USO DE MODELOS INFLACIONADOS DE ZEROS NA ANÁLISE DE QUESTIONÁRIOS DE FREQUÊNCIA ALIMENTAR

Giovana Fumes

Dissertação apresentada à Universidade Estadual Paulista "Júlio de Mesquita Filho" para a obtenção do título de Mestre em Biometria.

BOTUCATU
São Paulo - Brasil
Outubro – 2009

USO DE MODELOS INFLACIONADOS DE ZEROS NA ANÁLISE DE QUESTIONÁRIOS DE FREQUÊNCIA ALIMENTAR

Giovana Fumes

Orientador: Prof. Dr. **José Eduardo Corrente**

Dissertação apresentada à Universidade Estadual Paulista "Júlio de Mesquita Filho" para a obtenção do título de Mestre em Biometria.

BOTUCATU
São Paulo - Brasil
Outubro – 2009

FICHA CATALOGRÁFICA ELABORADA PELA SEÇÃO TÉCNICA DE AQUISIÇÃO E TRATAMENTO DA INFORMAÇÃO
DIVISÃO TÉCNICA DE BIBLIOTECA E DOCUMENTAÇÃO - CAMPUS DE BOTUCATU - UNESP
BIBLIOTECÁRIA RESPONSÁVEL: Selma Maria de Jesus

Fumes, Giovana.
Uso de modelos inflacionados de zeros na análise de questionários de freqüência alimentar / Giovana Fumes. – Botucatu : [s.n.], 2009.

Dissertação (mestrado) – Universidade Estadual Paulista, Instituto de Biociências, Botucatu, 2009.
Orientador: José Eduardo Corrente
Assunto CAPES: 40500004

1. Nutrição 2. Dieta 3. Alimentos - Consumo

CDD 612.3

Palavras-chave: Distribuição binominal negativa; Distribuição de Poisson; Distribuição inflacionada de zero; Questionário de frequência alimentar; Superdispersão.

Dedicatória

À Deus, Senhor da minha vida,

Aos meus amados pais, Israel e Alvacir.

"Posso, tudo posso Naquele que me fortalece

e nada e ninguém no mundo vai me fazer desistir.

Quero, tudo quero sem medo entregar meus projetos,

deixar-me guiar nos caminhos que Deus desejou pra mim e ali estar."

Tudo Posso (Celina Borges)

Agradecimentos

À Deus, razão da minha vida e a força que me impulsiona a voar alto.

Aos meus pais, Israel e Alvacir, por sempre acreditarem em mim e me incentivarem a buscar a realização dos meus sonhos.

Às minhas irmãs, Jaqueline e Juliane, pela amizade, pela compreensão nas horas difíceis e pelas orações.

Aos meus sobrinhos, Cauê, Felipe e Beatriz, alegrias eternas do meu coração.

Ao meu cunhado Reginaldo, por toda amizade e apoio.

À minha amada avó Helena, pelas palavras simples e sábias.

Aos meus tios, Lau e Noía, pelo exemplo e incentivo no estudo.

Aos meus amigos de comunidade, em especial, Gláucia e Rose, pela acolhida e oração.

Ao meu orientador, Prof. José Eduardo Corrente, por me incentivar a lutar pelos meus sonhos, acreditar em mim e pela sua dedicada orientação na elaboração desse estudo.

À Profª Maria del Pilar Diaz, pelas contribuições valiosas no trabalho e conversas amigas.

À todos os professores e funcionários do Departamento de Bioestatística - IB - Unesp - Botucatu, pelo ensino e amizade.

Aos meus amigos do mestrado em Biometria do Instituto de Biociências, pela partilha nos estudos e na luta diária.

À Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) pelo apoio financeiro.

Sumário

Lista de Figuras

Página

Lista de Tabelas

Página

USO DE MODELOS INFLACIONADOS DE ZEROS NA ANÁLISE DE QUESTIONÁRIOS DE FREQUÊNCIA ALIMENTAR

Autor: GIOVANA FUMES
Orientador: Prof. Dr. JOSÉ EDUARDO CORRENTE

RESUMO

Um instrumento amplamente utilizado para descrever a dieta habitual de um grupo populacional é o Questionário de Frequência Alimentar (QFA). Esta ferramenta é composta basicamente por uma lista de alimentos e a frequência com a qual cada alimento é consumido. Por se tratar da frequência de consumo, os dados gerados são dados de contagem, os quais podem ser modelados segundo uma distribuição de Poisson na presença de covariáveis. Quando existe uma variabilidade maior do que a esperada pelo modelo, tem-se um fenômeno chamado superdispersão. Nesses casos, uma alternativa seria ajustar os dados a um modelo Binomial Negativo através de uma modificação na função de variância. Porém, esta alta variabilidade pode ser causada pelos dados conterem um número excessivo de zeros que, no caso de dados de um QFA, seria caracterizado pelo não consumo de alguns alimentos. Para esses casos, existem modelos alternativos chamados modelos inflacionados de zeros,

