outliers* extremos, no caso (II) foi considerado também uma situação de dados "contaminados", em que 5% dos valores de cada amostra simulada foram substituídos por dados gerados de uma outra distribuição "normal transformada" com um intercepto $\beta_0 = 9$ e demais parâmetros como em (II), resultando valores aberrantes em relação aos dados inicialmente gerados.

As Tabelas 4.3, 4.4 e 4.5 apresentam os percentis medianos e médios, os vieses relativos em porcentagem, a raiz quadrada do erro quadrático médio relativo calculados a partir dos modelos Box-Cox *t* e Box-Cox Cole-Green com intercepto aleatório e método NCI, para os dados simulados seguindo a distribuição "normal transformada".

Tabela 4.3: Cenário 2: Percentis medianos e médios, vieses relativos em porcentagem e raiz quadrada do erro quadrático médio relativo. Dados gerados a partir de uma distribuição "normal transformada"(I).

Percentil		5	10	25	50	75	90	95
Verdadeiro (mediano)		74,79	80,29	89,96	101,66	114,21	126,53	134,26
Modelo								
BCCG	$\overline{\widehat{\psi}}$	76,22	81,23	90,34	101,69	114,47	127,34	135,73
	Viés $(\widehat{\psi})$	1,91	1,16	0,42	0,03	0,22	0,64	1,10
	$\sqrt{EQMR(\widehat{\psi})}$	0,03	0,02	0,01	0,01	0,01	0,02	0,02
BCT	$\overline{\widehat{\psi}}$	76,30	81,29	90,38	101,68	114,40	127,19	135,54
	Viés $(\widehat{\psi})$	2,02	1,24	0,46	0,02	0,16	0,52	0,96
	$\sqrt{EQMR(\widehat{\psi})}$	0,03	0,02	0,01	0,01	0,01	0,02	0,02
BCT $\tau = 4$	$\overline{\widehat{\psi}}$	75,19	80,30	89,63	101,27	114,43	127,75	136,43
	Viés $(\widehat{\psi})$	0,54	0,01	$-0,37$	$-0,38$	0,19	0,96	1,62
	$\sqrt{EQMR(\widehat{\psi})}$	0,02	0,02	0,01	0,01	0,01	0,02	0,02
BCT $\tau = 3$	$\overline{\widehat{\psi}}$	75,28	80,42	89,79	101,51	114,75	128,15	136,89
	Viés $(\widehat{\psi})$	0,66	0,16	$-0,18$	$-0,15$	0,48	1,28	1,96
	$\sqrt{EQMR(\widehat{\psi})}$	0,02	0,02	0,01	0,01	0,01	0,02	0,03
BCT $\tau = 2$	$\overline{\widehat{\psi}}$	76,64	81,77	91,11	102,76	115,90	129,17	137,82
	Viés $(\widehat{\psi})$	2,47	1,84	1,28	1,09	1,48	2,09	2,65
	$\sqrt{EQMR(\widehat{\psi})}$	0,03	0,02	0,02	0,02	0,02	0,03	0,03
BCT $\tau = 1,5$	$\overline{\widehat{\psi}}$	76,07	81,18	90,48	102,08	115,18	128,40	137,02
	Viés $(\widehat{\psi})$	1,72	1,10	0,57	0,41	0,85	1,48	2,06
	$\sqrt{EQMR(\widehat{\psi})}$	0,03	0,02	0,01	0,01	0,01	0,02	0,03
Verdadeiro (médio)		76,84	82,28	92,11	103,86	116,63	129,00	136,87
NCI	$\overline{\widehat{\psi}}$	76,92	82,46	92,18	103,91	116,52	128,86	136,67
	Viés $(\widehat{\psi})$	0,11	0,22	0,08	0,05	$-0,09$	$-0,11$	$-0,14$
	$\sqrt{EQMR(\widehat{\psi})}$	0,02	0,02	0,01	0,01	0,01	0,02	0,02

A Tabela 4.3, referente ao cenário de leve assimetria à direita, revela que o método NCI apresentou os menores vieses relativos. Os modelos BCCG e BCT apresentaram vieses moderados, não superando 2% na maioria dos casos. A Tabela 4.4, referente ao caso de assimetria à direita acentuada, mostra resultados bem diferentes: o método NCI apresentou o pior desempenho, subestimando todos os percentis médios. O modelo BCT com τ estimado mostrou a melhor performance. O desempenho do método NCI tornou-se ainda mais insatisfatório no cenário em que há simultaneamente assimetria à direita acentuada e presença de pontos aberrantes (Tabela 4.5). Novamente, o método

NCI subestimou todos os percentis médios. O modelo BCT com τ estimado permaneceu sendo o que apresenta melhor performance. Destaca-se ainda que a estimativa média do parâmetro τ no modelo BCT foi alta, o que sugere um modelo próximo ao BCCG. Não houve problemas de convergência neste cenário.

Tabela 4.4: Cenário 2: Percentis medianos e médios, vieses relativos em porcentagem e raiz quadrada do erro quadrático médio relativo. Dados gerados a partir de uma distribuição "normal transformada"(II).

Percentil		5	10	25	50	75	90	95
Verdadeiro (mediano)		304,41	364,97	496,26	694,4	967,44	1298,55	1543,38
Modelo								
BCCG	$\overline{\widehat{\psi}}$	313,29	373,49	501,32	695,56	965,69	1297,98	1549,43
	Viés $(\widehat{\psi})$	2,92	2,33	1,02	0,17	$-0,18$	$-0,04$	0,39
	$\sqrt{EQMR(\widehat{\psi})}$	0,07	0,06	0,04	0,03	0,04	0,05	0,06
BCT	$\overline{\widehat{\psi}}$	312,56	372,96	501,17	696,20	967,67	1302,05	1555,66
	Viés $(\widehat{\psi})$	2,68	2,19	0,99	0,26	0,02	0,02	0,27
	$\sqrt{EQMR(\widehat{\psi})}$	0,07	0,06	0,04	0,03	0,04	0,05	0,06
BCT $\tau = 4$	$\overline{\widehat{\psi}}$	298,87	359,86	490,68	692,84	978,70	1336,78	1610,99
	Viés $(\widehat{\psi})$	$-1,82$	$-1,40$	$-1,13$	$-0,23$	1,16	2,94	4,38
	$\sqrt{EQMR(\widehat{\psi})}$	0,07	0,06	0,04	0,03	0,04	0,06	0,08
BCT $\tau = 3$	$\overline{\widehat{\psi}}$	294,03	355,27	487,49	693,32	986,32	1355,15	1639,66
	Viés $(\widehat{\psi})$	$-3,41$	$-2,66$	$-1,77$	$-0,16$	1,95	4,36	6,24
	$\sqrt{EQMR(\widehat{\psi})}$	0,07	0,06	0,04	0,03	0,05	0,07	0,09
BCT $\tau = 2$	$\overline{\widehat{\psi}}$	315,77	380,34	519,35	734,57	1039,60	1421,38	1713,96
	Viés $(\widehat{\psi})$	3,73	4,21	4,65	5,78	7,46	9,46	11,05
	$\sqrt{EQMR(\widehat{\psi})}$	0,08	0,07	0,06	0,07	0,08	0,11	0,13
BCT $\tau = 1,5$	$\overline{\widehat{\psi}}$	313,57	376,62	511,97	720,31	1013,86	1379,99	1659,75
	Viés $(\widehat{\psi})$	3,01	3,19	3,16	3,73	4,80	6,27	7,54
	$\sqrt{EQMR(\widehat{\psi})}$	0,08	0,07	0,05	0,05	0,06	0,08	0,10
Verdadeiro (médio)		442,08	526,85	710,79	986,74	1361,78	1814,55	2147,29
NCI	$\overline{\widehat{\psi}}$	425,80	510,15	685,06	947,80	1302,80	1733, 26	2053,56
	Viés $(\widehat{\psi})$	$-3,68$	$-3,17$	$-3,62$	$-3,95$	$-4,33$	$-4,48$	$-4,37$
	$\sqrt{EQMR(\widehat{\psi})}$	0,08	0,05	0,05	0,05	0,06	0,07	0,07

Em resumo, sob o cenário em que os dados satisfazem os pressupostos do método NCI, este pode ter tanto melhor ou pior desempenho que os modelos medianos a depender do grau de assimetria dos dados. Em situação de alta assimetria e, principalmente, se acompanhada de cauda direita pesada, os modelos medianos são uma

alternativa recomendável ao método NCI.

Tabela 4.5: Cenário 2: Percentis medianos e médios, vieses relativos em porcentagem e raiz quadrada do erro quadrático médio. Dados gerados a partir de uma distribuição "normal transformada" contaminada (III).

Percentil		5	10	25	50	75	90	95
Verdadeiro (mediano)		320,87	384,82	521,16	728,18	1013,36	1359,03	1617,61
Modelo								
BCCG	$\overline{\widehat{\psi}}$	326,76	390,04	524,36	728,95	1013,85	1364,82	1631,32
	Viés $(\widehat{\psi})$	1,84	1,36	0,61	0,11	0,05	0,43	0,85
	$\sqrt{EQMR(\widehat{\psi})}$	0,07	0,06	0,04	0,03	0,04	0,06	0,07
BCT	$\overline{\widehat{\psi}}$	327,62	390,84	525,00	729,36	1013,74	1363,99	1629,69
	Viés $(\widehat{\psi})$	2,10	1,56	0,74	0,16	0,04	0,37	0,75
	$\sqrt{EQMR(\widehat{\psi})}$	0,07	0,06	0,04	0,03	0,04	0,06	0,07
BCT $\tau = 4$	$\overline{\widehat{\psi}}$	313,62	377,64	515,64	728,76	1030,30	1407,98	1698,03
	Viés $(\widehat{\psi})$	$-2,26$	$-1,87$	$-1,06$	0,08	1,67	3,60	4,97
	$\sqrt{EQMR(\widehat{\psi})}$	0,07	0,06	0,04	0,03	0,04	0,07	0,08
BCT $\tau = 3$	$\overline{\widehat{\psi}}$	306,56	370,87	509,95	726,91	1036,67	1427,49	1729,12
	Viés $(\widehat{\psi})$	$-4,46$	$-3,62$	$-2,15$	$-0,17$	2,30	5,04	6,89
	$\sqrt{EQMR(\widehat{\psi})}$	0,08	0,06	0,05	0,03	0,05	0,08	0,10
BCT $\tau = 2$	$\overline{\widehat{\psi}}$	332,86	401,62	549,65	779,46	1105,78	1515,43	1830,50
	Viés $(\widehat{\psi})$	3,74	4,37	5,47	7,04	9,12	11,51	13,16
	$\sqrt{EQMR(\widehat{\psi})}$	0,08	0,07	0,07	0,08	0,10	0,13	0,15
BCT $\tau = 1,5$	$\overline{\widehat{\psi}}$	328,61	395,17	537,98	758,43	1069,66	1458,24	1755,52
	Viés $(\widehat{\psi})$	2,41	2,69	3,23	4,15	5,56	7,30	8,53
	$\sqrt{EQMR(\widehat{\psi})}$	0,07	0,06	0,05	0,05	0,07	0,09	0,11
Verdadeiro (médio)		493,43	590,02	794,73	1099,84	1513,69	2012,91	2386,86
NCI	$\overline{\widehat{\psi}}$	455,44	546,40	736,23	1024,19	1417,84	1901,15	2264,45
	Viés $(\widehat{\psi})$	$-7,70$	$-7,39$	$-7,36$	$-6,88$	$-6,33$	$-5,55$	$-5,13$
	$\sqrt{EQMR(\widehat{\psi})}$	0,10	0,09	0,08	0,08	0,07	0,07	0,08

4.4 Cenário 3: Dados gerados a partir da distribuição gama

Nos cenários considerados anteriormente os dados foram gerados sob pressupostos dos modelos BCT(Cenário 1, Tabela 4.1), BCCG (Cenário 1, Tabela 4.2) e NCI(Cenário

2, Tabelas 4.3 e 4.4). No cenário considerado a seguir, os dados são gerados sob um modelo que não se encaixa nas suposições de nenhum dos modelos considerados nesta tese. Neste cenário os dados são gerados sob o seguinte modelo:

$$\begin{aligned} Y_{ij}|\gamma_i &\overset{ind}{\sim} \text{Gama}(\mu_i, \phi), \\ \log(\mu_i) &= \beta_0 + \gamma_i, \\ \gamma_i &\overset{ind}{\sim} N(0, \lambda^2), \end{aligned}$$

$i = 1, \ldots, 500$, $j = 1, 2, 3$. Os valores atribuídos aos parâmetros são $\beta_0 = 1$, $\phi = 2$ e $\lambda^2 = 0,25$, os quais proporcionam um cenário com uma assimetria acentuada à direita. Para gerar pontos discrepantes, 95% dos dados foram obtidos dessa distribuição gama e 5% das observações foram provenientes de outra distribuição gama com os mesmos valores dos parâmetros, exceto pelo parâmetro β_0, que agora é 4.

A Tabela 4.6 apresenta os percentis medianos e médios, os vieses relativos em porcentagem, a raiz quadrada do erro quadrático médio relativo calculados a partir dos modelos Box-Cox *t* e Box-Cox Cole-Green com intercepto aleatório e método NCI, para os dados simulados seguindo a distribuição gama com pontos *outliers*. Ressalta-se que os modelos BCT foram os que apresentaram os menores valores de vieses relativos. Os modelos baseados na distribuição normal apresentaram um pior ajuste nos percentis superiores, por exemplo, no modelo BCCG, o viés relativo para o ajuste do quantil de 95% é de 92,84%. O método NCI substima os percentis da distribuição "verdadeira" das médias de consumo.

As porcentagens de amostras descartadas foram: 6,8%, para o modelo BCCG; BCT com τ estimado, 10%; BCT com $\tau = 4$, 5,8%; BCT com $\tau = 3$, 9,8%; BCT com $\tau = 2$, 8,8% e BCT com $\tau = 1,5$, 29,4%.

Tabela 4.6: Cenário 3: Percentis medianos e médios, vieses relativos em porcentagem e raiz quadrada do erro quadrático médio relativo. Dados gerados a partir de uma gama contaminada.

Percentil		5	10	25	50	75	90	95
Verdadeiro (mediano)		0,69	0,92	1,43	2,22	3,33	4,72	5,76
Modelo								
BCCG	$\overline{\widehat{\psi}}$	0,61	0,84	1,42	2,55	4,64	7,99	11,11
	Viés $(\widehat{\psi})$	10,81	$-9,14$	$-1,06$	14,91	39,11	69,53	92,84
	$\sqrt{EQMR(\widehat{\psi})}$	0,19	0,15	0,08	0,18	0,43	0,76	1,01
BCT	$\overline{\widehat{\psi}}$	0,89	1,10	1,55	2,28	3,37	4,79	5,91
	Viés $(\widehat{\psi})$	29,34	18,85	8,47	2,85	1,06	1,50	2,60
	$\sqrt{EQMR(\widehat{\psi})}$	0,31	0,21	0,11	0,05	0,04	0,08	0,11
BCT $\tau = 4$	$\overline{\widehat{\psi}}$	0,88	1,08	1,52	2,22	3,26	4,60	5,65
	Viés $(\widehat{\psi})$	28,20	17,30	6,34	0,14	$-2,30$	$-2,54$	$-1,93$
	$\sqrt{EQMR(\widehat{\psi})}$	0,31	0,20	0,09	0,04	0,05	0,07	0,09
BCT $\tau = 3$	$\overline{\widehat{\psi}}$	0,88	1,08	1,52	2,23	3,28	4,63	5,70
	Viés $(\widehat{\psi})$	27,80	17,11	6,47	0,50	$-1,76$	$-1,81$	$-1,11$
	$\sqrt{EQMR(\widehat{\psi})}$	0,29	0,19	0,09	0,04	0,04	0,06	0,07
BCT $\tau = 2$	$\overline{\widehat{\psi}}$	0,92	1,12	1,57	2,30	3,36	4,73	5,81
	Viés $(\widehat{\psi})$	32,78	21,50	10,06	3,51	0,78	0,34	0,82
	$\sqrt{EQMR(\widehat{\psi})}$	0,34	0,23	0,12	0,05	0,04	0,06	0,07
BCT $\tau = 1,5$	$\overline{\widehat{\psi}}$	0,90	1,10	1,56	2,29	3,36	4,76	5,86
	Viés $(\widehat{\psi})$	30,54	19,73	9,00	3,01	0,82	0,89	1,72
	$\sqrt{EQMR(\widehat{\psi})}$	0,28	0,19	0,10	0,04	0,04	0,05	0,07
Verdadeiro (médio)		2,27	2,73	3,74	5,29	7,47	10,17	12,26
NCI	$\overline{\widehat{\psi}}$	1,25	1,64	2,53	4,00	6,19	9,06	11,31
	Viés $(\widehat{\psi})$	$-44,67$	$-39,81$	$-32,38$	$-24,40$	$-17,19$	$-10,98$	$-7,79$
	$\sqrt{EQMR(\widehat{\psi})}$	0,45	0,40	0,33	0,25	0,17	0,12	0,10

4.5 Conclusões

O modelo Box-Cox t com efeito aleatório proposto nesta tese para a estimação de percentis da distribuição dos consumos usuais (medianos) apresentou desempenho satisfatório nos diversos cenários simulados. Este modelo mostrou performance muito superior ao método NCI nos cenários de assimetria positiva acentuada, especialmente

na presença de cauda direita pesada. O método NCI apresentou melhores resultados que os modelos medianos (BCT e BCCG) apenas na situação em que simultaneamente os dados foram gerados de acordo com os seus pressupostos e sob cenário de leve assimetria.

Deve ser notado que, sob os modelos medianos, há a possibilidade de não convergência no processo numérico de maximização da função de verossimilhança. Nesses casos, recomenda-se inspeção da função de verossimilhança perfilada através de uma malha de valores para τ, ou seja, para o parâmetro referente ao número de graus de liberdade.

CAPÍTULO 5

Classe das distribuições Box-Cox simétricas

Este capítulo tem como objetivo propor uma nova classe de distribuições proveniente de uma transformação baseada em Box & Cox (1964), a qual envolve uma classe de distribuições simétricas truncadas. Tais distribuições são denominadas nesta tese como Box-Cox simétricas (BCS). Aqui é apresentada uma breve revisão sobre a classe de distribuições simétricas e log-simétricas, a definição da classe de distribuições Box-Cox simétricas, algumas propriedades, inferência, percentis, momentos e abordagens para uma análise sobre o peso nas caudas. Por fim, aplicações a dados de consumo de 33 nutrientes são apresentadas e uma comparação com enfoques alternativos é discutida. Detalhes técnicos são apresentados em apêndices.

5.1 Distribuições Simétricas

Diz-se que a variável aleatória W tem distribuição simétrica, com suporte em $\mathbb{R}$, parâmetro de locação $\mu \in \mathbb{R}$ e de escala $\sigma > 0$, se sua função densidade de probabilidade é dada por

$$v(w, \mu, \sigma; r) = \frac{1}{\sigma} r\left(\left(\frac{w-\mu}{\sigma}\right)^2\right), \qquad w \in \mathbb{R}, \tag{5.1}$$

para alguma função $r(\cdot)$, denominada função geradora de densidades, com $r(u) > 0$, para $u > 0$ e $\int_0^\infty u^{-1/2} r(u) \mathrm{d}u = 1$. Essa condição é necessária para que $v(w, \mu, \sigma; r)$ seja uma função densidade de probabilidade. Assim, denota-se $W \sim S(\mu, \sigma^2; r)$, e diz-se que

W é uma variável aleatória simétrica.

Seguem algumas propriedades das distribuições simétricas (Cysneiros, 2004).