tanto para o caso Poisson como para a Binomial Negativa. Dessa forma, o trabalho presente teve como objetivo estudar os modelos inflacionados de zeros para dados de contagem e aplicá-los a dados de um QFA, de modo a verificar os fatores que influenciam no consumo ou não de determinados alimentos por idosos pertencentes a uma cidade de porte médio do interior do Estado de São Paulo, Brasil. A variável resposta utilizada foi a freqüência mensal do consumo de alimentos provenientes do QFA e as covariáveis associadas eram vinculadas às características sociodemográficas, de prevenção e morbidades referidas. Dos 67 itens obtidos no QFA, alguns foram selecionados para este estudo segundo consulta a profissionais da área nutricional. A priori para verificar qual o modelo mais adequado para as porcentagens de zeros dos dados, foi calculado o valor predito para o ajuste do modelo nulo em cada um dos modelos utilizados: Poisson, Binomial Negativo, Poisson Inflacionado de Zeros e Binomial Negativo Inflacionado de Zeros. Também foi calculada uma taxa de extra-variação, baseada na variância do Modelo de Poisson Inflacionado de Zeros para verificar o quanto da superdispersão era ocasionada pela presença dos zeros nos dados. A comparação entre os modelos usuais e inflacionados foi feita pelo teste de Vuong e as comparações entre os modelos inflacionados foram feitas seguindo o critério de Akaike. Para um conjunto de dados com até 10% de zeros os modelos usuais de Poisson e Binomial Negativo se mostraram adequados. Os modelos inflacionados apresentaram-se adequados na porcentagem de 10% a 50% de zeros. Acima dessa porcentagem, os ajustes dos modelos inflacionados de zeros não se mostraram razoáveis e outros modelos têm sido propostos na literatura. É necessário ressaltar que as variáveis respostas apresentaram uma grande amplitude, que variou de zero a cento e vinte vezes o consumo ao mês para determinados alimentos, podendo ter limitado o ajuste dos modelos propostos. Como resultado da aplicação desses modelos aos dados do QFA para alguns alimentos selecionados, pode-se evidenciar que a probabilidade de não consumo está associada ao sexo, ao estado nutricional, ao funcionamento do intestino, ao consumo diário de água e à prática de atividade física. Já a probabilidade de consumo dos alimentos estudados se mostrou associada a to-

das as covariáveis consideradas no estudo. Conclui-se que, os modelos inflacionados de zeros forneceram uma caracterização integrada do consumo e do não consumo de alimentos nessa população de idosos; porém, esses modelos inflacionados resolveram parcialmente o problema, visto que são ajustados para uma certa porcentagem de observações nulas. Para grandes percentuais de zeros outros modelos devem ser aplicados, sendo necessários estudos posteriores mais detalhados para investigar a relação existente entre o modelo inflacionado a ser utilizado e a porcentagem de zeros.

THE USE OF ZERO-INFLATED MODELS IN THE ANALYSIS OF FOOD FREQUENCY QUESTIONNAIRES

Author: GIOVANA FUMES
Adviser: Prof. Dr. JOSÉ EDUARDO CORRENTE

SUMMARY

The Food Frequency Questionnaire (FFQ) is a widely used instrument to describe the usual diet of population groups. This tool basically consists of a list of food types and the frequency with which each type is consumed. Because it concerns intake frequency, the data generated are count data, which can be modeled according to a Poisson distribution in the presence of covariates. When a greater variety than that expected for the model exists, a phenomenon referred to as overdispersion occurs. In these cases, an alternative would be to fit the data to a Negative Binomial model through a modification in the variance function. However, such high variety may be caused because the data contain an excessive number of zeros, which, in the case of data from a FFQ, would be characterized by the non-consumption of certain foods. In these cases, there are alternative models called zero-inflated models for both the Poisson and the Negative-Binomial cases. Hence, the present investigation

aimed to study zero-inflated models for count data and apply them to data from a FFQ so as to observe the factors that influence or not the intake of certain foods by elderly residents of a medium-sized city in inner São Paulo state, Brazil. The response variable used was the monthly intake frequency of foods from the FFQ, and the associated covariates were related to the socio-demographic, prevention and morbidity characteristics reported. Of the 67 items obtained in the FFQ, some were selected for this study according to consultation with nutrition professionals. At first, in order to evaluate which model was the most suitable for the zero percentages in the data, the value predicted for fitting the null model in each of the models used, namely, Poisson, Negative Binomial, Zero-Inflated Poisson and Zero-Inflated Negative Binomial, was estimated. An extra-variation rate was also calculated based on the variance of the Zero-Inflated Poisson Model in order to evaluate how much overdispersion was caused by the presence of zeros in the data. The comparison between the usual and inflated models was performed by Vuong's test, and the comparisons between the inflated models were carried out according to Akaike's criteria. For a data set of up to 10% of zeros, the usual Poisson and Negative Binomial models showed to be adequate. The inflated models showed to be adequate in the percentage of 10% to 50% of zeros. Above such percentage, the fitting of the zero-inflated models did not show to be reasonable, and other models have been proposed in the literature. It is necessary to point out that the response variables presented large amplitude, which varied from zero to one hundred and twenty-fold the monthly intake for certain food types, and that may have limited the fitting of the models proposed. As a result of the application of such models to the FFQ data, it can be pointed out that the probability of non-consumption is associated with gender, nutritional status, bowel functioning, daily water consumption and physical activity performance. However, the probability of consumption of the foods studied showed to be associated with all the covariates considered in the study. It is concluded that zero-inflated models provided an integrated characterization of the consumption and non-consumption of foods in this population of elderly individuals; however, these

inflated models partly resolved the problem, considering that they are fitted for a certain percentage of null observations. Other models must be applied for large percents of zeros, and further more detailed studies are required to investigate the relationship between the inflated model to be used and the percentage of zeros.

1 INTRODUÇÃO

O Questionário de Frequência Alimentar (QFA) tem como objetivo a avaliação da dieta habitual de grupos populacionais. Ele contém basicamente duas ferramentas: uma lista de alimentos e a frequência habitual com a qual o indivíduo consome determinado alimento. A frequência de consumo está registrada em unidades de tempo: dias, semanas, meses ou anos. O formato é de perguntas simples com respostas fechadas, com opções de zero a dez vezes de consumo para cada alimento (Slater et al., 2003).

Os dados provenientes de um QFA são contagens, originárias da frequência com que cada alimento é consumido habitualmente pelo indivíduo. Uma estratégia de análise é supor que a distribuição desses dados siga uma distribuição de Poisson.

A distribuição de Poisson tem sido amplamente utilizada na análise de dados de contagem nas diversas áreas de conhecimento como: saúde pública, epidemiologia, sociologia, dentre outras. Entretanto, em dados de contagem pode ocorrer uma variabilidade maior do que a esperada. Esse fenômeno se chama superdispersão e, nesses casos, opta-se em utilizar uma distribuição que melhor acomode essa extra variação. Tal distribuição é chamada binomial negativa.

Um outro fenômeno usualmente verificado na análise deste tipo de dados é o não consumo de determinados alimentos. Este não consumo pode ocorrer pelo fato de, no período de coleta, o indivíduo não ter consumido determinado alimento ou, porque, tal alimento não faça parte da dieta habitual do indivíduo. Isso gera um excesso de zeros, podendo levar a uma superdispersão nos dados, que nem sempre é possível ajustá-los segundo uma distribuição binomial negativa. Para contornar

este problema, os modelos de Poisson e binomial negativo inflacionados de zeros são propostos.

A seguir apresentar-se-á uma revisão dos modelos usuais para ajustes de dados de contagens.

1.1 A distribuição de Poisson

A distribuição de Poisson surgiu no ano de 1837, quando Simeon Denis Poisson publicou uma aproximação limite da distribuição binomial, dando origem a essa distribuição que herdaria o seu nome. Mais tarde, em 1898, Bortkiewicz caracterizou seus princípios básicos que consistem na independência e mesma probabilidade de ocorrência entre os eventos. Ele mostrou também que essa distribuição pode ser usada nos casos em que o número de tentativas de um evento é muito grande e a probabilidade de ocorrência é muito baixa (Jonhson & Kotz, 1969).

Uma maneira de se obter a distribuição de Poisson é utilizar o processo de Poisson, que explicita os princípios básicos de independência e mesma probabilidade de ocorrência entre os eventos. Para isto, considere um dado intervalo de números reais e suponha que as contagens ocorram através desse intervalo. Subdividindo esse intervalo em subintervalos de comprimentos suficientemente pequenos, pode-se supor que:

1. a probabilidade de mais de uma contagem em um subintervalo seja nula;

2. a probabilidade de uma contagem em um subintervalo seja a mesma para todos os subintervalos;

3. a contagem em cada subintervalo seja independente de outros subintervalos.

Se o número médio de contagens no intervalo for $\mu > 0$ (também chamado de *taxa de ocorrência*), mostra-se que a variável aleatória Y que conta o número de ocorrências no intervalo tem distribuição de Poisson com parâmetro μ, denotada por $Y \sim P(\mu)$, e sua distribuição de probabilidade é dada por (Meyer,

1974):

$$P(Y = y) = \frac{e^{-\mu}\mu^y}{y!}, \quad \text{para} \qquad y = 0, 1, 2, \ldots$$

com $E(Y) = Var(Y) = \mu$.

Considerando a distribuição de Poisson como uma família de distribuições, tem-se que ela pertence à família exponencial a um parâmetro. Segundo Bickel & Doksum (1977) uma distribuição de probabilidade $f(y, \theta)$ pertence à família exponencial a um parâmetro se existem funções c e d dependentes do parâmetro θ, T e S dependentes de y, tal que ela possa ser escrita de forma:

$$f(y, \theta) = \exp\{c(\theta)T(y) + d(\theta) + S(y)\}I_A(y),$$

em que A é um conjunto de valores que não depende do parâmetro.

No caso de uma variável aleatória $Y \sim P(\mu)$, tem-se que:

$$f(y; \mu) = \exp[\ln(f(y; \mu))] = \exp(y \ln \mu - \mu - \ln y!)I_{\{0,1,2,\ldots\}}(y),$$

em que

$$c(\mu) = \ln \mu, \quad d(\mu) = -\mu, \quad T(y) = y \quad \text{e} \quad S(y) = -\ln y!.$$

Considerando agora um modelo em que a variável resposta tenha uma distribuição de Poisson, a qual pertence a família exponencial a um parâmetro, pode-se utilizar a idéia de Modelos Lineares Generalizados para ajustar essa variável a um conjunto de variáveis explanatórias.