(i) Se $W \sim S(\mu, \sigma^2; r)$, então a função característica de W é dada por $\psi_W(t) = e^{it\mu}\varphi(t^2\sigma^2)$, $t \in \mathbb{R}$, para alguma função φ, com $\varphi(u) \in \mathbb{R}$, para $u > 0$. Quando existem, $\mathrm{E(W)} = \mu$ e $\mathrm{Var(W)} = \xi\sigma^2$, em que $\xi > 0$ é uma constante dada por $\xi = -2\varphi'(0)$, com $\varphi'(0) = d\varphi(u)/du|_{u=0}$ e que não depende dos parâmetros μ e σ (Fang et al., 1990). Se $u^{\frac{-(k+1)}{2}} r(u)$ for integrável, então o k-ésimo momento de W existe (Kelker, 1970).

(ii) Se $W \sim S(\mu, \sigma^2; r)$, então $a + bW \sim S(a + b\mu, b^2\sigma^2; r)$, em que $a, b \in \mathbb{R}$, com $b \neq 0$, ou seja, a distribuição de qualquer combinação linear de uma variável aleatória com distribuição simétrica é também simétrica. Em particular, se $W \sim S(\mu, \sigma^2; r)$, então $S = (W - \mu)/\sigma \sim S(0, 1; r)$, com função de densidade $v(s) = r(s^2)$.

(iii) Berkane & Bentler (1986), considerando uma distribuição simétrica padrão $S \sim S(0, 1; r)$ e a existência de seus momentos, mostraram que a função característica de S pode ser expandida como

$$\psi_S(t) = \sum_{k=0}^{\infty} i^k \mu'_k \frac{t^k}{k!}, \tag{5.2}$$

em que $\mu'_k = \mathrm{E(S^k)} = i^{-\mathrm{k}}\psi_{\mathrm{S}}^{(\mathrm{k})}(0)$, com $\psi_S^{(k)}(0)$ denotando a k-ésima derivada de $\psi_S^k(t)$ avaliada em $t = 0$. Então

$$\mu'_k = \begin{cases} 0, & \text{se} \quad k \text{ é ímpar}, \\ \dfrac{(2m)!}{2^m m!}(\mu'_2)^m(k(m) + 1), & \text{se} \quad k = 2m, m = 1, 2, \ldots, \end{cases}$$

sendo

$$k(m) = \frac{\varphi^{(m)}(0)}{\{\varphi^{(1)}(0)\}^m} - 1,$$

em que $\varphi^{(r)}(0)$ é a r-ésima derivada da função φ, avaliada em zero. Os coeficientes $k(m)$, $m = 1, 2, \ldots$ são conhecidos como parâmetros de momentos e generalizam o coeficiente de curtose $\gamma_2 = 3\{k(2) + 1\}$ de uma distribuição $S(\mu, \sigma^2; r)$ (Muirhead, 1982).

Cambanis et al. (1981) observaram que a família de distribuições simétricas coincide com a classe de distribuições elípticas univariadas. A partir dos trabalhos de Kelker (1970) muitos estudos sobre distribuições elípticas univariadas e multivariadas sur-

giram. Alguns trabalhos que discutem aspectos destas distribuições são Rao (1990), Anderson & Fang (1987), Arellano-Valle (1994, Capítulo 2) e Cysneiros (2004).

A seguir, algumas funções geradoras de densidades (5.1) que são utilizadas para compor esta classe são explicitadas:

(i) Normal: $r(u) = (2\pi)^{-1/2}\exp\{-u/2\}$;

(ii) Exponencial dupla: $r(u) = \sqrt{2}/2\exp\{-\sqrt{2}|u|^{1/2}\}$;

(iii) Exponencial potência: $r(u) = [\tau\exp\{-1/2|u|^{\tau/2}/|p(\tau)|^{\tau}\}]/[p(\tau)2^{1+1/\tau}\Gamma(1/\tau)]$, em que $\tau > 0$ e $p(\tau)^2 = 2^{-2/\tau}\Gamma(1/\tau)[\Gamma(3/\tau)]^{-1}$; $\tau = 1$ e $\tau = 2$, correspondem às funções geradoras de densidades das distribuições exponencial dupla e normal, respectivamente;

(iv) Cauchy: $r(u) = \{\pi(1+u)\}^{-1}$;

(v) t-Student: $r(u) = \tau^{\tau/2}\{B(1/2,\tau/2)\}^{-1}(\tau+u)^{-(\tau+1)/2}$, $\tau > 0$, em que $B(\cdot,\cdot)$ é a função beta; $\tau = 1$ corresponde à função geradora de densidade da distribuição Cauchy;

(vi) Logística tipo I: $r(u) = c\exp\{-u\}(1+\exp\{-u\})^{-2}$, em que $c \approx 1,484300029$ é a constante normalizadora, obtida da relação $\int_0^\infty u^{-1/2}r(u)\mathrm{d}u = 1$;

(vii) Logística tipo II: $r(u) = \exp\{-u^{1/2}\}(1+\exp\{-u^{1/2}\})^{-2}$;

(viii) Slash canônica (Goméz et al., 2007): $r(u) = [1/(\sqrt{2\pi}u)](1-\exp\{-u/2\})$, para $u \neq 0$, e $r(u) = 1/(2\sqrt{2\pi})$, caso contrário;

(ix) Slash (Goméz et al., 2007): $r(u) = \Psi((q+1)/2, u/2)q2^{q/2-1}/(\sqrt{\pi}u^{(q+1)/2})$, com $q > 0$, $\Psi(\cdot,\cdot)$ uma função gama incompleta inferior, definida por $\Psi(a,x) = \int_0^x t^{a-1}\exp\{-t\}\mathrm{d}t$; quando $q = 1$ esta coincide com a função geradora da distribuição slash canônica.

Nota-se que esta classe de distribuições inclui a distribuição normal, bem como distribuições simétricas com caudas mais leves (por exemplo, logística tipo I) e mais pesadas (como a t-Student) que a distribuição normal.

5.2 Distribuições Log-simétricas

Diz-se que uma variável aleatória estritamente positiva L tem distribuição log-simétrica (Vanegas & Paula, 2014a; Vanegas & Paula, 2014b), com parâmetros $\eta \in \mathbb{R}$ e $\sigma > 0$, se sua função densidade de probabilidade é dada por

$$v(l,\eta,\sigma;r) = \frac{1}{\sigma l}r\left(\log^2\left(\frac{l}{\eta}\right)^{\frac{1}{\sigma}}\right), \qquad l > 0, \tag{5.3}$$

para uma função geradora de densidades $r(\cdot)$, com $r(u) > 0$, para $u > 0$ e $\int_0^\infty u^{-1/2} r(u) \mathrm{d}u = 1$. Assim, denota-se $L \sim LS(\eta, \sigma^2; r)$, e diz-se que L é uma variável aleatória log-simétrica. Essa classe inclui as distribuições log-normal, log-t-Student, log-exponencial potência, log-logística tipos I e II entre outras. A distribuição (5.3) generaliza a distribuição log-normal e introduz uma classe de distribuições mais flexíveis para descrever dados contínuos, positivos e assimétricos.

Se $L \sim LS(\eta, \sigma^2; r)$, então pode-se verificar que

(i) $W = \log(L) \sim S(\mu, \sigma^2; r)$, com $\mu = \log \eta$. Quando existem, E(W) = μ e Var(W) $\propto \sigma^2$. Desse modo, tem-se, por exemplo, que se L segue uma distribuição log-normal, log-t-Student e log-exponencial potência, W segue uma distribuição normal, t-Student e exponencial potência, respectivamente;

(ii) a mediana e o intervalo interquartílico de L são dados por η e $1,5$ senh$(\sigma s_{0,75})$, respectivamente, em que senh$(\cdot)$ é a função seno hiperbólico e $s_{0,75}$ é o terceiro quartil de uma variável aleatória $S \sim S(0, 1; r)$.

(iii) $(L/\eta)^{\frac{1}{\sigma}} \sim LS(1, 1; r)$, uma distribuição log-simétrica padrão;

(iv) $dL \sim LS(d\eta, \sigma^2; r)$, para toda constante $d > 0$;

(v) $L^d \sim LS(\eta^d, d^2\sigma^2; r)$, para toda constante $d \neq 0$;

(vi) L/η e η/L são variáveis aleatórias identicamente distribuídas;

(vii) o quantil de ordem α de L é dado por $q_\alpha = \eta \exp\{\sigma s_\alpha\}$, em que s_α é o quantil de ordem α de uma variável aleatória $S \sim S(0, 1; r)$;

(viii) o coeficiente de variação interquartílico é dado por $\varpi = (q_{0,75} - q_{0,25})/(q_{0,75} + q_{0,25}) = \tanh(\sigma s_{0,75})$, em que tanh$(\cdot)$ é a função tangente hiperbólica e q_α o quantil de ordem α de L.

5.3 A classe de distribuições Box-Cox simétricas

Seja Y uma variável aleatória positiva e contínua. A classe de distribuições Box-Cox simétricas é definida a partir da transformação [1]

$$Z = Z(Y, \mu, \sigma, \nu, \tau) = \begin{cases} \dfrac{1}{\sigma\nu}\left[\left(\dfrac{Y}{\mu}\right)^\nu - 1\right], & \text{se} \quad \nu \neq 0, \\ \dfrac{1}{\sigma}\log\left(\dfrac{Y}{\mu}\right), & \text{se} \quad \nu = 0, \end{cases} \tag{5.4}$$

[1] Para facilitar a leitura, algumas fórmulas que aparecem no Capítulo 1 serão repetidas aqui.

em que $\mu > 0$, $\sigma > 0$, $-\infty < \nu < \infty$ e Z tem uma distribuição simétrica padrão truncada no intervalo

$$A(\sigma,\nu) = \begin{cases} \left(-\frac{1}{\sigma\nu}, \infty\right), & \text{se} \quad \nu > 0, \\ \left(-\infty, -\frac{1}{\sigma\nu}\right), & \text{se} \quad \nu < 0, \\ (-\infty, \infty), & \text{se} \quad \nu = 0, \end{cases} \tag{5.5}$$

e denota-se $Z \sim S(0,1,A(\sigma,\nu);r)$, com $r(\cdot)$ sendo a função geradora de densidades definida em (5.1) para $\nu \neq 0$ e em (5.3), caso contrário. Assim, diz-se que Y tem distribuição BCS com parâmetros μ, σ e ν, e escreve-se $Y \sim BCS(\mu,\sigma,\nu;r)$, se Z dado em (5.4) é tal que $Z \sim S(0,1,A(\sigma,\nu);r)$.

A função densidade de probabilidade de Y é dada por

$$f_Y(y) = f_Z(z)\left|\frac{dz}{dy}\right| = \frac{y^{\nu-1}}{\mu^\nu\sigma} f_Z(z), \tag{5.6}$$

em que $z = Z(y)$, com $Z(\cdot)$ dado em (5.4) e

$$f_Z(z) = \frac{f_S(z)}{F_S\left(\frac{1}{\sigma|\nu|}\right)}, \tag{5.7}$$

com $f_S(\cdot)$ e $F_S(\cdot)$ sendo as funções de densidade de probabilidade e de distribuição acumulada de uma variável aleatória $S \sim S(0,1;r)$, respectivamente [2]. Note que $f_S(s) = r(s^2)$, e assim tem-se de (5.1), (5.4) e (5.7) que (5.6) por ser reescrita como

$$f_Y(y) = \begin{cases} \frac{y^{\nu-1}}{\mu^\nu\sigma} \frac{r\left(\left\{\frac{1}{\sigma\nu}\left[\left(\frac{y}{\mu}\right)^\nu - 1\right]\right\}^2\right)}{R\left(\frac{1}{\sigma|\nu|}\right)}, & \text{se} \quad \nu \neq 0, \\ \frac{1}{y\sigma} r\left(\left\{\frac{1}{\sigma}\log\left(\frac{y}{\mu}\right)\right\}^2\right), & \text{se} \quad \nu = 0, \end{cases} \tag{5.8}$$

em que $R(w) = \int_{-\infty}^{w} r(u^2)\mathrm{d}u$, para $w \in \mathbb{R}$.

A função distribuição acumulada de Y é dada por

$$F_Y(y) = \begin{cases} \frac{F_S(z)}{F_S\left(\frac{1}{\sigma\,|\,\nu\,|}\right)}, & \text{se} \quad \nu \leq 0, \\ \frac{F_S(z) - F_S\left(-\frac{1}{\sigma\,|\,\nu\,|}\right)}{F_S\left(\frac{1}{\sigma\,|\,\nu\,|}\right)}, & \text{se} \quad \nu > 0. \end{cases} \tag{5.9}$$

[2] Assume-se que $1/\sigma|\nu|$, quando $\sigma|\nu| = 0$, é interpretado como $\lim_{\sigma\nu \to 0}(1/\sigma|\nu|) = \infty$ e, nesse caso, $F(1/\sigma|\nu|) = 1$.

Nota-se que a classe inclui as distribuições Box-Cox *t* (Rigby & Stasinopoulos, 2006a) detalhada no Capítulo 1, Box-Cox Cole-Green (Stasinopoulos et al., 2008) e Box-Cox exponencial potência (Voudouris et al., 2012), permitindo ainda outras funções geradoras de densidades, as quais estão explicitadas na Seção 5.1.

A Figura 5.1 mostra funções de densidade de probabilidade das distribuições Box-Cox Cole-Green (BCCG), Box-Cox *t* (BCT), Box-Cox exponencial potência (BCPE) e Box-Cox slash (BCSlash) para uma particular escolha dos parâmetros. Aparentemente, as distribuições BCT e BCSlash apresentam caudas mais pesadas do que as demais explicitadas. A Figura 5.2 ilustra a função de densidade de probabilidade da distribuição Box-Cox slash para várias combinações de valores dos parâmetros (na Seção 1.3 tem-se para Box-Cox *t*). Note que σ está relacionado a dispersão e q controla o peso da cauda da distribuição. Detalhes sobre o peso das caudas de distribuições simétricas e Box-Cox simétricas serão discutidos na Seção 5.3.5.

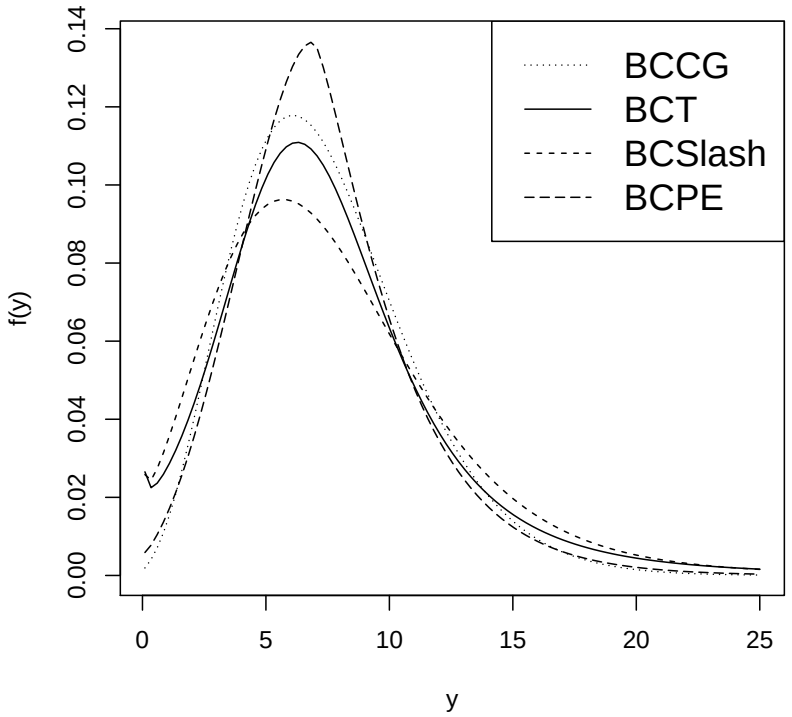
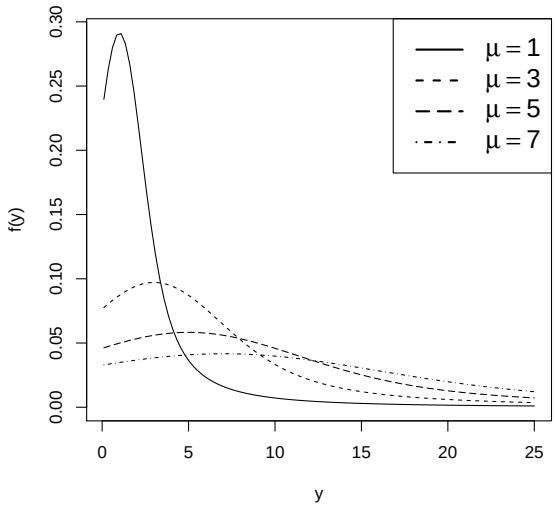
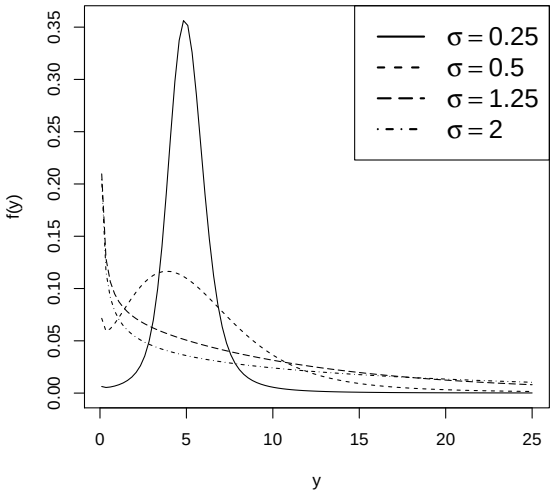


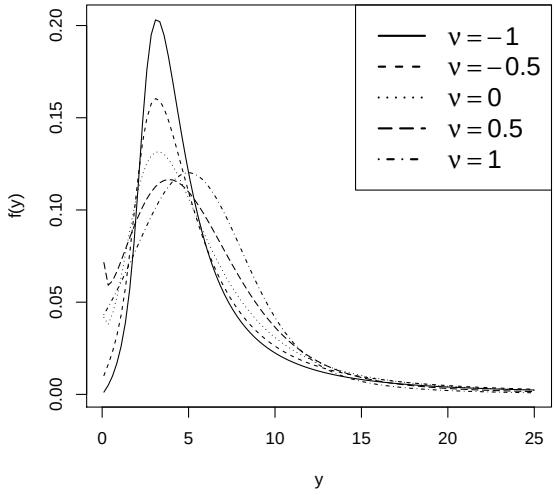
Figura 5.1: Gráficos das funções de densidade probabilidade das distribuições BCCG, BCT ($\tau = 4$), BCSlash($q = 4$), BCPE($\tau = 1,5$) para $\mu = 7$; $\sigma = 0,5$; $\nu = 0,5$.



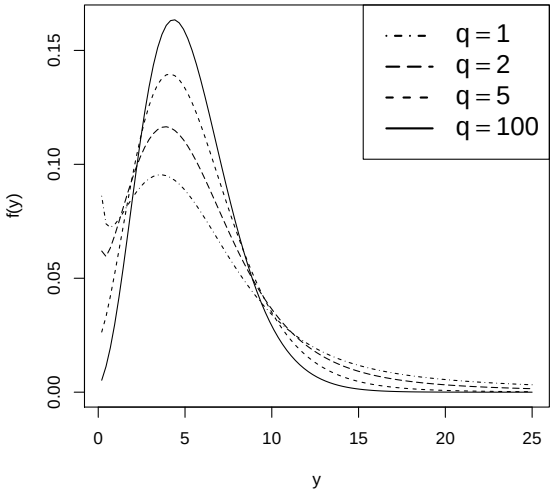
(a) $\sigma = 1; \nu = 1; q = 1$.