De acordo com McCullagh & Nelder (1989), um modelo linear generalizado é definido por três componentes:

1. Uma variável resposta Y, pertencente à Família Exponencial Canônica de Distribuições, com média μ e parâmetro de escala constante, conhecido, $\phi > 0$, escrita na forma:

$$f(y; \theta, \phi) = \exp\left\{\frac{1}{a(\phi)}[y\theta - b(\theta)] + c(y; \phi)\right\},$$

sendo $b(.)$ e $c(.)$ funções conhecidas e $a(\phi) = \dfrac{\phi}{w}$, com w o peso *a priori* . Desse modo, pode-se mostrar que:

$$E(Y) = b'(\theta) = \mu, \quad Var(Y) = a(\phi)b''(\theta) = a(\phi)V(\mu) = a(\phi)V,$$

em que θ é denominado parâmetro natural ou canônico e $V = \dfrac{d\mu}{d\theta}$ chamada de função de variância, que depende unicamente da média.

2. Um conjunto de p variáveis explicativas que entram no modelo na forma de um modelo linear (componente sistemático). O preditor linear de cada observação η_j é dado por:

$$\eta_j = \sum_{j=1}^{p} x_j \beta_j,$$

em que β_j's são parâmetros desconhecidos associados às covariáveis $x_j's$, para $j = 1, \ldots, p$.

3. Uma função de ligação que liga os componentes aleatório e sistemático. Uma função de ligação $g(.)$ é inversível e diferenciável, relacionando a média da distribuição ao preditor linear, ou seja,

$$g(\mu) = \eta_j = \sum_{j=1}^{p} x_j \beta_j, \qquad j = 1, 2, \ldots, p.$$

Sendo Y uma variável aleatória com distribuição de Poisson com parâmetro μ, sua forma escrita como uma família exponencial canônica é dada por:

$$f(y; \theta, \phi) = \exp(y\theta - e^{\theta} - \ln y!).$$

Assim,

$$\theta = \ln \mu \Longrightarrow \mu = e^{\theta}, \quad b(\theta) = e^{\theta} = \mu, \quad a(\phi) = 1 \quad \text{e} \quad c(y, \phi) = -\ln y!.$$

A função de ligação canônica é dada por $\ln(\mu)$, ou seja,

$$\eta = \theta \Longleftrightarrow g(\mu) = \ln(\mu).$$

Tem-se ainda que:

$$E(Y) = b'(\theta) = e^{\theta} = \mu \quad \text{e} \quad Var(Y) = b''(\theta)a(\phi) = e^{\theta} = \mu.$$

1.1.1 Estimação dos parâmetros na distribuição de Poisson

Seja uma amostra aleatória $Y_1, Y_2, \ldots, Y_n$ da distribuição de Poisson com parâmetro μ. Tem-se que μ pode ser estimado pelo Método de Máxima Verossimilhança, em que a função de verossimilhança é dada por:

$$l(\mu, \boldsymbol{y}) = \prod_{i=1}^{n} f(y_i, \mu) = \sum_{i=1}^{n} l(\mu; y_i) = \sum_{i=1}^{n} [y_i \ln \mu - \mu - \ln y_i!],$$

sendo $f(.)$ a distribuição de probabilidade de Y_i, $i = 1, \ldots, n$.

O fato da distribuição de Poisson pertencer à Família Exponencial, assegura que o logaritmo da função de verossimilhança satisfaz as condições para a obtenção de um máximo global como solução das equações de verossimilhança. Derivando, então, a função de verossimilhança, tem-se:

$$\frac{dl}{d\mu} = \sum_{i=1}^{n} \left[\frac{y_i}{\mu} - 1 \right] = 0 \quad \Rightarrow \quad \hat{\mu} = \overline{y}.$$

Considerando agora o ajuste de um modelo de Poisson adicionado a um conjunto de p variáveis explanatórias, tem-se um sistema de equações de verossimilhança, dado por $U_\beta = \dfrac{dl}{d\beta}$ (Demétrio, 2001). Então:

$$U_j = \sum_{i=1}^{n} \frac{\partial l(\theta_i; y_i, \phi)}{\partial \beta_j} = \sum_{i=1}^{n} \frac{1}{a(\phi)} (y_i - \mu_i) \frac{1}{V(\mu_i)} \frac{d\mu_i}{d\eta_i} x_{ij}, \qquad j = 0, 1, \ldots, p.$$

No caso da distribuição de Poisson, tem-se que:

$$U_j = \sum_{i=1}^{n} (y_i - \mu_i) \frac{1}{\mu_i} x_{ij}.$$

As equações de $U_j = 0$, para $j = 1, 2, ..., p$, são resolvidas numericamente por processos iterativos do tipo Newton-Raphson, o qual é baseado numa aproximação de Taylor para uma função $f(x)$ nas vizinhanças de um ponto x_0. Assim, para estimação dos parâmetros $\boldsymbol{\beta' s}$ tem-se:

$$\boldsymbol{\beta}^{(m+1)} = \boldsymbol{\beta}^{(m)} + (\boldsymbol{I}_0^{-1})^{(m)} \boldsymbol{U}^{(m)},$$

em que $\boldsymbol{\beta}^{(m)}$ e $\boldsymbol{\beta}^{(m+1)}$ são vetores dos parâmetros estimados nos passos (m) e $(m+1)$, $\boldsymbol{U}^{(m)}$ é o vetor escore de derivadas parciais de primeira ordem de $f(x)$, com elementos

$\frac{\partial l}{\partial \beta_j}$, avaliado no passo (m) e $(\boldsymbol{I}_0^{-1})^m$ a inversa negativa da matriz de derivadas parciais de segunda ordem de $f(x)$, com elementos $\frac{-\partial^2 l}{\partial \beta_j \partial \beta_k}$, avaliada no passo m.