(b) $\mu = 5; \nu = 0,5; q = 2$.



(c) $\mu = 5; \sigma = 0,5; q = 2$.



(d) $\mu = 5; \sigma = 0,5; \nu = 0,5$.

Figura 5.2: Gráficos de função densidade probabilidade da distribuição BCSlash(μ, σ, ν, q).

5.3.1 Algumas propriedades

Se $Y \sim BCS(\mu, \sigma, \nu; r)$, $\mu > 0$, $\sigma > 0$ e $\nu \in \mathbb{R}$ então pode-se verificar que

(i) $\frac{Y}{\mu} \sim BCS(1, \sigma, \nu; r)$;

(ii) $dY \sim BCS(d\mu, \sigma, \nu; r)$, para toda constante $d > 0$;

(iii) $\left(\frac{Y}{\mu}\right)^{\frac{1}{\sigma}} \sim BCS(1, 1, \sigma\nu; r)$;

(iv) $\left(\frac{Y}{\mu}\right)^{\nu} \sim BCS(1, \sigma\nu, 1; r)$, se $\nu > 0$;

(v) se $\nu = 1$, então $Y \sim BCS(\mu, \sigma, 1; r)$ tem uma distribuição simétrica truncada com parâmetros μ e $\mu^2\sigma^2$ e suporte em $(0, \infty)$;

(vi) se $\nu = 0$, então $Y \sim BCS(\mu, \sigma, 0; r) = LS(\log(\mu), \sigma^2; r)$.

Nota-se em (i) que μ é parâmetro de escala e em (iii) tem-se que σ é parâmetro de forma. Adicionalmente, de (vi) constata-se que a classe BCS generaliza a classe LS. As propriedades (i), (iii) e (iv) resumem-se a $(Y/\mu)^d \sim BCS(1, d\sigma, \nu/d; r)$, para toda constante $d > 0$.

5.3.2 Percentis

O quantil de ordem α de $Y \sim BCS(\mu, \sigma, \nu; r)$, denotado por y_α, é definido a partir do correspondente quantil de uma distribuição $Z \sim S(0, 1, A(\sigma, \nu); r)$, denotado por z_α. Pode-se mostrar que

$$y_\alpha = \begin{cases} \mu(1 + \sigma\nu z_\alpha)^{\frac{1}{\nu}}, & \text{se} \quad \nu \neq 0, \\ \mu \exp(\sigma s_\alpha), & \text{se} \quad \nu = 0, \end{cases}$$

em que

$$z_\alpha = \begin{cases} F_S^{-1}\left[\alpha F_S\left(\frac{1}{\sigma \mid \nu \mid}\right)\right], & \text{se} \quad \nu < 0, \\ F_S^{-1}\left[1 - (1 - \alpha)F_S\left(\frac{1}{\sigma \mid \nu \mid}\right)\right], & \text{se} \quad \nu > 0, \\ F_S^{-1}(\alpha) = s_\alpha, & \text{se} \quad \nu = 0, \end{cases}$$

com $F_S^{-1}(\cdot)$ sendo a inversa da função distribuição acumulada (definida em (5.9)) de uma variável aleatória que segue uma distribuição simétrica padrão, $S \sim S(0, 1; r)$. Nota-se que os quantis de Y são todos proporcionais a μ. Em particular, a mediana de Y é dada por

$$y_{1/2} = \begin{cases} \mu(1 + \sigma\nu z_{1/2})^{\frac{1}{\nu}}, & \text{se} \quad \nu \neq 0, \\ \mu \exp(\sigma s_{1/2}), & \text{se} \quad \nu = 0. \end{cases}$$

Note que, se a região de truncamento tem probabilidade negligenciável, tem-se $z_{1/2} \approx 0$ e $\mu \approx y_{1/2}$, ou seja, o parâmetro μ é aproximadamente igual à mediana da distribuição $BCS(\mu, \sigma, \nu; r)$. Ainda, μ é a mediana de Y quando $\nu = 0$ (transformação log).

Um coeficiente de variação para uma variável aleatória Y baseado nos percentis (Rigby & Stasinopoulos, 2006b), denotado por CV_Y, é definido como

$$CV_Y = \frac{3}{4}\frac{y_{0,75} - y_{0,25}}{y_{0,5}}.$$

Ignorando a região de truncamento, quando $\nu \neq 0$, tem-se que $CV_Y \approx 0,75\{[1 + \sigma\nu s_{0,75}]^{1/\nu} - [1 - \sigma\nu s_{0,75}]^{1/\nu}\}$; e para $\nu \approx 0$ tem-se que $CV_Y \approx 1,5 \operatorname{senh}(\sigma s_{0,75})$, uma função crescente em σ (vale a igualdade quando $\nu = 0$), em que $s_{0,75}$ é o terceiro quartil de uma variável S que segue uma distribuição simétrica padrão, $S \sim S(0, 1; r)$ e senh$(\cdot)$ é a função seno hiperbólico. Portanto, σ pode ser visto como um parâmetro de dispersão relativa.

Assim, tem-se que os parâmetros μ, σ e ν podem ser interpretados como escala (relacionado à mediana), dispersão relativa (associado ao coeficiente de variação baseado nos percentis), e à assimetria (dado pela transformação potência para simetria), respectivamente. Adicionalmente, um parâmetro extra pode ser incorporado, como por exemplo, o parâmetro referente ao número de graus de liberdade na distribuição Box-Cox t, o qual controla o peso da cauda.

5.3.3 Inferência

Um método de estimação que pode ser utilizado para encontrar estimativas para os parâmetros de uma distribuição BCS é o da maximização da verossimilhança (Sen et al., 2010, p.57). Seja Y uma variável que segue uma distribuição BCS. Considere uma amostra aleatória de tamanho N de Y, $Y_1, Y_2, \ldots, Y_N$, e sejam $y_1, y_2, \ldots, y_N$, os valores observados. De (5.8) o logaritmo da função de verossimilhança pode ser escrito como

$$\begin{aligned}
\ell(\mu, \sigma, \nu) &= (\nu - 1)\sum_{i=1}^{N} \log y_i - N\nu \log \mu - N \log \sigma + \sum_{i=1}^{N} \log f_S(z_i) - N \log F_S\left(\frac{1}{\sigma|\nu|}\right) \\
&= (\nu - 1)\sum_{i=1}^{N} \log y_i - N\nu \log \mu - N \log \sigma + \sum_{i=1}^{N} \log r(z_i^2) - N \log R\left(\frac{1}{\sigma|\nu|}\right),
\end{aligned}$$

em que $z_i = Z(y_i, \mu, \sigma, \nu)$. As derivadas de primeira ordem da função de verossimilhança em relação a cada um dos parâmetros são dadas por

$$\frac{\partial \ell}{\partial \mu} = -\frac{N\nu}{\mu} - \sum_{i=1}^{N} \omega_i z_i \frac{\partial z_i}{\partial \mu},$$

$$\frac{\partial \ell}{\partial \sigma} = \begin{cases} -\dfrac{N}{\sigma} - \displaystyle\sum_{i=1}^{N} \omega_i z_i \frac{\partial z_i}{\partial \sigma} + \frac{N}{\sigma^2|\nu|} \frac{f_S\left(\frac{1}{\sigma|\nu|}\right)}{F_S\left(\frac{1}{\sigma|\nu|}\right)}, & \text{se} \quad \nu \neq 0, \\ -\dfrac{N}{\sigma} - \displaystyle\sum_{i=1}^{N} \omega_i z_i \frac{\partial z_i}{\partial \sigma}, & \text{se} \quad \nu = 0, \end{cases}$$

$$\frac{\partial \ell}{\partial \nu} = -\sum_{i=1}^{N} \log y_i - N \log \mu - \sum_{i=1}^{N} \omega_i z_i \frac{\partial z_i}{\partial \nu} - \frac{N}{\sigma \nu^2} \text{sinal}(\nu) \frac{f_S\left(\frac{1}{\sigma|\nu|}\right)}{F_S\left(\frac{1}{\sigma|\nu|}\right)}, \quad \text{se} \quad \nu \neq 0;$$

se $\nu = 0$ a última parcela em $\partial \ell / \partial \nu$ deve ser substituída por seu limite quando $\nu \to 0$. As estimativas de máxima verossimilhança de μ e σ, para um valor fixado de ν, são soluções do sistema de equações

$$\mu = \begin{cases} \dfrac{1}{(N\sigma\nu)^{1/\nu}} \left[\displaystyle\sum_{i=1}^{N} \omega_i z_i y_i^{\nu} \right]^{1/\nu}, & \text{se} \quad \nu \neq 0, \\ \left[\displaystyle\prod_{i=1}^{N} y_i^{\omega_i} \right]^{1/\sum_{i=1}^{N} \omega_i}, & \text{se} \quad \nu = 0, \end{cases}$$

$$\sigma = \begin{cases} \dfrac{1}{N} \displaystyle\sum_{i=1}^{N} \omega_i z_i \frac{1}{\nu} \left[\left(\frac{y_i}{\mu} \right)^{\nu} - 1 \right] + \frac{1}{|\nu|} \frac{f_S\left(\frac{1}{\sigma|\nu|}\right)}{F_S\left(\frac{1}{\sigma|\nu|}\right)}, & \text{se} \quad \nu \neq 0, \\ \dfrac{1}{N} \displaystyle\sum_{i=1}^{N} \omega_i z_i \log\left(\frac{y_i}{\mu} \right), & \text{se} \quad \nu = 0, \end{cases}$$

em que $\omega_i = \omega(z_i)$, com $\omega(z) = -2r'(z^2)/r(z^2)$ uma função de peso dependente de $r(\cdot)$. Nota-se que a estimação por máxima verossimilhança dos parâmetros envolve médias ponderadas e geométricas das contribuições de cada observação y_i com pesos $\omega(z_i)$. A Tabela (5.1) apresenta $\omega(z)$ para diversas distribuições na classe BCS. A depender da escolha de $r(\cdot)$, tal função de peso pode ser decrescente em Y (por exemplo, para a distribuição t-Student), o que significa dizer que observações discrepantes terão peso pequeno na estimação dos parâmetros. Nesse sentido, o procedimento de estimação de μ e σ é robusto.

Tabela 5.1: Funções de peso para algumas distribuições simétricas.

Distribuições	$\omega(z)$
Normal	1
Exponencial dupla	$\sqrt{2}/\lvert z^2\rvert^{1/2}$
Exponencial Potência	$\tau\lvert z^2\rvert^{\tau/2-1}/(2\lvert p(\tau)\rvert^{\tau})$
Cauchy	$2/(1+z^2)$
t-Student	$(\tau+1)/(\tau+z^2)$
Logística Tipo I	$(-2(\exp\{-z^2\}-1))/(\exp\{-z^2\}+1)$
Logística Tipo II	$(\exp\{-\sqrt{z^2}\}-1)/(z(\exp\{-\sqrt{z^2}\}+1))$
Slash canônica	$2/z^2 - (\exp\{-z^2/2\})/(1-\exp\{-z^2/2\})$
Slash [3]	$\Psi((q+3)/2, z^2/2)/\Psi((q+1)/2, z^2/2)(2/z^2)$

O conjunto de equações de verossimilhança não apresenta solução explícita e pode ser resolvido por métodos iterativos. No *software R*, a rotina `gamlss` apresenta a

[3] $\Psi(\cdot,\cdot)$ função gama incompleta definida na Seção 5.1

implementação das distribuições Box-Cox *t* (BCT), Cole-Green (BCCG) e exponencial potência (BCPE), as quais são resolvidas através dos métodos de Newton-Raphson, escore de Fisher ou quasi-Newton, através dos algoritmos CG e RS (Rigby & Stasinopoulos, 2005).

Intervalos de confiança e testes de hipóteses podem ser realizados sobre os parâmetros das distribuições da família Box-Cox simétricas utilizando os estimadores de máxima verossimilhança (Sen et al., 2010, p.245). Desse modo, considerando $\boldsymbol{\theta} = (\theta_1, \theta_2, \theta_3) = (\mu, \sigma, \nu)$ o vetor de parâmetros, um intervalo de confiança assintótico para cada componente do vetor de parâmetros, com coeficiente de confiança α, é dado por

$$IC(\theta_k, \alpha) = \left(\widehat{\theta_k} - \Phi^{-1}\left(1 - \frac{\alpha}{2}\right)\frac{\widehat{\sigma_k}}{\sqrt{N}}; \widehat{\theta_k} + \Phi^{-1}\left(1 - \frac{\alpha}{2}\right)\frac{\widehat{\sigma_k}}{\sqrt{N}}\right),$$

em que $\widehat{\theta_k}$ representa a estimativa de máxima verossimilhança de θ_k, para $k = 1, \ldots, 3$, $\Phi(\cdot)$ é a função de distribuição acumulada de uma distribuição normal padrão e $\widehat{\sigma_k}$ é o k-ésimo elemento da inversa da diagonal da matriz de informação observada $\Sigma(\boldsymbol{\theta}) = \frac{\partial^2 \ell(\theta)}{\partial\theta\partial\theta^\top}$ avaliada em $\widehat{\boldsymbol{\theta}}$, a estimativa de máxima verossimilhança de $\boldsymbol{\theta}$.

Os testes assintóticos de Wald, razão de verossimilhanças e escore podem ser feitos para testar hipóteses de interesse sobre os parâmetros (Sen et al., 2010, p.263).

Adicionalmente, intervalos de confiança podem ser estabelecidos para os quantis das distribuições Box-Cox simétricas. Para tal procedimento, uma expansão em série de Taylor (Sen et al., 2010, p.17-18) por meio de um polinômio de primeira ordem para aproximação do quantil pode ser realizada, utilizando o ponto dado pelas estimativas de máxima verossimilhança do vetor de parâmetros da distribuição BCS. Assim,

$$\begin{aligned}\widehat{y_\alpha} &\approx y_\alpha + (\widehat{\mu} - \mu)(1 + \widehat{\sigma\nu} z_\alpha)^{\frac{1}{\widehat{\nu}}} + (\widehat{\sigma} - \sigma)\widehat{\mu} z_\alpha (1 + \widehat{\sigma\nu} z_\alpha)^{\frac{1}{\widehat{\nu}} - 1} + \\ &+ \frac{(\widehat{\nu} - \nu)\widehat{\mu}(1 + \widehat{\sigma\nu} z_\alpha)^{\frac{1}{\widehat{\nu}}}}{\widehat{\nu}}\left[-\frac{\log \widehat{\mu}(1 + \widehat{\sigma\nu} z_\alpha)}{\widehat{\nu}} + \frac{\widehat{\sigma} z_\alpha}{(1 + \widehat{\sigma\nu} z_\alpha)}\right],\end{aligned}$$

e por Sen et al. (2010, p.210), tem-se que

$$\mathrm{Var}(\widehat{y_\alpha}) \approx \mathrm{E}^\top \Sigma(\widehat{\boldsymbol{\theta}})\mathrm{E},$$

sendo $E^\top = (\epsilon_1, \epsilon_2, \epsilon_3)^\top$ composta pelos elementos

$$\begin{aligned}
\epsilon_1 &= \frac{\partial y_\alpha}{\partial \mu} = (1 + \widehat{\sigma\nu} z_\alpha)^{\frac{1}{\widehat{\nu}}}, \\
\epsilon_2 &= \frac{\partial y_\alpha}{\partial \sigma} = \widehat{\mu} z_\alpha (1 + \widehat{\sigma\nu} z_\alpha)^{\frac{1}{\widehat{\nu}} - 1}, \\
\epsilon_2 &= \frac{\partial y_\alpha}{\partial \nu} = \frac{\widehat{\mu}(1 + \widehat{\sigma\nu} z_\alpha)^{\frac{1}{\widehat{\nu}}}}{\widehat{\nu}} \left[-\frac{\log \widehat{\mu}(1 + \widehat{\sigma\nu} z_\alpha)}{\widehat{\nu}} + \frac{\widehat{\sigma} z_\alpha}{(1 + \widehat{\sigma\nu} z_\alpha)} \right],
\end{aligned}$$

e $\Sigma(\boldsymbol{\theta})$ a matriz de informação observada.

5.3.4 Momentos, assimetria e curtose

Se $Y \sim BCS(\mu, \sigma, \nu; r)$ e k é inteiro, supondo que os momentos de uma distribuição S simétrica existam e sejam finitos, então, se $\nu \neq 0$

$$\mathrm{E}(Y^k) = \mu^k \mathrm{E}\left((1 + \sigma\nu S)^{\frac{k}{\nu}} I_{A(\sigma,\nu)}(S) \right), \tag{5.10}$$

em que $I_{A(\sigma,\nu)}(\cdot)$ é a função indicadora do conjunto $A(\sigma, \nu)$, com $A(\sigma, \nu)$ dado em (5.5), e S é uma variável aleatória que segue uma distribuição simétrica padrão. Se $\nu = 0$, tem-se que

$$\lim_{\nu \to 0} (1 + \sigma\nu s)^{\frac{k}{\nu}} = \exp\{k\sigma s\},$$

para todo $s \in \mathbb{R}$, e assim obtém-se

$$\mathrm{E}(Y^k) = \mu^k \mathrm{E}(\exp\{k\sigma S\}). \tag{5.11}$$

Assim, quando a região de truncamento é desprezível ou $\nu = 0$, os momentos podem ser encontrados através da função geradora de momentos de uma variável que segue uma distribuição simétrica padrão (vide (5.2)).

Para o caso em que $\nu \neq 0$, aproximações para a média e a variância podem ser calculadas, por meio de uma expansão em Taylor (Sen et al., 2010, p.17-18); detalhes são relatados no Apêndice A.5. Desse modo, considerando a região de truncamento desprezível e valores pequenos para σ, tem-se que

$$\mathrm{E}(Y) \approx \mu \left[1 + \frac{\sigma^2(1 - \nu)}{2} \right] \quad \text{e} \quad \mathrm{Var}(Y) \approx \mu^2 \sigma^2 \left[1 - \frac{\sigma^2(1 - \nu)^2}{4} \right].$$

Assim, para $\nu \neq 0$, sendo região de truncamento desprezível e σ pequeno, o valor esperado será dado aproximadamente por μ, o qual, nestas condições, coincide com a mediana, e a variância será aproximadamente igual à de uma variável aleatória

simétrica com parâmetro de escala dado por $\mu^2\sigma^2$.

Os valores aproximados, considerando a região de truncamento, referentes ao primeiro e segundo momentos de variáveis que seguem a distribuição Box-Cox Cole-Green estão explicitados no Apêndice A.5.

A partir das definições de momentos dadas em (5.10) e (5.11), as medidas de assimetria γ_1 e curtose γ_2 definidas por

$$\gamma_1 = \frac{E[(Y - E(Y))^3]}{[E(Y - E(Y))^2]^{3/2}} = \frac{E(Y^3) - 3E(Y^2)E(Y) + 2[E(Y)]^3}{[E(Y^2) - [E(Y)]^2]^{3/2}}$$

e

$$\gamma_2 = \frac{E[(Y - E(Y))^4]}{[E(Y - E(Y))^2]^2} = \frac{E(Y^4) - 4E(Y)E(Y^3) + 6[E(Y)]^2E(Y^2) - 3[E(Y)]^4}{[E(Y^2) - [E(Y)]^2]^2},$$

respectivamente, podem ser explicitadas. Os valores aproximados dos momentos de terceira e quarta ordem para variáveis que seguem as distribuições Box-Cox Cole-Green para os cálculos de tais coeficientes estão explícitos no Apêndice A.6.

5.3.5 Peso da Cauda

Em teoria de valores extremos é muito comum a análise de uma medida conhecida como índice da cauda, a qual avalia a taxa de decaimento das caudas de distribuições. A seguir, algumas definições que são utilizadas nessa área são apresentadas e um breve estudo sobre o decaimento da cauda direita de algumas distribuições que compõem a classe das Box-Cox simétricas é apresentado.