Como estas derivadas de segunda ordem não são obtidas facilmente, pode-se substituir a matriz de informação observada I_0 pela matriz esperada de Fisher $\Im$, que consiste na substituição dos valores das derivadas pelos valores esperados das derivadas parciais de segunda ordem. No caso da distribuição de Poisson, o método de Fisher coincide com o de Newton-Raphson, não havendo problema na substituição realizada, assegurando a convergência do processo iterativo. Assim,

$$\boldsymbol{\beta}^{(m+1)} = \boldsymbol{\beta}^{(m)} + (\mathfrak{S}^{-1})^{(m)} \boldsymbol{U}^{(m)},$$

sendo que $\Im$ tem elementos dados por $\Im_{jk} = E\left[\frac{-\partial^2 l}{\partial \beta_j \partial \beta_k}\right]$, que é a matriz de covariâncias dos $U_j's$ para $j, k = 1, 2, \ldots, p$.

Multiplicando ambos os termos da equação anterior por $\Im$, tem-se:

$$\mathfrak{S}^{(m)} \boldsymbol{\beta}^{(m+1)} = \mathfrak{S}^{(m)} \boldsymbol{\beta}^{(m)} + \boldsymbol{U}^{(m)}. \tag{1}$$

O valor esperado da matriz de informação de Fisher é dado por:

$$\Im_{jk} = E(U_j U_k) = \sum_{i=1}^{n} \frac{1}{a_i(\phi)} \frac{1}{V(\mu_i)} \left(\frac{d\mu_i}{d\eta_i}\right)^2 x_{ij} x_{ik},$$

fazendo $a_i(\phi) = \frac{\phi}{w_i}$, com $\phi > 0$ constante, w_i os pesos *a priori* e $W_i = \frac{w_i}{V(\mu_i)} \left(\frac{d\mu_i}{d\eta_i}\right)^2$, para $i = 1, 2, \ldots, n$. Na forma matricial, pode-se escrever a matriz de Informação de Fisher como:

$$\mathfrak{S} = \frac{1}{\phi} \boldsymbol{X}^T \boldsymbol{W} \boldsymbol{X},$$

em que $\boldsymbol{X}$ é a matriz do modelo e $\boldsymbol{W} = diag(W_1, W_2, ..., W_n)$. Como tem-se ligação canônica, $W_i = w_i V(\mu_i)$, pois $\frac{d\theta_i}{d\mu_i} = \frac{d\eta_i}{d\mu_i} = V^{-1}(\mu_i)$ para $i = 1, 2, \ldots, n$, o vetor escore $\boldsymbol{U}$ pode ser reescrito como:

$$\boldsymbol{U} = \frac{1}{\boldsymbol{\phi}} \boldsymbol{X}^T \boldsymbol{W} \boldsymbol{\Delta} (\boldsymbol{y} - \boldsymbol{\mu}),$$

com $\boldsymbol{\Delta} = diag\left\{\frac{d\eta_1}{d\mu_1}, \frac{d\eta_2}{d\mu_2}, ..., \frac{d\eta_n}{d\mu_n}\right\} = diag\{g'(\mu_1), g'(\mu_2), ..., g'(\mu_n)\}$. Substituindo $\boldsymbol{\mathfrak{F}}$ e $\boldsymbol{U}$ na equação (1) e rearranjando os termos, tem-se:

$$\boldsymbol{X}^T\boldsymbol{W}^{(m)}\boldsymbol{X}\boldsymbol{\beta}^{(m+1)} = \boldsymbol{X}^T\boldsymbol{W}^{(m)}[\boldsymbol{X}\boldsymbol{\beta}^{(m)} + \boldsymbol{\Delta}^{(m)}(\boldsymbol{y} - \boldsymbol{\mu})^{(m)}],$$

e, fazendo $\boldsymbol{z}^{(m)} = \boldsymbol{X}\boldsymbol{\beta}^{(m)} + \boldsymbol{\Delta}^{(m)}(\boldsymbol{y}-\boldsymbol{\mu})^{(m)} = \boldsymbol{\eta}^{(m)} + \boldsymbol{\Delta}^{(m)}(\boldsymbol{y}-\boldsymbol{\mu})^{(m)}$, que é chamada de variável dependente ajustada, tem-se:

$$\boldsymbol{\beta}^{(m+1)} = (\boldsymbol{X}^T\boldsymbol{W}^{(m)}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{W}^{(m)}\boldsymbol{z}^{(m)}.$$

Observa-se que o estimador obtido pelo modelo proposto independe de ϕ, e tem a forma da solução das equações normais para o modelo linear obtida pelo Método dos Mínimos Quadrados Ponderados, exceto que, neste modelo, o processo é iterativo.

O método usual para iniciar o processo é especificar um valor inicial β_0 e sucessivamente alterá-lo com as novas estimativas até a convergência ser obtida. O valor inicial geralmente utilizado é a média das respostas observadas. Dessa forma, definindo-se um β_0, pode-se resumir o algoritmo de estimação nos seguintes passos:

1. Obter as estimativas

$$\eta_i^{(m)} = \sum_{j=1}^{p} x_{ij}\beta_j^{(m)} \quad \text{e} \quad \mu_i^{(m)} = g^{-1}(\eta_i^{(m)}), \quad \text{para} \quad i = 1, 2, \ldots, n \quad \text{e} \quad m = 1, 2, \ldots$$

2. Obter a variável dependente ajustada e os pesos *a priori*

$$\begin{aligned} z_i^{(m)} &= \eta_i^{(m)} + (y_i - \mu_i^{(m)})g'(\mu_i^{(m)}) \quad \text{e} \\ W_i^{(m)} &= \frac{w_i}{V(\mu_i^{(m)})[g'(\mu_i^{(m)})]^2}, \quad \text{para} \quad i = 1, 2, \ldots, n \quad \text{e} \quad m = 1, 2, \ldots \end{aligned}$$

3. Calcular

$$\boldsymbol{\beta}^{(m+1)} = (\boldsymbol{X}^T\boldsymbol{W}^{(m)}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{W}^{(m)}\boldsymbol{z}^{(m)},$$

voltar ao passo (1) com $\boldsymbol{\beta}^{(m)} = \boldsymbol{\beta}^{(m+1)}$ e repetir o processo até a convergência, obtendo-se $\widehat{\boldsymbol{\beta}} = \boldsymbol{\beta}^{(m+1)}$.

Um critério para convergência do parâmetro pode ser:

$$\sum_{j=1}^{p}\left(\frac{\beta_j^{(m)}-\beta_j^{(m+1)}}{\beta_j^{(m)}}\right)^2<\xi,$$

tomando-se para ξ um valor suficientemente pequeno.

1.2 A função *Deviance*

Para medir a qualidade do ajuste de um Modelo Linear Generalizado, utiliza-se a função *deviance*. Segundo McCullagh & Nelder (1989), o ajuste de um modelo a um conjunto de dados consiste em substituir os valores observados de y pelos valores estimados por um modelo proposto com o menor número possível de parâmetros envolvidos. A questão é que, ao se fazer este ajuste, faz-se necessário medir se há uma discrepância relativamente pequena, entre o que se observa e o que se modela, para saber se de fato esta modelagem representa bem o que foi observado.

Admitindo-se uma combinação satisfatória entre a variável resposta e a função de ligação, deseja-se estudar a quantidade de parâmetros da parte linear necessários para explicar o modelo para a descrição razoável dos dados. Um número grande de covariáveis pode explicar bem o modelo, mas pode vir a complicar sua interpretação. Por outro lado, um número pequeno de covariáveis pode ter uma interpretação mais simples, e não revelar uma justificativa coerente para os dados. O que se deseja então, é encontrar um modelo que seja satisfatório em termos de interpretação e com o menor número de parâmetros.

Para um conjunto de n observações pode ser ajustado um modelo contendo até n parâmetros. O modelo mais simples a ser ajustado é o *modelo nulo*, que tem um único parâmetro representado por um valor comum a todos os parâmetros. Esse modelo atribui toda a variação entre as respostas ao componente aleatório. No outro extremo, pode-se construir um *modelo saturado* ou *completo* no qual tem-se n parâmetros, um para cada observação. Ele atribui toda a variação ao componente sistemático, reproduzindo os próprios dados.

Os modelos nulo e saturado representam os dois extremos. O que

se deseja é um modelo intermediário, que possua o menor número possível de parâmetros mas explique bem a resposta. Dessa forma, busca-se um modelo com p parâmetros linearmente independentes, o qual denomina-se *modelo corrente* ou *modelo sob pesquisa*.

Nelder & Wedderburn (1972) propõe como medida de discrepância a chamada *deviance*, que é dada por:

$$S_p = 2(\widehat{l_n} - \widehat{l_p}),$$

sendo $\widehat{l_n}$ e $\widehat{l_p}$ os máximos do logaritmo da função de verossimilhança para os modelos saturado e corrente, com n parâmetros para o modelo saturado e p parâmetros para o modelo corrente. Supondo $a_i = \dfrac{\phi}{w_i}$, tem-se que:

$$\widehat{l_n} = \frac{1}{\phi}\sum_{i=1}^{n}\{w_i[y_i\widetilde{\theta}_i - b(\widetilde{\theta}_i)] + c(y_i;\phi)\} \quad \text{e} \quad \widehat{l_p} = \frac{1}{\phi}\sum_{i=1}^{n}\{w_i[y_i\widehat{\theta}_i - b(\widehat{\theta}_i)] + c(y_i;\phi)\},$$

em que as estimativas do parâmetro canônico sob o modelo saturado é dada por $\widetilde{\theta}_i = \widetilde{\theta}(y_i)$ e do modelo corrente é dada por $\widehat{\theta}_i = \widehat{\theta}(\widehat{\mu_i})$, para $i = 1, 2, \ldots, n$. Dessa forma, tem-se que:

$$S_p = \frac{1}{\phi}\sum_{i=1}^{n} 2w_i\{y_i[\widehat{\theta}_i - \widehat{\theta}_i] - b(\widetilde{\theta}_i) + b(\widehat{\theta}_i)\} = \frac{1}{\phi}D_p,$$

em que S_p é chamada de *scaled deviance* e D_p a *deviance*.