Usualmente para calcular o índice da cauda utiliza-se um limite, que é uma razão entre probabilidades da distribuição ser maior que valores localizados no extremo da cauda direita da distribuição.

Segundo Resnick (2007, p.20), uma função é dita de variação regular se assintoticamente ela se comporta como uma função potência. Por definição, uma função $M : \mathbb{R}^+ \to \mathbb{R}^+$ é de variação regular no infinito com índice ϱ, denota-se $M \in RV_\varrho$, se, para $y > 0$,

$$\lim_{t\to\infty} \frac{M(ty)}{M(t)} = y^\varrho,$$

em que ϱ é chamado de expoente de variação ou índice de variação regular. A função canônica com índice de variação ϱ é y^ϱ. Se $\varrho = 0$, a função M é dita de variação lenta.

Uma função $M : \mathbb{R}^+ \to \mathbb{R}^+$ é dita de variação rápida ou de variação regular com

índice $-\infty$, e denota-se $M \in RV_{-\infty}$, se para todo $y > 0$,

$$\lim_{t\to\infty} \frac{M(ty)}{M(t)} = y^{-\infty} := \begin{cases} \infty, & \text{se} \quad 0 < y < 1, \\ 1, & \text{se} \quad y = 1, \\ 0, & \text{se} \quad y > 1. \end{cases}$$

Uma distribuição com função de distribuição F é dita ter cauda direita pesada se $\overline{F} := 1 - F$ é uma função de variação regular com índice de variação regular negativo, $\varrho < 0$, dado por $\varrho = -1/\varsigma$, isto é,

$$\lim_{t\to\infty} \frac{\overline{F}(ty)}{\overline{F}(t)} = y^{-\frac{1}{\varsigma}}$$

(Rodrigues & Gomes, 2009). O parâmetro ς é denominado índice da cauda.

Utilizando a regra de L'Hôpital, o limite que define o índice da cauda pode ser reescrito em função da densidade de probabilidade f como

$$\lim_{t\to\infty} \frac{\overline{F}(ty)}{\overline{F}(t)} = \lim_{t\to\infty} \frac{1 - F(ty)}{1 - F(t)} = y \lim_{t\to\infty} \frac{f(ty)}{f(t)}.$$

Desse modo, se uma variável aleatória $W \sim S(\mu, \sigma^2; r)$, este limite pode ser escrito através de sua função geradora de densidades $r(\cdot)$ como

$$\mathcal{L}_S(w; r) = w \lim_{t\to\infty} \frac{r\left(\left(\frac{tw-\mu}{\sigma}\right)^2\right)}{r\left(\left(\frac{t-\mu}{\sigma}\right)^2\right)}. \tag{5.12}$$

Por de Haan (1970, Corolário 1.2.1(2 e 3)), tem-se ainda que o índice da cauda de uma distribuição simétrica $S(\mu, \sigma; r)$ é invariante sob transformação de locação-escala. Então (5.12) pode ser obtido independente de μ e σ através da expressão

$$\mathcal{L}_S(w; r) = w \lim_{t\to\infty} \frac{r(t^2 w^2)}{r(t^2)}.$$

Para uma variável aleatória que segue uma distribuição $L \sim LS(\eta, \sigma^2; r)$, o índice da cauda é dado por

$$\mathcal{L}_{LS}(l; r) = \lim_{t\to\infty} \frac{r\left(\left(\frac{\log tl-\eta}{\sigma}\right)^2\right)}{r\left(\left(\frac{\log t-\eta}{\sigma}\right)^2\right)}. \tag{5.13}$$

Assim, para uma variável aleatória $Y \sim BCS(\mu, \sigma, \nu; r)$ o índice pode ser calculado

através das relações dadas em (5.12) e (5.13) como

$$\mathcal{L}_{BCS}(y;r) = \begin{cases} \mathcal{L}_S(y^\nu;r), & \text{se} \quad \nu > 0, \\ y^\nu, & \text{se} \quad \nu < 0, \\ \mathcal{L}_{LS}(y;r), & \text{se} \quad \nu = 0, \end{cases} \tag{5.14}$$

detalhes podem ser vistos no Apêndice A.7.

A Tabela 5.2 apresenta o índice da cauda para algumas distribuições simétricas e Box-Cox simétricas [4]. Os índices de cauda relatados foram calculados com o auxílio do *software* Maple 13, detalhes em *http://www.maplesoft.com*.

Tabela 5.2: Índice da cauda de algumas distribuições simétricas e Box-Cox simétricas.

Distribuições	Simétricas	BCS ($\nu > 0$)	BCS ($\nu = 0$)	BCS ($\nu < 0$)
Normal	0	0	0	$1/\lvert\nu\rvert$
Exponencial dupla	0	0	$\sigma/\sqrt{2}$	$1/\lvert\nu\rvert$
Exponencial Potência				
$\tau > 1$	0	0	0	$1/\lvert\nu\rvert$
$\tau = 1$	0	0	$\sigma/\sqrt{2}$	$1/\lvert\nu\rvert$
$\tau < 1$	0	0	1	$1/\lvert\nu\rvert$
Cauchy	1	$1/\nu$	1	$1/\lvert\nu\rvert$
t-Student	$1/\tau$	$1/(\nu\tau)$	1	$1/\lvert\nu\rvert$
Logística Tipo I	0	0	0	$1/\lvert\nu\rvert$
Logística Tipo II	0	0	σ	$1/\lvert\nu\rvert$
Slash canônica	1	$1/\nu$	1	$1/\lvert\nu\rvert$
Slash	$1/q$	$1/(\nu q)$	1	$1/\lvert\nu\rvert$

Na Tabela 5.2 tem-se alguns exemplos da relação dada em (5.14). Para $\nu > 0$, as distribuições Box-Cox normal (Cole-Green), exponenciais dupla e potência e logísticas tipos I e II apresentam índice da cauda igual a zero, indicando uma variação rápida, as demais apresentam uma variação regular, explicitando a presença de cauda direita pesada, destaque para as distribuições Box-Cox t e slash, nas quais um parâmetro extra controla o peso nas caudas, para cada valor fixado de ν. A coluna "BCS ($\nu = 0$)" refere-se às distribuições log-simétricas e, assim, observa-se que as distribuições log-normal, log-logística tipo I e exponencial potência quando $\tau > 1$ apresentam uma variação rápida, já as distribuições log-t-Student, log-slash e seus casos particulares, log-Cauchy e log-slash canônica, e a distribuição log-exponencial potência quando $\tau < 1$, são de variação lenta; as demais têm índices de cauda positivos, indicando cauda direita pesada. É importante destacar que, para as distribuições log-simétricas, o parâmetro extra (por exemplo, o número de graus de liberdade no caso da distribuição log-t-Student) não tem efeito no controle do índice da cauda. Para $\nu < 0$, todas as

[4] O cálculo do índice da cauda da distribuição log-exponencial potência com $\tau > 1$ foi feito apenas para $\tau \in \mathbb{Q}$. Para a distribuição slash, apenas para $q \in \mathbb{N}^*$.

distribuições apresentam cauda direita pesada e o índice da cauda depende apenas do parâmetro de transformação ν.

Uma abordagem alternativa para comparar o peso da cauda de distribuições é considerada por Rigby et al. (2014, Capítulo 12). Aqui, o enfoque será dado somente sobre peso da cauda direita da distribuição. Se duas variáveis aleatórias Y_1 e Y_2, com funções densidade de probabilidade $f_{Y_1}(y)$ e $f_{Y_2}(y)$, respectivamente, e $\lim_{y\longrightarrow\infty} f_{Y_1}(y) = \lim_{y\longrightarrow\infty} f_{Y_2}(y) = 0$, então Y_2 tem cauda mais pesada do que Y_1 se e somente se $\lim_{y\longrightarrow\infty}(\log f_{Y_2}(y) - \log f_{Y_1}(y)) = \infty$. Os autores apresentam três possibilidades para o comportamento assintótico (y grande) do logaritmo da função densidade de probabilidade: $-k_2(\log y)^{k_1}$, $-k_4 y^{k_3}$ ou $-k_6 \exp(k_5 y)$, com $k's$ positivos. As três formas são decrescentes de acordo com o peso da cauda, assim a primeira forma tem cauda mais pesada que a segunda e a terceira apresenta a cauda mais leve do que as duas anteriores. Para a primeira forma, o decrescimento de k_1 resulta em aumento do peso da cauda, enquanto que o decrescimento de k_2 para um k_1 fixo resulta em aumento do peso da cauda. Similarmente para as outras duas formas.

Segundo Rigby et al. (2014), o peso da cauda direita das distribuições pode ser dividido em quatro formas: cauda não pesada: $k_3 \geq 1$; cauda pesada, isto é, mais pesada do que de qualquer distribuição exponencial, porém mais leve do que a cauda de qualquer distribuição do "tipo Pareto": $k_1 > 1$ ou $0 < k_3 < 1$; cauda do "tipo Pareto": $k_1 = 1$ e $k_2 > 1$; cauda mais pesada do que qualquer distribuição "tipo Pareto": $k_1 = 1$ e $k_2 = 1$.

A Tabela 5.3 mostra a forma assintótica da cauda direita do logaritmo da função de densidade para algumas distribuições simétricas e Box-Cox simétricas. Nota-se que, quando $\nu > 0$, as distribuições Box-Cox t e slash têm cauda pesada como a de uma distribuição do "tipo Pareto" com um parâmetro extra controlando o peso da cauda para cada valor fixo de ν; as distribuições Box-Cox Cole-Green, logísticas tipos I e II e exponencial potência podem ter caudas direitas não pesadas ou caudas mais pesadas, porém mais leves que a cauda da distribuição do "tipo Pareto", a depender do parâmetro de transformação; no caso da distribuição Box-Cox exponencial potência para $\tau \neq 1$, o peso da cauda depende do produto deste parâmetro com um parâmetro extra. Quando $\nu = 0$, as distribuições log-normal, log-logística tipo I e log-exponencial potência para $\tau > 1$ têm caudas direita pesadas, mas mais leves do que uma distribuição do "tipo Pareto", enquanto as distribuições log-t-Student, log-exponencial potência para $\tau < 1$ e log-slash têm caudas mais pesadas que a cauda das distribuições do "tipo Pareto". As distribuições log-exponencial dupla (log-exponencial potência com $\tau = 1$) e log-logística tipo II apresentam um comportamento de cauda como o de uma distribuição do "tipo Pareto". Quando $\nu < 0$, todas as distribuições Box-Cox simétricas apresentadas na Tabela 5.3 têm uma cauda direita do "tipo Pareto".

Tabela 5.3: Comportamento assintótico da cauda direita do logaritmo da função de densidade de algumas distribuições simétricas e Box-Cox simétricas.

Distribuições	Simétricas	BCS ($\nu > 0$)	BCS ($\nu = 0$)	BCS ($\nu < 0$)
Normal	$k_3 = 2, k_4 = 1/(2\sigma^2)$	$k_3 = 2\nu, k_4 = 1/(2\mu^{2\nu}\sigma^2\nu^2)$	$k_1 = 2, k_2 = 1/(2\sigma^2)$	$k_1 = 1, k_2 = \lvert\nu\rvert + 1$
Exponencial dupla	$k_3 = 1, k_4 = \sqrt{2}/\sigma$	$k_3 = \nu, k_4 = \sqrt{2}/(\mu^\nu\sigma\nu)$	$k_1 = 1, k_2 = \sqrt{2}/\sigma + 1$	$k_1 = 1, k_2 = \lvert\nu\rvert + 1$
Exponencial potência				
$\tau > 1$	$k_3 = \tau, k_4 = 1/(2p(\tau)^\tau\sigma^\tau)$	$k_3 = \nu\tau, k_4 = 1/(2p(\tau)^\tau\mu^{\nu\tau}\sigma^\tau\nu^\tau)$	$k_1 = \tau, k_2 = 1/(2p(\tau)^\tau\sigma^\tau)$	$k_1 = 1, k_2 = \lvert\nu\rvert + 1$
$\tau = 1$	$k_3 = 1, k_4 = \sqrt{2}/\sigma$	$k_3 = \nu, k_4 = \sqrt{2}/(\mu^\nu\sigma\nu)$	$k_1 = 1, k_2 = \sqrt{2}/\sigma + 1$	$k_1 = 1, k_2 = \lvert\nu\rvert + 1$
$\tau < 1$	$k_3 = \tau, k_4 = 1/(2p(\tau)^\tau\sigma^\tau)$	$k_3 = \nu\tau, k_4 = 1/(2p(\tau)^\tau\mu^{\nu\tau}\sigma^\tau\nu^\tau)$	$k_1 = 1, k_2 = 1$	$k_1 = 1, k_2 = \lvert\nu\rvert + 1$
Cauchy	$k_1 = 1, k_2 = 2$	$k_1 = 1, k_2 = \nu + 1$	$k_1 = 1, k_2 = 1$	$k_1 = 1, k_2 = \lvert\nu\rvert + 1$
t-Student	$k_1 = 1, k_2 = \tau + 1$	$k_1 = 1, k_2 = \nu\tau + 1$	$k_1 = 1, k_2 = 1$	$k_1 = 1, k_2 = \lvert\nu\rvert + 1$
Logística tipo I	$k_3 = 2, k_4 = 1/\sigma^2$	$k_3 = 2\nu, k_4 = 1/(\mu^{2\nu}\sigma^2\nu^2)$	$k_1 = 2, k_2 = 1/\sigma^2$	$k_1 = 1, k_2 = \lvert\nu\rvert + 1$
Logística tipo II	$k_3 = 1, k_4 = 1/\sigma$	$k_3 = \nu, k_4 = 1/(\mu^\nu\sigma\nu)$	$k_1 = 1, k_2 = 1/\sigma + 1$	$k_1 = 1, k_2 = \lvert\nu\rvert + 1$
Slash canônica	$k_1 = 1, k_2 = 2$	$k_1 = 1, k_2 = \nu + 1$	$k_1 = 1, k_2 = 1$	$k_1 = 1, k_2 = \lvert\nu\rvert + 1$
Slash ($q \in \mathbb{N}^*$)	$k_1 = 1, k_2 = q + 1$	$k_1 = 1, k_2 = \nu q + 1$	$k_1 = 1, k_2 = 1$	$k_1 = 1, k_2 = \lvert\nu\rvert + 1$

Nota: Cauda não pesada: $k_3 \geq 1$; cauda pesada, porém mais leve do que a cauda de qualquer distribuição do "tipo Pareto": $k_1 > 1$ ou $0 < k_3 < 1$; cauda do "tipo Pareto": $k_1 = 1$ e $k_2 > 1$; cauda mais pesada do que qualquer distribuição "tipo Pareto": $k_1 = 1$ e $k_2 = 1$.

5.4 Comparação entre classes de distribuições Box-Cox simétricas e simétricas transformadas

Uma estratégia usual para lidar com dados contínuos no qual tem-se uma assimetria positiva é empregar uma transformação de Box-Cox, e assumir que os dados transformados seguem uma distribuição normal.[5] A distribuição normal pode ser substituída por uma distribuição da classe das distribuições simétricas contínuas após a transformação de Box-Cox; detalhes em Cordeiro & Andrade (2011). Tais distribuições são intituladas como simétricas transformadas. Formalmente, esta abordagem não corresponde a assumir uma distribuição coerente para os dados, pois o suporte da variável transformada não é toda reta real, a menos que o parâmetro de transformação seja igual a zero. De fato, o suporte da variável transformada é de $(-1/\nu, \infty)$ se $\nu > 0$ e $(-\infty, -1/\nu)$ se $\nu < 0$. Além disso, os parâmetros do modelo são interpretados como características dos dados transformados, e não dos dados originais. A abordagem da classe de distribuições BCS não apresenta tais limitações: uma distribuição genuína é assumida para os dados e os parâmetros são interpretáveis em termos de características dos dados originais, e não com os dados transformados.

Na próxima seção, uma comparação entre abordagens alternativas é feita por meio de uma aplicação em dados reais. Na maioria dos casos, será visto que a classe Box-Cox simétrica de distribuições proporciona melhor ajuste do que a classe das simétricas transformadas.

5.5 Aplicações e comparações entre enfoques alternativos

Esta seção apresenta aplicações das distribuições Box-Cox simétricas na análise do consumo de micro e macronutrientes. Os dados utilizados são compostos por 368 recordatórios 24 horas (R24h), referentes ao primeiro consumo relatado pelos 136 homens e 232 mulheres que compõe a amostra descrita com detalhes no Capítulo 3. A escolha pelo primeiro R24h é usual na área nutricional quando deseja-se estudar a distribuição de consumo de alguns nutrientes, em especial dos macronutrientes, que são nutrientes necessários em grande quantidade para o bom funcionamento do organismo, como por exemplo, os carboidratos e as proteínas. Nesta seção, as distribuições foram ajustadas para os consumos dos 22 micronutrientes, já explorados com outra abordagem no Capítulo 3, e de 11 macronutrientes.

Para cada nutriente, assume-se que os dados $Y_1, \ldots, Y_n$ são independentes. Os

[5] A transformação Box-Cox é dada por (5.4) com $\mu = \sigma = 1$.

seguintes modelos foram ajustados para os dados: Box-Cox *t* (BCT) com o parâmetro referente ao número de graus de liberdade fixado em 4 e sendo estimado a partir dos dados; Box-Cox Cole-Green (BCCG), o qual corresponde ao modelo BCT quando $\tau \to \infty$; modelos *skew* normal (SN) e *t* (ST) (Azzalini, 2005); e os modelos simétricos transformados com erros normais (TN) e *t* (TT) (Cordeiro & Andrade, 2011). Para os modelos ST e TT, foi considerado somente o caso no qual o número de graus de liberdade é fixo, e novamente considerou-se $\tau = 4$, porque ocorreram problemas numéricos na estimação de τ juntamente com os demais parâmetros dos modelos. Em todos os casos, o método utilizado foi o de máxima verossimilhança. Para tal procedimento, a rotina presente no `gamlss`, na linguagem de programação *R* foi usada para encontrar as estimativas dos parâmetros das distribuições BCCG, BCT, SN e ST. Para as distribuições TN e TT, a função optim do *R* e o PROC NLP do programa SAS (SAS, 1990) foram utilizados.

As medidas utilizadas para medir a qualidade de ajuste dos modelos foram o critério de Akaike e as medidas propostas por Anderson-Darling (AD, ADR e AD2R) (Luceño, 2005). AD é uma medida global de falta de ajuste, enquanto que ADR e AD2R são medidas mais sensíveis para falta de ajuste na cauda direita de uma distribuição, sendo que a AD2R dá um peso maior na cauda direita do que a ADR, detalhes no Apêndice A.8.

As Tabelas de 5.4 a 5.8 apresentam os critérios de Akaike e Anderson-Darling para todos os modelos ajustados para dados de consumo de 22 e 11 micro e macronutrientes, respectivamente. As células em branco nas tabelas indicam que o algoritmo utilizado na estimação por máxima verossimilhança não convergiu ou produziu estimativas não realistas. As tabelas mostram informações importantes. Primeiramente, os conjuntos de dados abrangem uma ampla gama de distribuições, proporcionando distribuições de caudas mais leves à mais pesadas. Isso pode ser visto através dos valores estimados pelo parâmetro referente ao número de graus de liberdade no modelo Box-Cox *t*, o qual variou de $2,2$ a $187,4$. Em segundo lugar, não houve problemas de convergência no ajuste dos modelos Box-Cox *t* (com número de graus de liberdade fixo ou estimado), simétricos transformados normal e *t* com $\tau = 4$. A estimação por máxima verossimilhança sob o modelo *skew* normal não alcançou convergência em 14 casos, seguido pelo modelo Box-Cox Cole-Green (10 casos) e o *skew t* (9 casos). Em terceiro lugar, o modelo de Box-Cox *t* com τ estimado apresentou um desempenho melhor do que com τ fixado em quase todos os casos. Em quarto lugar, de acordo com o critério de Akaike para os dados de micronutrientes, o modelo de Box-Cox *t* com τ estimado conseguiu o melhor ajuste em 12 casos, seguido pelo modelo Box-Cox *t* com τ fixado (5 casos). O mesmo padrão foi observado para os dados de macronutrientes. Em quinto lugar, de acordo com os critérios de Anderson-Darling, em todos os casos, os modelos Box-Cox Cole-Green, *skew* normal e normal transformado não apresentaram o melhor

ajuste. Em contraste, o modelo de Box-Cox t com τ estimado foi o melhor modelo na maioria dos casos. No geral, o modelo de Box-Cox t com τ estimado apresentou um desempenho melhor do que os demais.

Tabela 5.4: Critérios de Akaike para comparação entre o ajuste das distribuições Box-Cox t (BCT) e Cole-Green (BCCG), *skew* normal (SN) e t (ST) e simétricas transformadas normal (TN) e t (TT) para distribuição de micronutrientes.

Micronutriente (m)	τ estimado		$\tau = 4$			$\tau \longrightarrow \infty$		
	$\widehat{\tau}$	BCT	BCT	ST	TT	BCCG	SN	TN
Vitamina A (mcg)	7,2	5807,4	5810,5	5825,6	5810,6			5822,7
Vitamina D (mcg)	6,8	1688,8	1688,7	1708,3	1693,7		1908,5	1698,9
Vitamina E (mg)	6,9	1812,8	1814,6	1816,5	1814,6	1824,3		1824,3
Vitamina K (mcg)	7,8	4354,3	4358,5		4358,6			4368,4
Vitamina C (mg)	2,2	4022,5	4030,9		4034,2			4120,9
Vitamina B1 (mg)	6,8	709,7	710,9	709,7	710,9	720,9		720,9
Vitamina B2 (mg)	6,5	701,6	702,7	700,2	702,7	716,1		716,1
Vitamina B3 (mg)	6,8	2722,4	2723,3	2724,6	2723,3	2730,8	2920,3	2730,8
Vitamina B6 (mg)	50,2	853,2	863,2	861,0	863,2	851,3	945,9	851,3
Vitamina B12 (mcg)	2,5	1782,6	1787,1		1787,1			1887,3
Ácido pantotênico (mg)	8,4	1466,0	1470,0	1463,6	1470,0	1474,1	1622,7	1474,1
Folato (mcg)	4,9	4795,4	4794,1	4800,7	4794,1	4823,4		4823,4
Cálcio (mg)	13,3	5311,3	5318,9	5311,2	5318,9	5312,5	5472,2	5312,5
Fósforo (mg)	14,7	5548,5	5557,4	5550,0	5557,4	5549,0	5642,2	5549,0
Magnésio (mg)	8,6	4474,9	4479,0	4475,1	4479,0	4481,5	4637,3	4481,5
Ferro (mg)	5,9	2409,7	2410,1	2408,4	2410,1	2431,9		2431,9
Zinco (mg)	14,5	2185,3	2192,5		2192,5	2185,4	2278,6	2185,4
Cobre (mg)	5,5	566,1	566,0	576,4	566,0			593,6
Selênio (mcg)	5,2	3992,6	3991,9		3991,9			4020,8
Sódio (mg)	4,6	6525,4	6523,7	6542,3	6523,7			6572,3
Potássio (mg)	9,3	6144,5	6150,4		6150,4	6151,9	6300,2	6151,9
Manganês (mg)	2,5	1616,5	1618,9		1624,2			1656,2

Tabela 5.5: Critérios de Akaike para comparação entre os ajustes das distribuições Box-Cox *t* (BCT) e Cole-Green(BCCG), *skew* normal (SN) e *t* (ST) e simétricas transformadas normal (TN) e *t* (TT) para distribuição de macronutrientes.

Macronutrientes (m)	τ estimado		$\tau = 4$			$\tau \longrightarrow \infty$		
	$\widehat{\tau}$	BCT	BCT	ST	TT	BCCG	SN	TN
Proteína (g)	10,1	3659,5	3664,8	3660,9	3664,8	3662,5	3760,1	3662,5
Energia (kcal)	6,1	5861,9	5862,2		5862,3	5876,1	6003,7	5876,1
Fibra (g)	10,0	2652,2	2657,3	2654,4	2657,4	2655,6	2752,5	2655,6
Carboidrato (g)	10,5	4360,0	4367,0		4367,0	4366,5	4455,3	4366,5
Gordura total (g)	13,9	3587,0	3595,0	3591,9	3595,0	3587,5	3824,6	3587,5
Proteína animal (g)	4,9	3514,4	3512,8	3531,3	3515,3	3526,3	3645,2	3526,5
Proteína vegetal (g)	6,6	2963,3	2964,0	2963,1	2964,2	2972,6	3058,9	2972,6
Gordura saturada (g)	187,4	2819,6	2835,2	2823,3	2835,2	2817,7	2991,6	2817,7
Gordura monosaturada (g)	12,6	2857,1	2864,1	2864,1	2864,1	2858,4		2858,4
Gordura polisaturada (g)	7,1	2596,9	2599,7	2597,2	2599,7	2612,4	3020,2	2612,4
Colesterol (mg)	5,8	4724,7	4724,0	4752,0	4725,6		4933,7	4728,5

Tabela 5.6: Critérios de Anderson-Darling para comparação dos ajustes das distribuições Box-Cox *t* (BCT) e Cole-Green (BCCG), *skew* normal (SN) e *t* (ST) e simétricas transformadas normal (TN) e *t* (TT) para o consumo de micronutrientes.

Micronutriente	Medida	$\widehat{\tau}$	$\tau = 4$			$\tau \longrightarrow \infty$		
		BCT	BCT	ST	TT	BCCG	SN	TN
Vitamina A	AD	1,12	1,50	1,93	1,51			1,47
	ADR	0,64	0,78	1,01	0,80			0,94
	AD2R	6,26	5,45	> 100	5,74			> 100
Vitamina D	AD	0,25	0,34	0,93	0,54		11,61	0,64
	ADR	0,11	0,13	0,47	0,27		5,04	0,38
	AD2R	3,40	2,24	4,78	5,61		> 100	> 100
Vitamina E	AD	0,21	0,36	0,29	0,36	0,91		0,91
	ADR	0,13	0,20	0,17	0,20	0,52		0,52
	AD2R	3,39	3,21	11,59	3,21	61,61		61,43
Vitamina K	AD	0,60	0,86		0,86			1,28
	ADR	0,35	0,45		0,45			0,86
	AD2R	55,55	9,42		9,41			> 100
Vitamina C	AD	0,70	0,93		0,86			7,92
	ADR	0,35	0,42		0,40			4,23
	AD2R	6,27	42,83		26,54			> 100
Vitamina B1	AD	0,20	0,21	0,20	0,21	1,02		1,02
	ADR	0,10	0,13	0,11	0,13	0,47		0,47
	AD2R	2,83	4,83	6,26	4,84	> 100		> 100
Vitamina B2	AD	0,24	0,34	0,24	0,34	1,02		1,02
	ADR	0,13	0,18	0,14	0,18	0,46		0,47
	AD2R	2,10	4,12	4,68	4,12	> 100		> 100
Vitamina B3	AD	0,22	0,21	0,34	0,21	1,07	15,21	1,07
	ADR	0,15	0,14	0,22	0,14	0,56	7,28	0,56
	AD2R	3,46	4,98	6,16	4,99	17,27	> 100	17,25
Vitamina B6	AD	0,37	0,31	0,29	0,31	0,45	7,49	0,45
	ADR	0,18	0,17	0,11	0,17	0,21	3,76	0,21
	AD2R	4,12	7,99	4,87	8,10	4,45	> 100	4,45
Vitamina B12	AD	0,36	0,74		0,74			8,69
	ADR	0,26	0,55		0,55			5,08
	AD2R	8,89	20,29		20,53			> 100
Ácido Pantotênico	AD	0,24	0,51	0,27	0,51	0,65	11,14	0,65
	ADR	0,10	0,20	0,16	0,20	0,34	5,41	0,34
	AD2R	2,38	4,66	2,59	4,64	13,78	> 100	13,77
Folato	AD	0,19	0,20	0,25	0,20	1,83		1,82
	ADR	0,12	0,14	0,15	0,14	0,94		0,95
	AD2R	4,62	3,13	75,53	3,13	> 100		> 100
Cálcio	AD	0,28	0,53	0,17	0,53	0,48	12,17	0,48
	ADR	0,11	0,30	0,07	0,30	0,18	5,40	0,18
	AD2R	2,43	8,41	2,47	8,43	7,50	> 100	7,50

Tabela 5.7: Critérios de Anderson-Darling para comparação dos ajustes das distribuições Box-Cox *t* (BCT) e Cole-Green (BCCG), *skew* normal (SN) e *t* (ST) e simétricas transformadas normal (TN) e *t* (TT) para o consumo de micronutrientes.

Micronutriente	Medida	$\widehat{\tau}$	$\tau = 4$			$\tau \longrightarrow \infty$		
		BCT	BCT	ST	TT	BCCG	SN	TN
Fósforo	AD	0,25	0,63	0,34	0,63	0,36	6,85	0,36
	ADR	0,15	0,38	0,22	0,38	0,18	3,13	0,19
	AD2R	2,43	7,80	4,35	7,82	5,59	> 100	5,56
Magnésio	AD	0,31	0,59	0,35	0,59	0,66	10,45	0,66
	ADR	0,18	0,32	0,22	0,32	0,36	4,91	0,36
	AD2R	2,77	4,63	3,30	4,63	25,36	> 100	25,36
Ferro	AD	0,25	0,38	0,13	0,38	1,21		1,21
	ADR	0,10	0,18	0,07	0,18	0,45		0,45
	AD2R	2,35	3,60	31,82	3,61	> 100		> 100
Zinco	AD	0,16	0,30		0,30	0,38	6,53	0,38
	ADR	0,09	0,18		0,18	0,19	3,16	0,19
	AD2R	1,91	6,49		6,63	7,97	> 100	7,97
Cobre	AD	0,37	0,50	0,37	0,50			1,74
	ADR	0,21	0,26	0,25	0,26			0,99
	AD2R	16,14	5,68	> 100	5,69			> 100
Selênio	AD	0,21	0,29		0,29			1,70
	ADR	0,12	0,17		0,17			0,83
	AD2R	7,25	3,12		3,11			> 100
Sódio	AD	0,18	0,20	0,42	0,20			2,28
	ADR	0,09	0,10	0,28	0,10			1,24
	AD2R	8,26	5,46	> 100	5,46			> 100
Potássio	AD	0,34	0,73		0,73	0,57	7,98	0,57
	ADR	0,22	0,43		0,44	0,34	3,54	0,34
	AD2R	10,70	6,27		6,31	> 100	> 100	> 100
Manganês	AD	1,29	1,74		1,22			4,64
	ADR	1,03	1,32		0,91			2,85
	AD2R	52,72	66,35		37,23			56,04

Tabela 5.8: Critérios de Anderson-Darling para comparação dos ajustes das distribuições Box-Cox *t* (BCT) e Cole-Green (BCCG), *skew* normal (SN) e *t* (ST) e simétricas transformadas normal (TN) e *t* (TT) para o consumo de macronutrientes.

Macronutrientes	Medida	$\widehat{\tau}$	$\tau = 4$			$\tau \longrightarrow \infty$		
		BCT	BCT	ST	TT	BCCG	SN	TN
Proteína	AD	0,23	0,45	0,30	0,45	0,54	7,35	0,54
	ADR	0,14	0,27	0,17	0,27	0,30	3,57	0,30
	AD2R	3,60	6,30	3,73	6,34	10,93	> 100	10,93
Energia	AD	0,19	0,23		0,23	1,13	9,42	1,10
	ADR	0,10	0,13		0,13	0,58	4,69	0,54
	AD2R	1,77	3,14		3,15	> 100	> 100	> 100
Fibra	AD	0,21	0,47	0,33	0,47	0,51	7,02	0,51
	ADR	0,11	0,23	0,19	0,24	0,26	3,35	0,26
	AD2R	2,03	5,85	3,66	6,28	11,27	> 100	11,26
Carboidrato	AD	0,26	0,66		0,66	0,39	5,45	0,39
	ADR	0,13	0,38		0,38	0,17	2,32	0,17
	AD2R	5,21	6,96		7,06	> 100	> 100	> 100
Gordura Total	AD	0,41	0,68	0,52	0,68	0,61	14,95	0,62
	ADR	0,24	0,40	0,27	0,40	0,36	6,90	0,36
	AD2R	5,69	6,17	7,50	6,18	14,44	> 100	14,46
Proteína Animal	AD	0,35	0,33	0,34	0,31	1,46	9,09	1,43
	ADR	0,16	0,16	0,12	0,15	0,70	4,44	0,68
	AD2R	3,19	3,79	2,81	4,79	23,33	> 100	22,34
Proteína Vegetal	AD	0,25	0,28	0,15	0,28	1,08	7,09	1,08
	ADR	0,10	0,15	0,07	0,15	0,44	3,43	0,43
	AD2R	1,98	4,45	2,02	4,70	39,66	> 100	39,88
Gordura Saturada	AD	0,16	0,61	0,51	0,61	0,17	11,38	0,17
	ADR	0,08	0,30	0,22	0,30	0,08	4,97	0,08
	AD2R	3,06	8,43	3,51	8,47	3,28	> 100	3,27
Gordura Monosaturada	AD	0,33	0,50	0,54	0,50	0,62		0,62
	ADR	0,11	0,23	0,18	0,23	0,26		0,26
	AD2R	4,40	4,77	4,86	4,85	29,04		29,06
Gordura Polisaturada	AD	0,57	0,90	0,39	0,90	1,02	28,25	1,02
	ADR	0,26	0,41	0,24	0,41	0,51	12,97	0,51
	AD2R	3,52	4,00	61,74	3,99	85,78	> 100	86,57
Colesterol	AD	0,46	0,46	0,63	0,29		14,08	1,16
	ADR	0,29	0,34	0,37	0,16		6,50	0,51
	AD2R	7,55	10,47	6,50	6,39		> 100	11,01

Uma análise sobre os dados referentes aos consumos de proteína animal e energia utilizando as distribuições Box-Cox t, Cole-Green e exponencial potência é apresentada. As Tabelas 5.9 e 5.10 mostram medidas descritivas, estimativas dos parâmetros e medidas de qualidade de ajuste, *box-plot* ajustados (Hubert & Vandervieren, 2008) são apresentados na Figura 5.3. As medidas descritivas e os *box-plot* ajustados indicam alta dispersão, assimetria à direita e observações discrepantes para ambos os conjuntos de dados, ressaltando-se a presença de pontos aberrantes extremos na distribuição de consumo de energia.

Para os dois conjuntos de dados as estimativas de μ e ν sob os três modelos são similares e as de μ são próximas à mediana dos dados. Nota-se ainda que os critérios de Akaike são semelhantes para os modelos BCT e BCPE e ambos menores que para o modelo BCCG. As medidas de Anderson Darling indicam que o modelo BCT produz melhor ajuste, especialmente na cauda. De fato, $AD2R = 3,19$ para o modelo BCT, enquanto que $AD2R = 23,33$ e $AD2R = 3,42$ para os modelos BCCG e BCPE, respectivamente, para os dados de consumo de proteína animal. Padrão similar observou-se para os dados de consumo de energia.

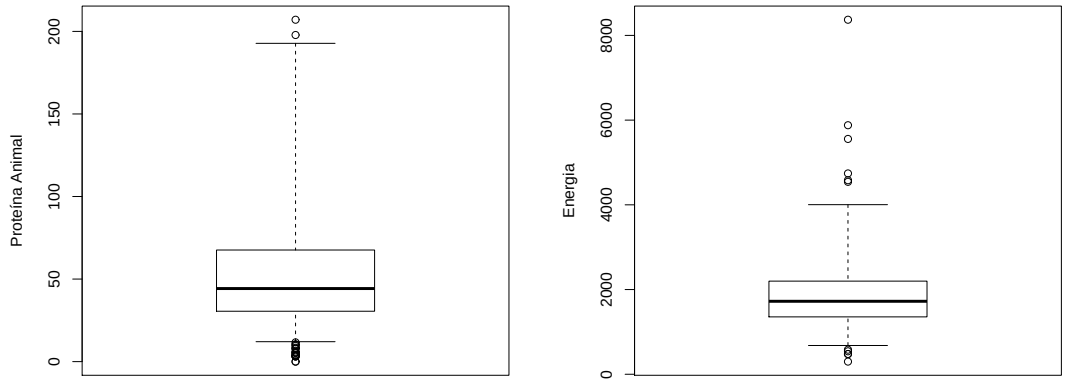


Figura 5.3: $Box - plot$ ajustados do consumo de proteína animal e energia.

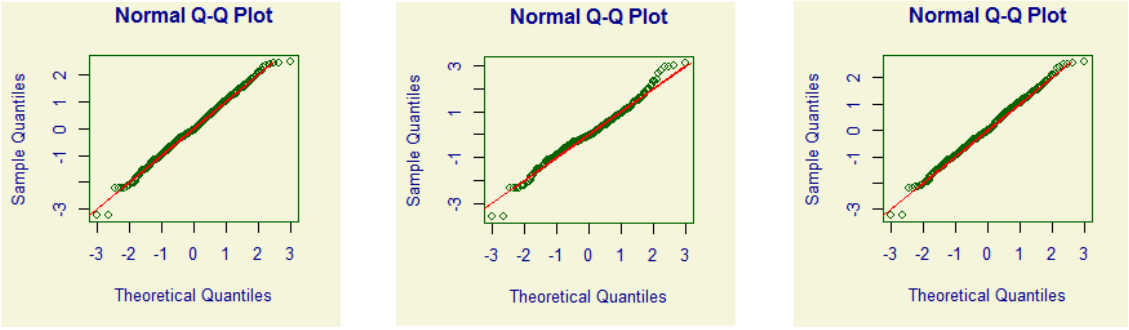
Tabela 5.9: Medidas descritivas, estimativas dos parâmetros dos modelos BCT, BCCG e BCPE; e medidas de comparação AD, ADR e AD2R para distribuição de consumo de proteína animal.

Descritivas (mg)	Mínimo	1ºQuartil	Mediana	Média (DP)	3ºQuartil	Máximo
	0,02	30,51	44,26	52,27 (34,02)	67,59	207,10
Distribuição	BCT		BCCG		BCPE	
Parâmetros	Estimativa	EP	Estimativa	EP	Estimativa	EP
μ	45,74	1,46	46,37	1,50	44,73	1,36
σ	0,55	0,03	0,67	0,02	0,70	0,03
ν	0,42	0,07	0,44	0,06	0,42	0,07
τ	4,90	0,92			1,24	0,12
AIC	3514,39		3526,30		3511,43	
AD	0,35		1,46		0,51	
ADR	0,16		0,70		0,26	
AD2R	3,19		23,33		3,42	

Tabela 5.10: Medidas descritivas, estimativas dos parâmetros dos modelos BCT, BCCG e BCPE; e medidas de comparação AD, ADR e AD2R para distribuição de consumo de energia.