Para o modelo de Poisson, tem-se que $a(\phi) = 1$ e a *deviance* assume o mesmo valor da *scaled deviance*, que é dada por:

$$D_p = S_p = 2\sum_{i=1}^{n}[y_i \ln\left(\frac{y_i}{\widehat{\mu_i}}\right) - (y_i - \widehat{\mu_i})].$$

A *deviance* cresce ou decresce de acordo com a entrada das covariáveis no modelo. Quanto maior for o número de covariáveis, menor o valor da *deviance*, mas a complexidade da interpretação dos dados aumenta. Quanto melhor for o ajuste do modelo, menor será o valor da *scaled deviance*. Portanto, na prática procura-se uma *deviance* moderada e um modelo que contemple os parâmetros necessários para ajustar os dados.

Para se testar a qualidade do ajuste de um modelo linear generalizado, o valor da *deviance* deve ser comparado a um valor de uma distribuição de probabilidade conhecida. A distribuição da *deviance* é desconhecida, mas, no caso da distribuição de Poisson, pode-se mostrar que ela tem uma distribuição assintótica qui-quadrado com $(n-p)$ graus de liberdade (Nelder & Wedderburn, 1972).

1.3 A distribuição binomial negativa

A origem da distribuição binomial negativa se dá no ano de 1679 com Pascal e Fermat, sendo também conhecida como distribuição de Pascal (Jonhson & Kotz, 1969). A distribuição binomial negativa surge como uma generalização da distribuição geométrica, que é definida como o tempo de espera do primeiro sucesso numa sequência de ensaios de Bernoulli com probabilidade p. De fato, a distribuição binomial negativa nada mais é do que a distribuição do tempo de espera do k-ésimo sucesso nessa mesma sequência de ensaios de Bernoulli com probabilidade p (Meyer, 1974).

Seguindo a mesma idéia, supõe-se que um experimento seja continuado até que um particular evento A ocorra na k-ésima vez (Meyer, 1974; Ross & Preece, 1985). Se $P(A) = p$ e $P(\overline{A}) = q = 1 - p$ em cada repetição, define-se a variável aleatória Y como sendo o número de vezes necessárias a fim de que A possa ocorrer exatamente k vezes. A distribuição de probabilidade de Y é dada por:

$$P(Y = y) = \begin{pmatrix} y+k-1 \\ k-1 \end{pmatrix} p^k q^y, \quad k = 0, 1, 2, \ldots.$$

Fazendo $p = \dfrac{k}{\mu + k}$ e $q = 1 - p$, a equação anterior pode ser reescrita como:

$$P(Y = y) = \frac{\Gamma(y+k)}{y!\Gamma(k)} \left(1 + \frac{\mu}{k}\right)^{-k} \left(\frac{\mu}{k+\mu}\right)^{y}, \quad y = 0, 1, 2, \ldots,$$

em que $\Gamma(.)$ é a função gama definida por:

$$\Gamma(z) = \int_0^\infty e^{-t} t^{z-1} dt, \quad \text{para} \quad z > 0.$$

Uma outra forma de obtenção da distribuição binomial negativa é dada em McCullagh & Nelder (1989). Seja uma variável aleatória Y cuja distribuição condicionada à uma variável aleatória Z, tenha uma distribuição de Poisson com média Z. Supondo agora que Z seja uma variável aleatória com distribuição gama com parâmetros μ e θ, definida por:

$$f(z) = \frac{\left(\frac{z}{\theta\mu}\right)^{\frac{1}{\theta}} \exp\left(\frac{-z}{\theta\mu}\right)}{\Gamma\left(\frac{1}{\theta}\right)} \frac{1}{z} \quad \text{para} \quad z > 0,$$

tem-se que a distribuição conjunta de Y e Z é dada por:

$$f(y,z) = f(z).f(y/z) = \frac{\left(\frac{z}{\theta\mu}\right)^{\frac{1}{\theta}} \exp\left(\frac{-z}{\theta\mu}\right)}{\Gamma\left(\frac{1}{\theta}\right)} \frac{1}{z} \frac{e^{-z}z^y}{y!} \quad \text{para} \quad y = 0,1,2,\ldots \quad \text{e} \quad z > 0.$$

Então, a distribuição marginal de Y pode ser obtida por:

$$\begin{aligned} f(y) &= \int_0^\infty \frac{e^{-z-\frac{z}{\theta\mu}} z^y \left(\frac{z}{\theta\mu}\right)^{\frac{1}{\theta}}}{y!\Gamma\left(\frac{1}{\theta}\right) z} dz = \frac{\left(\frac{1}{\theta\mu}\right)^{\frac{1}{\theta}}}{y!\Gamma\left(\frac{1}{\theta}\right)} \int_0^\infty e^{-z\left(1+\frac{1}{\theta\mu}\right)} z^{y+\frac{1}{\theta}-1} dz = \\ &= \frac{\Gamma\left(y+\frac{1}{\theta}\right)}{y!\Gamma\left(\frac{1}{\theta}\right)} \left(\frac{1}{1+\theta\mu}\right)^{\frac{1}{\theta}} \left(\frac{\mu}{\frac{1}{\theta}+\mu}\right)^y \quad \text{para} \quad y = 0,1,2,\ldots. \end{aligned}$$