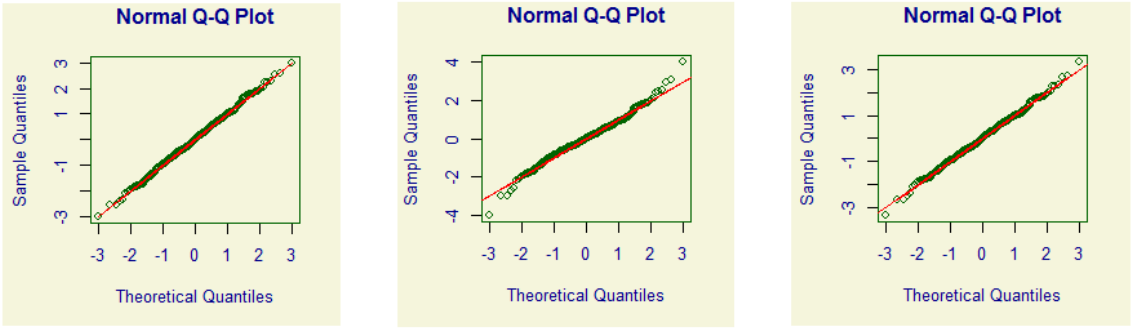
Descritivas (kcal)	Mínimo	1ºQuartil	Mediana	Média (DP)	3ºQuartil	Máximo
	298,80	1356,00	1723,00	1868,00 (838,35)	2197,00	8370,00
Distribuição	BCT		BCCG		BCPE	
Parâmetros	Estimativa	EP	Estimativa	EP	Estimativa	EP
μ	1725,00	34,48	1726,00	36,86	1724,00	34,14
σ	0,34	0,02	0,41	0,02	0,41	0,02
ν	0,05	0,12	0,07	0,10	0,06	0,11
τ	6,14	1,37			1,40	0,14
AIC	5861,85		5876,11		5863,99	
AD	0,19		1,13		0,25	
ADR	0,10		0,58		0,14	
AD2R	1,77		112,25		4,18	

Os resíduos quantílicos propostos por Dunn & Smyth (1996) também foram calculados. As Figuras 5.4 e 5.5 explicitam os resultados. Nota-se claramente uma fuga da normalidade dos resíduos no ajuste do modelo BCCG nos valores extremos. Na Figura 5.4, os comportamentos dos resíduos nos modelos BCT e BCPE se mostraram bem semelhantes, indicando um bom ajuste dos modelos. Já na Figura 5.5, os resíduos do modelo BCT tiveram um comportamento ligeiramente melhor em relação à normalidade do que os encontrados por meio do ajuste da distribuição BCPE.



(a) BCT; (b) BCCG; (c) BCPE.

Figura 5.4: Resíduos quantílicos para o ajuste dos modelos Box-Cox *t*, normal e exponencial potência para distribuição de consumo de proteína animal.



(a) BCT; (b) BCCG; (c) BCPE.

Figura 5.5: Resíduos quantílicos para o ajuste dos modelos Box-Cox *t*, normal e exponencial potência para a distribuição de consumo de energia.

Para os dados de consumo de energia, as estimativas do parâmetro de assimetria (ν) são próximas de zero e os erros padrão são relativamente grandes. Este fato sugere que modelos log-simétricos podem ser adequados. A Tabela 5.11 apresenta as estimativas dos parâmetros e as medidas de qualidade de ajuste para os modelos log-*t*, log-normal e log-exponencial potência. Comparando com os resultados da Tabela 5.10, somente a estimativa de μ sofreu uma pequena alteração, os critérios de Akaike e Anderson Darling foram similares, exceto para a medida AD2R nos modelos log-normal ($AD2R = 48,17$) e log-exponencial potência ($AD2R = 2,70$), indicando uma melhora no ajuste da cauda direita para tais modelos. Os gráficos de resíduos quantílicos foram feitos para os modelos log-simétricos e os comportamentos seguiram o mesmo padrão da Figura 5.5, sendo por este motivo omitidos. De modo geral, os modelos BCT e log-*t* apresentaram os melhores ajustes para a distribuição de consumo de energia.

Tabela 5.11: Estimativas dos parâmetros dos modelos log-*t*, log-normal e log-exponencial potência; critério de Akaike e medidas de comparação AD, ADR e AD2R para distribuição de consumo de energia.

Distribuição	log-*t*		log-normal		log-EP	
Parâmetros	Estimativa	EP	Estimativa	EP	Estimativa	EP
μ	1722,00	34,44	1717,00	36,74	1721,00	34,09
σ	0,34	0,02	0,41	0,02	0,41	0,02
τ	6,09	1,34			1,40	0,14
AIC	5860,01		5874,72		5862,27	
AD	0,19		1,16		0,26	
ADR	0,10		0,55		0,15	
AD2R	1,80		48,17		2,70	

5.6 Conclusões

Neste capítulo definiu-se a nova classe de distribuições Box-Cox simétricas para variáveis contínuas positivas, apresentaram-se algumas de suas propriedades e obtiveram-se momentos e quantis. Mostrou-se que essa classe tem como casos particulares as classes de distribuições log-simétricas, por exemplo, a distribuição log-normal, e simétricas truncadas com suporte nos reais positivos, além das distribuições Box-Cox *t*, Box-Cox Cole-Green e Box-Cox exponencial potência. A nova classe de distribuições permite interpretação dos parâmetros em termos de quantis (em particular da mediana), dispersão relativa e assimetria, o que a torna atrativa para modelagem de regressão. A classe de distribuições BCS é útil ainda para modelar dados assimétricos positivos com observações discrepantes, já que inclui distribuições com cauda direita pesada. De fato, este capítulo apresenta um estudo detalhado de peso da cauda direita das distribuições da classe BCS e comprova que esta inclui distribuições com cauda direita do "tipo-Pareto", bem como mais leve ou mais pesada que esta. Verifica-se ainda que, para algumas distribuições BCS, o método de máxima verossimilhança é robusto a observações discrepantes para a estimação de μ e de σ.

Aplicações de distribuições BCS e modelos alternativos à análise de dados de consumo de nutrientes foram apresentadas e comparações entre os diferentes enfoques foram discutidas. Constatou-se que modelos BCS proporcionaram melhores ajustes que os modelos alternativos.

CAPÍTULO 6

Considerações finais e propostas futuras

Esta tese apresenta uma proposta alternativa para a estimação da distribuição usual de consumo e da prevalência de inadequação alimentar através de um modelo Box-Cox *t* com efeitos aleatórios ou mistos, em especial para dados que apresentam alta assimetria e/ou presença de pontos discrepantes. A modelagem proposta não requer transformação nos dados e permite a interpretação de efeitos de covariáveis sobre quantis da distribuição de consumo. Um estudo com um banco de dados de consumo de 22 micronutrientes é apresentado e, na maioria dos casos, os valores do critério de Akaike foram menores do que para o modelo padrão (Tooze et al., 2010), e os valores estimados de prevalência de inadequação foram mais plausíveis com o observado no banco de dados. Por meio do estudo de simulação, constatou-se que o modelo Box-Cox *t* estima de forma adequada a distribuição de consumo usual populacional, em especial para casos de assimetria à direita e/ou na presença de pontos discrepantes.

Este trabalho propõe ainda uma nova classe de distribuições que envolve uma transformação do tipo Box-Cox e a classe de distribuições simétricas truncadas para variáveis contínuas e positivas, denominada classe das distribuições Box-Cox simétricas. Esta classe inclui as distribuições log-simétricas e as simétricas truncadas com suporte nos reais positivos. Algumas propriedades, os quantis e os momentos são obtidos e estudos sobre o decaimento da cauda são apresentados. A classe possibilita interpretação dos parâmetros em termos de quantis, o que a torna interessante para modelagem de regressão. Para algumas distribuições da classe Box-Cox simétrica, o método de máxima verossimilhança é robusto a observações discrepantes para a estimação dos parâmetros relacionados à mediana e à dispersão relativa. Aplicação a um banco de dados de consumo de 22 micronutrientes e 11 macronutrientes mostrou que os modelos Box-Cox

simétricos apresentaram melhores ajustes do que modelos alternativos.

Para estudos futuros pretende-se

(i) implementar as distribuições da classe Box-Cox simétrica no programa *R* por meio da rotina *gamlss*, assim como já estão disponíveis as distribuições Box-Cox *t*, Box-Cox Cole-Green e Box-Cox exponencial potência;

(ii) estudar a modelagem de regressão na classe de distribuições Box-Cox simétricas;

(iii) desenvolver modelos de regressão Box-Cox simétricos para análise do excesso de consumo de nutrientes que podem ser prejudiciais à saúde (Morimoto et al., 2012);

(iv) desenvolver modelos Box-Cox simétricos inflacionados de zeros. Estes podem ser úteis para modelagem de dados referentes a ingestão de alimentos que são consumidos episodicamente, os quais tipicamente apresentam uma quantidade significativa de zeros (não consumo);

(v) estudar métodos de diagnósticos para modelos de regressão Box-Cox simétricos e Box-Cox simétricos inflacionados de zeros.

APÊNDICE A

Apêndice

A.1 Estimação dos Modelos Aditivos Generalizados para Locação, Escala e Forma

Na linguagem de programação *R*, dois métodos de estimação para a classe dos Modelos Aditivos Generalizados para Locação, Escala e Forma (*Generalized Additive Models for Location, Scale and Shape (GAMLSS)*) estão implementados.

O primeiro é o da maximização da densidade a posteriori (*Maximium a Posteriori (MAP) Estimation*), que é equivalente a maximização de uma função de verossimilhança hierárquica ou penalizada, a qual utiliza um algoritmo de retroajuste (*backfitting*) para estimação dos parâmetros. A ideia central de um algoritmo de retroajuste é de um processo iterativo que busca minimizar uma função de perda (normalmente o erro quadrático) em relação a cada uma das funções (referente a uma variável preditora) até a convergência. Na ausência de um termo de penalização, o algoritmo resume-se ao procedimento de Newton-Raphson. Este procedimento pode ser encontrado na rotina `gamlss` (Rigby & Stasinopoulos, 2006a).

O segundo método de estimação é o da maximização da verossimilhança marginal aproximada, presente na rotina `gamlss.mx` (Stasinopoulos et al., 2008) que faz o uso do algoritmo EM (*Expectation-Maximization*) para o cálculo das estimativas dos parâmetros.

Nesta tese, o primeiro método foi utilizado para estimativa dos valores iniciais da Seção 4, sem a penalização; e o método da verossimilhança marginal foi usada para verificar a programação feita.

Na próxima seção encontram-se os métodos da verossimilhança marginal e per-

filada, os quais foram usados para o ajuste dos modelos descritos nos Capítulos 3 e 4.

A.1.1 Estimação por máxima verossimilhança marginal aproximada

Seja y_i o vetor de observações da variável resposta referente ao indivíduo i, com distribuição $f(y_i, \boldsymbol{\theta})$, em que $\boldsymbol{\theta}$ é o vetor de parâmetros associados a distribuição f e γ_i o efeito aleatório, com distribuição $h(\gamma_i, \lambda)$, em que λ é o parâmetro associado a distribuição h.

A função densidade de probabilidade conjunta de (y_i, γ_i) é da forma

$$f(y_i, \gamma_i) = \prod_{j=1}^{n_i} f_{ij}(y_{ij}|\gamma_i)h(\gamma_i),$$

e a função densidade de probabilidade marginal de y_i é dada por

$$f(y_i) = \int_{-\infty}^{\infty} \prod_{j=1}^{n_i} f_{ij}(y_{ij}|\gamma_i)h(\gamma_i)\mathrm{d}\gamma_\mathrm{i}.$$

Assim, a função densidade de probabilidade marginal de $y = (y_1^\top, y_2^\top, \ldots, y_N^\top)^\top$ é dada por

$$f(\boldsymbol{y}) = \prod_{i=1}^{N} f_i(y_i) = \prod_{i=1}^{N} \int_{-\infty}^{\infty} \prod_{j=1}^{n_i} f_i(y_{ij}|\gamma_i)h(\gamma_i)\mathrm{d}\gamma_\mathrm{i},$$

e o logaritmo da função de verossimilhança pode ser escrito como

$$\begin{aligned} \ell(\boldsymbol{\theta}) &= \log f(\boldsymbol{\theta}) = \log \prod_{i=1}^{N} f_i(y_i) = \sum_{i=1}^{N} \log f_i(y_i) = \\ &= \sum_{i=1}^{N} \log \left[\int_{-\infty}^{\infty} \prod_{j=1}^{n_i} f_i(y_{ij}|\gamma_i)h(\gamma_i)\mathrm{d}\gamma_\mathrm{i} \right]. \end{aligned} \tag{A.1}$$

A integral dada na equação (A.1) é aproximada pelo método de quadratura de Gauss-Hermite (detalhes no Apêndice A.3).

Desse modo, para se obter a estimação por máxima verossimilhança basta derivar a equação aproximada pela quadratura de Gauss-Hermite em relação aos parâmetros de interesse, e assim obtém-se o vetor escore. A solução dos estimadores aproximados de máxima verossimilhança são encontrados resolvendo o sistema de equações formado a partir de cada vetor escore igualado a zero (Casella & Berger, 2002, p.316). Quando

um processo iterativo é requerido, uma linguagem de programação é utilizada para estimar os parâmetros.

A.1.2 Verossimilhança perfilada

Existem situções nas quais se deseja realizar inferência envolvendo apenas alguns parâmetros do modelo. Em tais situações, os parâmetros de interesse são aqueles para os quais se deseja fazer a inferência, os demais são chamados de parâmetros de perturbação. Uma possível solução é utilizar uma pseudo-verossimilhança, que faz com que a função dos dados dependa somente dos parâmetros de interesse.

A ideia comumente utilizada é substituir o vetor de parâmetros de perturbação por uma estimativa consistente deste na verossimilhança original. A função resultante é conhecida como função de verossimilhança perfilada. A seguir, apresenta-se um resumo do método baseado no trabalho de Araújo Júnior (2006).

Considere um modelo paramétrico em que $\boldsymbol{\theta}$ é um vetor de parâmetros desconhecidos de dimensão p. Adotando a decomposição (τ, ϕ) do vetor $\boldsymbol{\theta}$, suponha que o interesse seja fazer inferência somente sobre o parâmetro τ. Assim, os parâmetros τ e ϕ são ditos de interesse e perturbação, respectivamente.

A função de verossimilhança perfilada pode ser obtida substituindo, na função de verossimilhança original, o vetor de parâmetros de perturbação ϕ por sua estimativa de máxima verossimilhança para valores fixados do parâmetro de interesse.

Assim, pode-se escrever $\widehat{\theta}_\tau = (\tau, \widehat{\phi}_\tau)$, em que $\widehat{\phi}_\tau$ é a solução em ϕ de $\partial\ell(\tau, \phi)/\partial\phi = 0$. A função de verossimilhança perfilada é definida por

$$L_p(\tau) = L(\tau, \widehat{\phi}_\tau).$$

Seja $\ell(\tau, \phi; y)$ o logaritmo da função de verossimilhança, que é unicamente maximizada com relação aos parâmetros dado y. Supondo que a estimativa $\widehat{\phi}_\tau$ é única, o logaritmo da função de verossimilhança τ é definido como

$$\ell_p(\tau) = l(\tau, \widehat{\phi}_\tau; y) = sup_\phi l(\tau, \phi; y).$$

A expressão acima sugere um procedimento de maximização em duas etapas. A primeira etapa consiste em achar o valor único $\widehat{\phi}_\tau$ que maximiza $\ell(\tau, \phi)$ com respeito a ϕ supondo τ constante. A segunda etapa visa encontrar o valor de τ que maximiza $\ell_p(\tau) = \log L_p(\tau)$.

Assim, o estimador de máxima verossimilhança perfilada $\widehat{\tau}$ é obtido no caso conti-

nuamente diferenciável como solução da equação

$$\frac{\partial \ell_p(\tau)}{\partial \tau} = 0.$$

Um procedimento semelhante a este foi realizado na construção do *grid* para o parâmetro referente ao número de graus de liberdade descrito no Capítulo 3. No caso, valores diferentes para o parâmetro referente ao número de graus de liberdade foram fixados e os demais parâmetros foram estimados, construindo um perfil dos valores da função de máxima verossimilhança.

A.2 Distribuição Box-Cox t

A.2.1 O logaritmo da verossimilhança (ℓ)

O logaritmo da função de verossimilhança de uma variável aleatória Y que segue uma distribuição $BCT(\mu, \sigma, \nu, \tau)$ é dado por

$$\ell = (\nu - 1)\log y - \nu \log \mu - \log \sigma + \log f_T(z) - \log F_T\left(\frac{1}{\sigma|\nu|}\right), \tag{A.2}$$

em que Z, $f_T(t)$ e $F_T(t)$ são definidos conforme (1.6), (1.7) e (1.9), respectivamente. O logaritmo da função de verossimilhança de um conjunto de dados composto por variáveis aleatórias independentes e identicamente distribuídas é dado por $\ell_d = \sum_{i=1}^{n} \ell_i$, em que ℓ_i é o logaritmo da função de verossimilhança de uma observação Y_i que segue uma distribuição Box-Cox t conforme definido em (A.2).

A.2.2 Derivadas de primeira ordem de ℓ

As derivadas de primeira ordem do logaritmo da função de verossimilhança (ℓ) com respeito a μ, σ, ν, τ são dadas por

$$\frac{\partial \ell}{\partial \mu} = \frac{wz}{\mu\sigma} + \frac{\nu}{\mu}(wz^2 - 1),$$

em que $w = (\tau + 1)/(\tau + z^2)$ e $z = Z(y)$ conforme (1.6),

$$\frac{\partial \ell}{\partial \sigma} = \begin{cases} \dfrac{1}{\sigma}(wz^2 - 1) + \dfrac{h(\sigma,\nu,\tau)}{\sigma^2|\nu|}, & \text{se} \quad \nu \neq 0, \\ \dfrac{1}{\sigma}(wz^2 - 1), & \text{se} \quad \nu = 0, \end{cases}$$

$$\frac{\partial \ell}{\partial \nu} = \frac{wz^2}{\nu} - \log\left(\frac{y}{\mu}\right)\left(wz^2 + \frac{wz}{\sigma\nu} - 1\right) + sinal(\nu)\frac{h(\sigma,\nu,\tau)}{\sigma\nu^2}, \quad \text{se} \quad \nu \neq 0,$$

se $\nu = 0$ a última parcela em $\partial\ell/\partial\nu$ deve ser substituída por seu limite quando $\nu \to 0$,

$$\frac{\partial \ell}{\partial \tau} = \begin{cases} \dfrac{\partial \ell}{\partial \tau} = -\dfrac{1}{2}\log\left(1 + \dfrac{z^2}{\tau}\right) + \dfrac{wz^2}{2\tau} + \dfrac{1}{2}\psi\left(\dfrac{(\tau+1)}{2}\right) - \dfrac{1}{2}\psi\left(\dfrac{\tau}{2}\right) - \dfrac{1}{2\tau} - j(\sigma,\nu,\tau), & \text{se} \quad \nu \neq 0, \\ \dfrac{\partial \ell}{\partial \tau} = -\dfrac{1}{2}\log\left(1 + \dfrac{z^2}{\tau}\right) + \dfrac{wz^2}{2\tau} + \dfrac{1}{2}\psi\left(\dfrac{(\tau+1)}{2}\right) - \dfrac{1}{2}\psi\left(\dfrac{\tau}{2}\right) - \dfrac{1}{2\tau}, & \text{se} \quad \nu = 0, \end{cases}$$

em que

$$h(\sigma,\nu,\tau) = \frac{f_T\left(\frac{1}{\sigma|\nu|}\right)}{F_T\left(\frac{1}{\sigma|\nu|}\right)},$$

e

$$j(\sigma,\nu,\tau) = \frac{\partial}{\partial\tau}\left[\log F_T\left(\frac{1}{\sigma|\nu|}\right)\right],$$

em que $\psi(s) = \frac{d}{ds}[\log\Gamma(s)]$ é a função digamma.

A.2.3 Valor esperado das derivadas de segunda ordem e cruzadas de ℓ

Os valores esperados das derivadas de segunda ordem e cruzadas do logaritmo da função da verossimilhança a respeito dos parâmetros μ, σ, ν e τ requerem a avaliação de termos da forma $E[w^sZ^r]$ para valores inteiros de s e r, em que $w = (\tau+1)/(\tau+Z^2)$.

Se a probabilidade da região de truncamento tiver probabilidade desprezível, Z segue aproximadamente uma distribuição t com τ graus de liberdade. Se $Z \sim t_\tau$ então $E(w^sZ^r) = 0$ se r é ímpar e $E(Z^2) = \tau/(\tau-2)$, $E(Z^4) = (3\tau^2)/[(\tau-2)(\tau-4)]$, $E(w) = 1$ e $E(w^2) = [(\tau+1)(\tau+2)]/[\tau(\tau+3)]$. Das identidades $(\tau+Z^2)w^sZ^r = (\tau+1)w^{s-1}Z^r$, para $r = 0, 2, 4$ e $s = 1, 2$ segue que: $E[wZ^2] = 1$, $E[wZ^4] = (3\tau)/(\tau-2)$, $E[wZ^6] = (15\tau^2)/[(\tau-2)(\tau-4)]$, $E[w^2Z^2] = (\tau+1)/(\tau+3)$, $E[w^2Z^4] = [3(\tau+1)]/(\tau+3)$, $E[w^2Z^6] = [15\tau(\tau+1)]/[(\tau-2)(\tau+3)]$.

Para os resultados a seguir, usa-se ainda a expansão em Taylor dada por $\log(Y/\mu) = (1/\nu)\log(1+\sigma\nu Z) \approx \sigma Z - (\sigma^2\nu Z^2)/2 + (\sigma^3\nu^2Z^3)/3 - (\sigma^4\nu^3Z^4)/4$ e o valor esperado é calculado retendo os termos de ordem menor em σ^j.

Os valores esperados aproximados das derivadas de segunda ordem do logaritmo da função de verossimilhança(ℓ) com respeito a μ, σ, ν, τ, para $\tau > 2$, são dados por

$$\mathrm{E}\left[\frac{\partial^2\ell}{\partial\mu^2}\right] \approx -\frac{\tau + 2\sigma^2\nu^2\tau + 1}{\mu^2\sigma^2(\tau+3)},$$

$$\mathrm{E}\left[\frac{\partial^2\ell}{\partial\sigma^2}\right] \approx -\frac{2\tau}{\sigma^2(\tau+3)},$$

$$\mathrm{E}\left[\frac{\partial^2\ell}{\partial\nu^2}\right] \approx -\frac{\sigma^2\tau(7\tau-9)}{4(\tau-2)(\tau+3)},$$

$$\mathrm{E}\left[\frac{\partial^2\ell}{\partial\tau^2}\right] \approx \frac{1}{4}\psi^{(1)}\left[\frac{(\tau+1)}{2}\right] - \frac{1}{4}\psi^{(1)}\left[\frac{\tau}{2}\right] + \frac{(\tau+5)}{2\tau(\tau+1)(\tau+3)},$$

em que $\psi^{(1)}(s) = \frac{d}{ds}\psi(s)$ é a função trigamma.

Os valores esperados aproximados das derivadas cruzadas de ℓ com respeito a (μ, σ, ν, τ) são dados por

$$\mathrm{E}\left[\frac{\partial^2\ell}{\partial\mu\partial\sigma}\right] \approx -\frac{2\nu\tau}{\mu\sigma(\tau+3)},$$

$$\mathrm{E}\left[\frac{\partial^2\ell}{\partial\mu\partial\tau}\right] \approx \frac{2\nu}{\mu(\tau+1)(\tau+3)},$$

$$\mathrm{E}\left[\frac{\partial^2\ell}{\partial\sigma\partial\tau}\right] \approx \frac{2}{\sigma(\tau+1)(\tau+3)},$$

$$\mathrm{E}\left[\frac{\partial^2\ell}{\partial\mu\partial\nu}\right] \approx \frac{\tau-3}{2\mu(\tau+3)},$$

$$\mathrm{E}\left[\frac{\partial^2\ell}{\partial\sigma\partial\nu}\right] \approx -\frac{\sigma\nu\tau}{\tau+3},$$

$$\mathrm{E}\left[\frac{\partial^2 \ell}{\partial \nu \partial \tau}\right] \approx \frac{\sigma^2 \nu(2\tau+1)}{(\tau+1)(\tau+3)(\tau-2)}.$$

A.3 Quadratura de Gauss-Hermite

O resumo descrito nesta seção baseou-se no trabalho de Carrasco (2012) e Peixoto (2008).

O cálculo diferencial e integral é constantemente utilizado como instrumento para resolução de problemas em diversas áreas do conhecimento, sendo bastante comum a existência de integrais de funções que não possuem antiderivada explícita ou cuja antiderivada não seja um processo simples de se obter. Assim, métodos de aproximação surgem como uma alternativa na busca da aproximação dos valores de tais integrais. Os métodos que aproximam integrais envolvem combinações lineares de avaliação do integrando, isto é

$$\int_a^b f(x)dx \approx c_1 f(x_1) + c_2 f(x_2) + \ldots + c_q f(x_q) = \sum_{i=1}^{q} c_i f(x_i), \quad -\infty < a < b < \infty, \quad \text{(A.3)}$$

em que os termos $x_1, x_2, \ldots, x_q$ são chamados de nós ou abscissas e $c_1, c_2, \ldots, c_q$ são os coeficientes que pertencem ao conjunto dos números reais. Tem-se ainda que na equação (A.3), para os casos nos quais $a = -\infty$ ou $b = \infty$, assume-se, sem perda de generalidade, que o intervalo é aberto neste(s) extremo(s). A equação (A.3) é uma aproximação da integral pela soma de áreas de retângulos de base c_i e altura $f(x_i)$, para $i = 1, \ldots, q$, conforme a Figura A.1 (Peixoto, 2008).

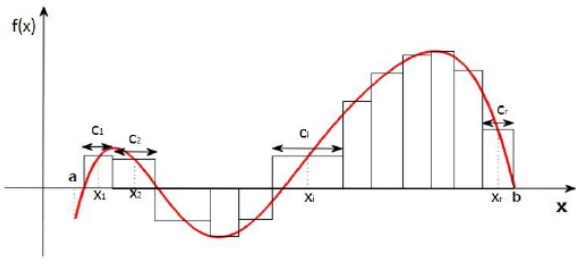


Figura A.1: Aproximação da integral por retângulos de base c_i e altura $f(x_i)$.

A quadratura de Gauss [1] considera as abscissas x_i da equação (A.3) como sendo as raízes de um polinômio ortogonal.

Uma sequência de polinômios $\{p_q(x)\}_{q=0}^{\infty}$ pertencentes ao espaço de todos os polinômios algébricos de um grau menor ou igual a q, é uma sequência de polinômios ortogonais em relação a uma função peso $w(x)$ definido no intervalo real $[a, b]$[2], se

- $p_q(x) = \sum_{i=0}^{q} A_{q,i}x^i$ possui grau exatamente q, isto é, $A_{q,q} \neq 0$;
- $< p_{q_1}(x), p_{q_2}(x) >= \int_a^b p_{q_1}(x)p_{q_2}(x)df(x) = \int_a^b p_{q_1}(x)p_{q_2}(x)w(x)d(x) = 0$,

sendo $f(x)$ uma função contínua. Os polinômios ortogonais mais comuns encontrados na literatura são os polinômios de Jacobi, de Hermite e de Leguerre. Neste estudo foram utilizados os polinômios ortogonais de Hermite.

Seja uma sequência de polinômios $\{p_q(x)\}_{q=0}^{\infty}$, definida com $w(x) = e^{-x^2}$ sobre o intervalo $(-\infty, \infty)$. Então, o polinômio de Hermite de grau q é dado por

$$H_q(x) = (-1)^q e^{x^2} \frac{d^q}{dx^q}\{e^{-x^2}\}.$$

O polinômio de Hermite $H_q(x)$ definido como $w(x) = e^{-x^2}$ está diretamente associado ao método de quadratura de Gauss. Esta associação é conhecida como quadratura de Gauss-Hermite e permite aproximar integrais da forma $\int_{-\infty}^{\infty} e^{-x^2}\mathrm{dx}$, como

$$\int_{-\infty}^{\infty} f(x)e^{-x^2}\mathrm{dx} \approx \sum_{\mathrm{i=1}}^{\mathrm{q}} v_{\mathrm{i}}\mathrm{f}(\mathrm{s_i}),$$

em que s_i, $i = 1, 2, \ldots, q$ são as raízes de $H_q(x)$ e v_i são os pesos associados a cada raiz do polinômio, dados por

$$v_i = \frac{2^{q-1}q!\sqrt{\pi}}{q^2[H_{q-1}(s_i)]^2}.$$

As raízes e pesos dos polinômios de Hermite podem ser calculados pelo programa *R* (versão 3.0.1), os quais encontram-se na rotina `statmod` na função `gauss.quad`, com opção `hermite`.

[1] Carl Friedrich Gauss (1777-1855) aprimorou as técnicas de Newton-Cotes (da obra *Methodus nova integralium valores per approxinationem inveniend: Commentationes Societatis Regiae Scientarium Gottingensis Recentiores*, 1814).

A.4 Programação em R do modelo Box-Cox t com efeito aleatório normal

```
library(gamlss)
library(statmod)
library(bbmle)

BCT.Int<-function(b,beta.fixo,sigma,nu,tau,lambda,X,Z,Y,log=TRUE){
zeta<-exp(lambda)
ll=sapply(b,function(bi){
preditor<-as.matrix(X)%*%beta.fixo+as.matrix(Z)%*%bi
mu=exp(preditor)
sigma=exp(sigma)
tau=exp(tau)
sum(dBCT(Y,mu=mu,sigma=sigma,nu=nu,tau=tau,log=TRUE))+
dnorm(bi,0,sd=1/zeta,log=TRUE)
})
if(log==FALSE){ll<-exp(ll)}
return(ll)}

gauss.hermite<-function(BCT.Int,n.pontos,beta.fixo,sigma,nu,tau,
lambda,X,Z,Y,log=FALSE){
pontos<-gauss.quad(n.pontos,kind="hermite")
integral<-sum(pontos$weights*BCT.Int(pontos$nodes,beta.fixo,
sigma,nu,tau,lambda,X,Z,Y,log=FALSE)/
exp(-pontos$nodes^2))
return(integral)
}

veroM<-function(modelo,formu.X,formu.Z,beta.fixo,sigma,nu,tau,lambda,
integral,pontos,dados){
dados.id<-split(dados,dados$dados.id)
ll<-c()
for(i in 1:length(dados.id)){
X<-model.matrix(as.formula(formu.X),data=dados.id[[i]])
Z<-model.matrix(as.formula(formu.Z),data=dados.id[[i]])
```

```
ll[i]<-gauss.hermite(modelo,n.pontos=pontos,X=X,Z=Z,
Y=dados.id[[i]]$dados.nutriente,
beta.fixo=beta.fixo,sigma=sigma,nu=nu,tau=tau,lambda=lambda,
log=FALSE)
}
return(sum(log(ll)))
}

mod.BCT<-function(b0,sigma0,nu0,tau0,lambda0,integral,pontos,dados){
ll<-veroM(modelo=BCT.Int,formu.X="~1",formu.Z="~1",beta.fixo=b0,
sigma=sigma0,nu=nu0,tau=tau0,lambda=lambda0,
integral=integral,pontos=pontos,
dados=dados)
return(-ll)
}

P.GH=mle2(mod.BCT,start=list(b0=valorinicial1,sigma0=valorinicial2,
nu0=valorinicial3,tau0=valorinicial4,lambda0=valorinicial5),
data=list(integral="GH",pontos=100,dados=dados))
summary(P.GH)
```

A.5 Valor esperado e variância da distribuição Box-Cox Cole-Green

De (5.10), considerando $\nu \neq 0$ e $k = 1$, tem-se que o valor esperado de uma variável aleatória $Y \sim BCS(\mu, \sigma, \nu; r)$ é dado por

$$\mathrm{E(Y)} = \mu \mathrm{E}\left((1 + \sigma\nu \mathrm{S})^{\frac{k}{\nu}} \mathrm{I}_{\mathrm{A}(\sigma,\nu)}(\mathrm{S})\right), \tag{A.4}$$

em que $I_{A(\sigma,\nu)}(\cdot)$ é a função indicadora do conjunto $A(\sigma, \nu)$, com $A(\sigma, \nu)$ dado em (5.5). Assim, para $\nu > 0$, (A.4) pode ser reescrita como

$$\mathrm{E(Y)} = \frac{\mu}{\mathrm{F_S}\left(\frac{1}{\sigma\nu}\right)} \int_{-\frac{1}{\sigma\nu}}^{\infty} (1 + \sigma\nu \mathrm{S})^{\frac{1}{\nu}} \mathrm{f_S(s)ds},$$

em que $F_S(\cdot)$ e $f_S(\cdot)$ são as funções de distribuição acumulada e de densidade de uma variável aleatória S que segue uma distribuição simétrica padrão.

Uma expansão em Taylor (Sen et al., 2010, p.17-18) pode ser útil para aproximar a expressão anterior. Desse modo, seja $f : \mathbb{R} \to \mathbb{R}$, uma função contínua diferenciável até ordem k, em um ponto x_o de $\mathbb{R}$, então:

$$f(x) = f(x_0) + \sum_{j=1}^{k} f^{(j)}(x_0)\frac{(x-x_0)^j}{j!} + R_k(x, x_0),$$

em que $f^{(j)}$ é a j-ésima derivada e o erro da aproximação é definido como

$$R_k(x, x_0) = \frac{(x-x_0)^k}{k!}\{f^{(k)}[hx_0 + (1-h)x] - f^{(k)}(x_0)\},$$

para algum $0 < h < 1$.

Nesse trabalho, define-se $f(s) = (1+\sigma\nu s)^{\frac{1}{\nu}}$ e o ponto zero para realizar a expansão. Assim, tem-se que a derivada de n-ésima ordem da função f é dada por

$$f^n(s) = \sigma^n(1-\nu)(1-2\nu)\cdots(1-(n-1)\nu)(1+\sigma\nu s)^{\frac{1}{\nu}-n} \Rightarrow f^n(0) = \sigma^n(1-\nu)(1-2\nu)\cdots(1-(n-1)\nu),$$

e aproximando a função por um polinômio de segunda ordem, tem-se

$$(1+\sigma\nu s)^{\frac{1}{\nu}} \approx 1 + \sigma s + \sigma^2\frac{(1-\nu)s^2}{2},$$

assim, pode-se reescrever (A.4) como

$$\begin{aligned} \mathrm{E(Y)} &\approx \frac{\mu}{F_S\left(\frac{1}{\sigma\nu}\right)}\int_{-\frac{1}{\sigma\nu}}^{\infty}\left[1 + \sigma s + \sigma^2\frac{(1-\nu)s^2}{2}\right]f_s(s)\mathrm{ds} = \\ &= \frac{\mu}{F_S\left(\frac{1}{\sigma\nu}\right)}\left[\int_{-\frac{1}{\sigma\nu}}^{\infty} f_s(s)\mathrm{ds} + \sigma\int_{-\frac{1}{\sigma\nu}}^{\infty} sf_s(s)\mathrm{d}s + \frac{\sigma^2(1-\nu)}{2}\int_{-\frac{1}{\sigma\nu}}^{\infty} s^2 f_s(s)\mathrm{ds}\right]. \end{aligned}$$

O procedimento para $\nu < 0$ é análogo. Portanto, o valor esperado pode ser aproximado por

$$E(Y) \approx \mu\left[1 + \frac{\sigma}{F_Z\left(\frac{1}{\sigma|\nu|}\right)}\left[E(Z) + \frac{\sigma^2(1-\nu)}{2}E(Z^2)\right]\right],$$

em que Z é uma variável aleatória que segue uma distribuição simétrica padrão truncada com suporte $A(\sigma, \nu)$, sendo $A(\sigma, \nu)$ dado em (5.5).

Seguindo novamente uma aproximação por um polinômio de segunda ordem é feita em torno do ponto zero, através da função $g(s) = (1+\sigma\nu s)^{\frac{2}{\nu}}$, o segundo momento

é dado aproximadamente pela expressão

$$E(Y^2) \approx \mu^2 \left[1 + \frac{\sigma}{F_S\left(\frac{1}{\sigma|\nu|}\right)} \left[2E(Z) + \sigma(2-\nu)E(Z^2) \right] \right],$$

e a variância pode ser aproximada por

$$\text{Var(Y)} \approx \frac{\mu^2\sigma^2}{\text{F}_\text{S}\left(\frac{1}{\sigma|\nu|}\right)} \left[\text{E(Z}^2\text{)} - \frac{1}{\text{F}_\text{S}\left(\frac{1}{\sigma|\nu|}\right)} \left[\text{E(Z)} + \frac{\sigma(1-\nu)}{2}\text{E(Z}^2\text{)} \right]^2 \right].$$

Particularizando para a distribuição Box-Cox Cole-Green, os valores aproximados da esperança e variância são dados, respectivamente, por

$$\text{E(Y)} \approx \begin{cases} \mu \left[1 + \dfrac{\sigma}{2} \left[\sigma(1-\nu) + \dfrac{|3\nu-1| \exp\left\{\frac{-1}{2\sigma^2\nu^2}\right\}}{\sqrt{2\pi}\Phi\left(\frac{1}{\sigma|\nu|}\right)} \right] \right], & \text{se} \quad \nu \neq 0, \\ \mu \exp\left\{\dfrac{\sigma^2}{2}\right\}, & \text{se} \quad \nu = 0, \end{cases}$$

e

$$\text{Var(Y)} \approx \begin{cases} \mu^2 \left[1 + \sigma \left[\sigma(2-\nu) + \dfrac{|3\nu-2| \exp\left\{\frac{-1}{2\sigma^2\nu^2}\right\}}{\nu\sqrt{2\pi}\Phi\left(\frac{1}{\sigma|\nu|}\right)} \right] \right] - \\ - \left[\mu \left[1 + \dfrac{\sigma}{2} \left[\sigma(1-\nu) + \dfrac{|3\nu-1| \exp\left\{\frac{-1}{2\sigma^2\nu^2}\right\}}{\sqrt{2\pi}\Phi\left(\frac{1}{\sigma|\nu|}\right)} \right] \right] \right]^2, & \text{se} \quad \nu \neq 0, \\ \mu^2[\exp\{2\sigma^2\} - \exp\{\sigma^2\}], & \text{se} \quad \nu = 0, \end{cases}$$

em que $\Phi(\cdot)$ é a função de distribuição acumulada de uma variável aleatória S que segue uma distribuição normal padrão. É importante ressaltar ainda que, para $\nu = 0$, os resultados apresentados são exatos.

A.6 Terceiro e quarto momentos da distribuição Box-Cox Cole-Green

Para obtenção dos valores aproximados de assimetria e curtose faz-se necessário o uso dos terceiro e quarto momentos (5.10 e 5.11). Desse modo, aproximações por Taylor utilizando-se polinômios de terceiro e quarto graus, respectivamente, seguindo o mesmo procedimento descrito no Apêndice A.5 podem ser feitos. Assim, seus valores aproximados, são dados respectivamente por

$$\begin{aligned}\mathrm{E}(\mathrm{Y}^3) &\approx \mu^3\left[1+\frac{3\sigma}{F_S\left(\frac{1}{\sigma|\nu|}\right)}\left[E(Z)+\frac{\sigma(3-\nu)}{2}E(Z^2)+\right.\right.\\ &\quad\left.\left.+\frac{\sigma^2(3-\nu)(3-2\nu)}{6}E(Z^3)\right]\right],\end{aligned}$$

$$\begin{aligned}\mathrm{E}(\mathrm{Y}^4) &\approx \mu^4\left[1+\frac{4\sigma}{F_S\left(\frac{1}{\sigma|\nu|}\right)}\left[E(Z)+\frac{\sigma(4-\nu)}{2}E(Z^2)+\right.\right.\\ &\quad+\frac{\sigma^2(4-\nu)(4-2\nu)}{6}E(Z^3)+\\ &\quad\left.\left.+\frac{\sigma^3(4-\nu)(4-2\nu)(4-3\nu)}{24}E(Z^4)\right]\right],\end{aligned}$$

em que Z é uma variável aleatória que segue uma distribuição simétrica padrão truncada com suporte $A(\sigma,\nu)$, sendo $A(\sigma,\nu)$ dado em (5.5).

Para a distribuição Box-Cox Cole-Green, os valores aproximados dos momentos de terceira e quarta ordem são dados por

$$\mathrm{E}(\mathrm{Y}^3)\approx\begin{cases}\mu^3\left[1+\dfrac{3\sigma^2(3-\nu)}{2}+\dfrac{3\sigma\exp\left\{\frac{-1}{2\sigma^2\nu^2}\right\}}{2\nu\Phi\left(\frac{1}{\sigma|\nu|}\right)\sqrt{2\pi}}\right.\\ \left.\left[3|\nu-1|+\dfrac{|3-\nu|(3-2\nu)(2\sigma^2\nu^2+1)}{3\nu}\right]\right], & \text{se}\quad \nu\neq 0,\\ \mu^3\exp\left\{\dfrac{9\sigma^2}{2}\right\}, & \text{se}\quad \nu=0,\end{cases}$$

$$\mathrm{E}(\mathrm{Y}^4)\approx\begin{cases}\mu^4\left[1+2\sigma^2(4-\nu)+\dfrac{4\sigma}{\Phi\left(\frac{1}{\sigma|\nu|}\right)}\left[\dfrac{\exp\left\{\frac{-1}{2\sigma^2\nu^2}\right\}}{\sqrt{2\pi}}\right.\right.\\ \left[\dfrac{|3\nu-4|}{2\nu}+\dfrac{|4-\nu|(4-2\nu)(2\sigma^2\nu^2+1)}{6\nu^2}\right]+\\ \left.\left.+\dfrac{\sigma^3(4-\nu)(4-2\nu)(4-3\nu)E(Z^4)}{24}\right]\right], & \text{se}\quad \nu\neq 0,\\ \mu^4\exp\left\{8\sigma^2\right\}, & \text{se}\quad \nu=0,\end{cases}$$

em que

$$\mathrm{E}(Z^4)=\int_{-\infty}^{\infty}\mathrm{s}^4\frac{\exp\left\{\frac{-1}{2\sigma^2\nu^2}\right\}}{\sqrt{2\pi}}\mathrm{I}_{\mathrm{A}(\sigma,\nu)}(\mathrm{s})\mathrm{ds},$$

Para $\nu=0$ as expressões são exatas.

A.7 Índice da cauda da classe de distribuições Box-Cox simétrica

Este apêndice mostra detalhes das relações estabelecidas entre o índice da cauda das classes de distribuições simétricas, log-simétricas e Box-Cox simétricas. Assim, seja uma variável aleatória $Y \sim BCS(\mu, \sigma, \nu; r)$, o índice pode ser calculado através das relações

$$\mathcal{L}_{BCS}(y; r) = \begin{cases} \mathcal{L}_S(y^\nu; r), & \text{se} \quad \nu > 0, \\ y^\nu, & \text{se} \quad \nu < 0, \\ \mathcal{L}_{LS}(y; r), & \text{se} \quad \nu = 0. \end{cases}$$

Prova. Se $Y \sim BCS(\mu, \sigma, \nu; r)$, com função densidade dada em (5.8) e $\nu \neq 0$, tem-se

$$\begin{aligned} \mathcal{L}_{BCS}(y; r) &= y \lim_{t \longrightarrow \infty} \frac{f(ty)}{f(t)} = y \lim_{t \longrightarrow \infty} \frac{\frac{(ty)^{\nu-1}}{\mu^\nu \sigma} r\left(\left\{\frac{1}{\sigma\nu}\left[\left(\frac{ty}{\mu}\right)^\nu - 1\right]\right\}^2\right)}{\frac{t^{\nu-1}}{\mu^\nu \sigma} r\left(\left\{\frac{1}{\sigma\nu}\left[\left(\frac{t}{\mu}\right)^\nu - 1\right]\right\}^2\right)} \\ &= y^\nu \lim_{t \longrightarrow \infty} \frac{r\left(\left\{\frac{1}{\sigma\nu}\left[\left(\frac{ty}{\mu}\right)^\nu - 1\right]\right\}^2\right)}{r\left(\left\{\frac{1}{\sigma\nu}\left[\left(\frac{t}{\mu}\right)^\nu - 1\right]\right\}^2\right)}. \end{aligned} \tag{A.5}$$

Quando $\nu > 0$, $t \longrightarrow \infty$ implica em $x = (t/\mu)^\nu \longrightarrow \infty$, e então de (A.5) tem-se,

$$\mathcal{L}_{BCS}(y; r) = y^\nu \lim_{x \longrightarrow \infty} \frac{r\left(\left\{\frac{xy^\nu - 1}{\sigma\nu}\right\}^2\right)}{r\left(\left\{\frac{x-1}{\sigma\nu}\right\}^2\right)} = \mathcal{L}_S(y^\nu; r).$$

Quando $\nu < 0$, $t \longrightarrow \infty$ implica em $x = (t/\mu)^\nu \longrightarrow 0$ e novamente de (A.5) obtém-se

$$\mathcal{L}_{BCS}(y; r) = y^\nu \lim_{x \longrightarrow 0} \frac{r\left(\left\{\frac{xy^\nu - 1}{\sigma\nu}\right\}^2\right)}{r\left(\left\{\frac{x-1}{\sigma\nu}\right\}^2\right)} = y^\nu.$$

Quando $\nu = 0$, tem-se

$$\begin{aligned} \mathcal{L}_{BCS}(y; r) &= y \lim_{t \longrightarrow \infty} \frac{f(ty)}{f(t)} = y \lim_{t \longrightarrow \infty} \frac{\frac{1}{ty\sigma} r\left(\left\{\frac{1}{\sigma}\log\left(\frac{yt}{\mu}\right)\right\}^2\right)}{\frac{1}{t\sigma} r\left(\left\{\frac{1}{\sigma}\log\left(\frac{t}{\mu}\right)\right\}^2\right)} \\ &= y \lim_{t \longrightarrow \infty} \frac{r\left(\left\{\frac{\log(yt) - \log\mu}{\sigma}\right\}^2\right)}{r\left(\left\{\frac{\log t - \log\mu}{\sigma}\right\}^2\right)} = \mathcal{L}_{LS}(y; r). \end{aligned}$$

A.8 Critérios de Anderson-Darling

Sejam $y_1, y_2, \ldots, y_n$ uma amostra aleatória de uma variável aleatória com função de distribuição acumulada $F(y)$ e seja k_y o número de observações $y_i < y$. A função de distribuição acumulada empírica é a função escada $\widehat{F}(y) = k_y/n$. Uma medida utilizada para estudar a bondade do ajuste é proposta por Labeyrie (1991), o qual propõe a seguinte distância ponderada:

$$\int_{-\infty}^{\infty} |\widehat{F}(y) - F(y)| W(F(y)) dy,$$

em que $W(y)$ é uma função de ponderação. Adotando $W(y) = 1/(1-y)$, a distância média ponderada é estimada por

$$L = \frac{1}{n} \sum_{i=1}^{n} \frac{|\widehat{F}(y_i) - F(y_i)|}{1 - F(y_i)}.$$

Uma medida de discrepância ou "distância" entre duas distribuições é proposta por Anderson & Darling (1952), baseada na noção usual de uma medida em um espaço de funções, dada por

$$n \int_{-\infty}^{\infty} [\widehat{F}(y) - F(y)]^2 \psi[F(y)] dF(y) \equiv W_n^2,$$

em que $\widehat{F}(y)$ é a função de distribuição acumulada empírica e $\psi[F(y)](\geq 0)$ é alguma função previamente definida. Considerando $\psi(t) = 1/[t(1-t)]$, $\psi(t) = 1/(1-t)$ e $\psi(t) = 1/(1-t)^2$, W_n^2 resulta em

$$AD = n \int_{-\infty}^{\infty} \frac{[\widehat{F}(y) - F(y)]^2}{F(y)[1 - F(y)]} dF(y),$$

$$ADR = n \int_{-\infty}^{\infty} \frac{[\widehat{F}(y) - F(y)]^2}{[1 - F(y)]} dF(y),$$

e

$$AD2R = n \int_{-\infty}^{\infty} \frac{[\widehat{F}(y) - F(y)]^2}{[1 - F(y)]^2} dF(y),$$

respectivamente. A medida AD propõe a ponderação para medir a qualidade de ajuste. A medida ADR é uma variante da medida AD, a qual ressalta a falta de ajuste na cauda direita. A medida $AD2R$ acentua ainda mais a falta de ajuste da cauda direita.

Como $\widehat{F}(y)$ é uma função escada com saltos iguais a $1/n$ nas estatísticas de ordem, as estatísticas de Anderson-Darling podem ser escritas em formas alternativas mais úteis

para propósitos computacionais, como

$$AD = -n - \frac{1}{n}\sum_{i=1}^{n}(2i-1)[\log z_i + \log(1 - z_{n+1-i})],$$

$$ADR = \frac{n}{2} - 2\sum_{i=1}^{n} z_i - \frac{1}{n}\sum_{i=1}^{n}(2i-1)\log(1 - z_{n+1-i}),$$

$$AD2R = 2\sum_{i=1}^{n}\log(1 - z_i) + \frac{1}{n}\sum_{i=1}^{n}\frac{(2i-1)}{(1 - z_{n+1-i})},$$

em que $z_i = F(y_{i:n})$ e $x_{i:n}$ é a i-ésima estatística de ordem (Luceño, 2005, p.906, Apêndice B).

Referências Bibliográficas

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle, *Second International Symposium on Information Theory*, pp. 267–281. Budapest: Akademinai Kiado.

Anderson, R. W. & Darling, D. A. (1952). Asymptotic theory of certain "goodness of fit"criteria based on stochastic processes, *The Annals of Mathematical Statistics* **23**(2): 193–212.

Anderson, T. & Fang, K. (1987). Cochran's theorem for elliptically contoured distributions, *Sankhya A* **49**(3): 305–315.

Araújo Júnior, A. G. (2006). *Ajustes para Verossimilhança Perfilada na Distribuição Birnbaum-Saunders*, Mestrado, Universidade Federal de Pernambuco.

Arellano-Valle, R. (1994). *Distribuições Elípticas: Propriedades, Inferência e Aplicações a Modelos de Regressão*, Tese de Doutorado, Universidade de São Paulo.

Azzalini, A. (2005). The skew-normal distribution and related multivariate families, *Scadinavian Journal of Statistics* **32**(2): 159–188.

Berkane, M. & Bentler, P. M. (1986). Moments of elliptically distributed random variates, *Statistics & Probability Letters* **4**(6): 333–335.

Block, G. (1982). A review of validations of dietary assessment methods, *American Journal of Epidemiology* **115**(4): 492–505.

Borrelli, R., Simonetti, M. S. & Fidanza, F. (1992). Inter and intra individual variability in food intake of elderly people in Perugia (Italy), *British Journal of Nutrition* **68**(1): 3–10.

Box, G. E. P. & Cox, D. R. (1964). An analysis of transformations, *Journal of the Royal Statistical Society* B **26**(2): 211–252.

Cambanis, S., Huang, S. & Simons, G. (1981). On the theory of elliptically contoured distributions, *Journal of Multivariate Analysis* **11**(3): 368–385.

Carrasco, J. (2012). *Modelo de Regressão Beta com Erro nas Variáveis*, Doutorado, Universidade de São Paulo.

Carriquiry, A. L. (1998). Assessing the prevalence of nutrient inadequacy, *Public Health Nutrition* **2**(1): 23–33.

Casella, G. & Berger, R. L. (2002). *Statistical Inference*, Duxbury, United States of America.

Cole, T. & Green, P. J. (1992). Smoothing reference centile curves: the LMS method and penalized likelihood, *Statistics in Medicine* **11**(10): 1305–1319.

Cordeiro, G. M. & Andrade, M. G. (2011). Transformed symmetric models, *Statistical Modelling* **11**(4): 371–388.

Cysneiros, F. J. A. (2004). *Métodos Restritos e Validação de Modelos Simétricos de Regressão*, Tese de Doutorado, Universidade de São Paulo.

de Haan, L. (1970). *On Regular Variation and Its Application to the Weak Convergence of Sample Extremes*, Mathematical Centre tracts, 32. Amsterdam: Mathematics Center.

Dodd, K. W., Guenther, P., Freedman, L., Subar, A., Kipnis, V., Midthune, D., Tooze, J. & Smith, S. (2006). Statistical methods for estimating usual intake of nutrients and foods: A review of the theory, *Journal of the American Dietetic Association* **106**(10): 1640–1650.

Dunn, P. K. & Smyth, G. K. (1996). Randomized quantile residuals, *Journal of Computational and Graphical Statistics* **5**(3): 236–244.

Fang, K. T., Kotz, S. & NG, K. W. (1990). *Symmetric Multivariate and Related Distributions*, Chapman and Hall, London.

Goméz, H. W., Quintana, F. A. & Torres, F. J. (2007). A new family of slash-distributions with elliptical contours, *Statistics and Probability Letters* **77**(7): 717–725.

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J. & Stahel, W. A. (1986). *Robust Statistics: The Approach Based on Influence Functions*, John Wiley & Sons, New York.

Hubert, M. & Vandervieren, E. (2008). An adjusted boxplot for skewed distributions, *Computational Statistics & Data Analysis* **52**(12): 5186–5201.

INS (2003). *Food and Nutrition Board. Dietary Reference Intakes. Applications in Dietary Planning*. Washington DC: National Academy Press.

Johnson, N. L., Kotz, S. & Balakrishnan (1982). *Continuous Univariate Distributions*, John Wiley & Sons, New York.

Kelker, D. (1970). Distribution theory of spherical distributions and a location-scale parameter generalization, *Sankhya A* **32**(4): 419–430.

Labeyrie, J. (1991). Times scales and statistical uncertainties in the predition of extreme environmental conditions, *Reability Engineering & System Safety* **32**(3): 243–266.

Luceño, A. (2005). Fitting the generalized Pareto distribution to data using maximum goodness-of-fit estimators, *Computational Statistics & Data Analysis* **51**(2): 904–917.

Marra, T., Pereira, L., Faria, C., Pereira, D., Martins, M. & Tirado, M. (2007). Avaliação das atividades de vida diária de idosos com diferentes níveis de demência, *Revista Brasileira de Fisioterapia* **11**(4): 267–273.

Morimoto, J. M., Marchioni, D. M. L., Cesar, C. L. G. & Fisberg, R. M. (2012). Statistical innovations improve prevalence estimates of nutrient risk populations: applications in São Paulo, Brazil, *Journal of the Academy of Nutrition and Dietetics* **112**(10): 1614–1618.

Muirhead, R. J. (1982). *Aspects of Multivariate Statistical Theory*, John Wiley & Sons, New York.

Peixoto, L. (2008). *Quadratura de Gauss Iterativa com Base nos Polinômios Ortogonais Clássicos*, Mestrado, Centro Federal de Educação Tecnológica de Minas Gerais.

Pinheiro, J. C. & Bates, D. M. (1995). Approximations to the log-likelihood function in the nonlinear mixed-effects model, *Journal of Computational and Graphical Statistics* **4**(1): 12–35.

R Core Team (2008). *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Rao, B. (1990). Remarks on univariate elliptical distributions, *Statistics & Probability Letters* **10**(4): 307–315.

Resnick, S. I. (2007). *Heavy-Tail Phenomena Probabilistic and Statistical Modeling*, Springer, New York.

Rigby, B., Stasinopoulos, M., Heller, G. & Voudouris, V. (2014). The Distribution Toolbox of GAMLSS, `http://www.gamlss.org`. Data do acesso: 09-12-2014.

Rigby, R. A. & Stasinopoulos, D. M. (2001). The GAMLSS project: A flexible approach to statistical modelling, *In New Trends in Statistical Modelling: Proceedings of the 16th International Workshop on Statistical Modelling, eds B. Klein & L. Korsholm,* pp. 337–345. Odense, Denmark.

Rigby, R. A. & Stasinopoulos, D. M. (2004). Smooth centile curves for skew and kurtotic data modelled using the Box–Cox power exponential distribution, *Statistics in Medicine* **23**(19): 3053–3076.

Rigby, R. A. & Stasinopoulos, D. M. (2005). Generalized additive models for location, scale and shape, *Journal of the Royal Statistical Society* C *(Applied Statistics)* **54**(3): 507–554.

Rigby, R. A. & Stasinopoulos, D. M. (2006a). *Statistical Modelling using GAMLSS in R*, http://www.gamlss.org, London.

Rigby, R. A. & Stasinopoulos, D. M. (2006b). Using the Box-Cox t distribution in GAMLSS to model skewness and kurtosis, *Statistical Modelling* **6**(3): 209–229.

Rodrigues, L. H. & Gomes, I. (2009). High quantile estimation and the port methodology, *REVSTAT* **7**(3): 245–254.

SAS (1990). *User's Guide*, 4 edn. version 6, Cary, v.2, 796p., North Carolina.

Sen, P. K., Singer, J. M. & Pedroso de Lima, A. C. (2010). *From Finite Sample to Asymptotic Methods in Statistics*, Cambridge: Cambridge University Press.

Slater, B., Marchioni, D. & Fisberg, R. (2004). Estimando a prevalência da ingestão inadequada de nutrientes, *Revista de Saúde Pública* **38**(4): 599–605.

Stasinopoulos, D. M., Rigby, R. A. & Akantziliotou, C. (2008). *Instructions on how to use the GAMLSS Package in R*, London. http://www.gamlss.org.

Suitor, C. W. & Gleason, P. M. (2002). Using dietary reference intake-based methods to estimate the prevalence of inadequate nutrient intake among school-aged children, *Journal of the American Dietetic Association* **102**(4): 530–536.

Tooze, J. A., Kipnis, V., Buckman, D. W., Carroll, R., Freedman, L., Guenther, P., Krebs-Smith, S., Subar, A. & K.W., D. (2010). A mixed-effects model approach for estimating the distribution of usual intake of nutrients: The NCI method, *Statistics in Medicine* **29**(27): 2857–2868.

Usuga Manco, O. C. (2013). *Modelos de Regressão Beta com Efeitos Aleatórios Normais e Não Normais para Dados Longitudinais*, Doutorado, Universidade de São Paulo.

Vanegas, L. H. & Paula, G. A. (2014a). Log-symmetric distributions: statistical properties and parameter estimation, *Brazilian Journal of Probability and Statistics* . Aceito para publicação. Disponível em `http://www.imstat.org/bjps/papers/BJPS272.pdf` (Acesso em 11/12/2014).

Vanegas, L. H. & Paula, G. A. (2014b). A semiparametric approach for joint modeling of median and skewness, *Test* . doi:10.1007/s11749-014-0401-7.

Voudouris, V., Gilchrist, R., Rigby, R. A., Sedgwick, J. & Stasinopoulos, D. M. (2012). Modelling skewness and kurtosis with BCPE density in GAMLSS, *Journal of Applied Statistics* **39**(6): 1279–1293.