A equação anterior pode ser reparametrizada, fazendo $k = \dfrac{1}{\theta}$:

$$f(y) = \frac{\Gamma(y+k)}{y!\Gamma(k)} \left(1+\frac{\mu}{k}\right)^{-k} \left(\frac{\mu}{k+\mu}\right)^y \quad \text{para} \quad y = 0,1,2,\ldots,$$

que é a forma de uma distribuição binomial negativa, em que k é o parâmetro de dispersão. Esta distribuição tem valor esperado $E(Y) = \mu$ e a variância $Var(Y) = \mu\left(1+\dfrac{\mu}{k}\right)$.

O modelo binomial negativo apresenta o mesmo valor esperado do modelo de Poisson, mas apresenta uma modificação na variância, a qual pode ser uma alternativa para modelos com superdispersão.

1.3.1 Estimação dos parâmetros na distribuição binomial negativa

Whitaker & Dickens (1972) apresentam uma alternativa de estimação

para os parâmetros da distribuição binomial negativa feita pelo Método dos Momentos. Considerando o modelo binomial negativo com parâmetros μ e k:

$$f(y) = \frac{\Gamma(y+k)}{y!\Gamma(k)} \left(\frac{k}{k+\mu}\right)^k \left(\frac{\mu}{\mu+k}\right)^y \quad \text{para} \quad y = 0, 1, 2, \ldots.$$

Para a estimação pelo método dos momentos, fazendo-se $p = \dfrac{k}{k+\mu}$, $q = \dfrac{\mu}{\mu+k}$ e $q = 1 - p$, obtém-se:

$$f(y) = \frac{\Gamma(y+k)}{y!\Gamma(k)} p^k q^y \quad \text{para} \quad y = 0, 1, 2, \ldots.$$

O primeiro momento centrado no zero é dado por:

$$E(Y) = \frac{kq}{p} = \mu.$$

Como o segundo momento centrado no zero é dado por:

$$E(Y^2) = \frac{kq}{p^2} + \frac{k^2q^2}{p^2},$$

tem-se que:

$$Var(Y) = \frac{kq}{p^2} = \mu + \frac{\mu^2}{k} = \sigma^2.$$

O Método dos Momentos propõe usar o valor da média amostral $\overline{y}$ para estimar os parâmetros μ e k. A partir do primeiro momento centrado no zero tem-se:

$$\widehat{\mu} = \overline{y}.$$

A variância σ^2 é igual a variância da média amostral $\sigma^2_{\overline{y}}$ vezes o tamanho amostral n, ou seja, $s^2 = ns^2_{\overline{y}}$. Dessa forma, obtém-se uma estimativa para o parâmetro k dada por:

$$\widehat{k} = \frac{\overline{y}^2}{ns^2_{\overline{y}} - \overline{y}}.$$

1.4 Excesso de zeros

Um conjunto de dados pode apresentar uma variabilidade maior do que a esperada pelos modelos probabilísticos padrões. Esse fenômeno é conhecido

como superdispersão, podendo ser ocasionada por diversos fatores. Um deles pode ser devido ao excesso de zeros nos dados (Borgatto, 2004). Para estes casos, pode-se propor uma modelagem feita a partir dos chamados modelos inflacionados de zeros, nos quais os dados apresentam um ajuste em duas partes: uma para as contagens nulas e outra para as não-nulas (Lambert, 1992).

As contagens nulas que geram o excesso de zeros podem ocorrer de duas formas distintas. A primeira delas se dá no caso em que há falta de uma determinada característica presente na população (denominados zeros estruturais); a outra razão, seria devido a ausência de determinada característica no período do estudo (denominados zeros amostrais). Por exemplo, considere o número de crianças nascidas numa região. Das mulheres consideradas nessa região, tem-se aquelas que não são mães. O fato de não ter filhos pode ser ocasionado por dois fatores: a mulher não é mãe porque não pode ter filhos (zeros estruturais) ou porque a mulher ainda não teve filhos (zeros amostrais). Nos dois casos a contagem é nula mas a natureza desta nulidade é distinta (Poston & McKibben, 2003).

Desse modo, pode-se atribuir uma probabilidade p para zeros estruturais e $1-p$ para amostrais, sendo a modelagem da resposta nula dada por:

$$P(Y=0) = p + (1-p)R(0),$$

em que $R(0)$ é a probabilidade da resposta ser igual a zero para o modelo usual de Poisson ou Binomial Negativo. Para as contagens não-nulas, tem-se:

$$P(Y>0) = (1-p)R(Y),$$

em que $R(Y)$ é a probabilidade da resposta não nula, que possui uma distribuição usual de Poisson ou Binomial Negativa (Chin & Quddus, 2003). Assim, o modelo inflacionado segue a estrutura:

$$P(Y=y) = \begin{cases} p + (1-p)R(0) & \text{para} \quad y = 0 \\ (1-p)R(Y) & \text{para} \quad y \geq 1. \end{cases}$$

em que R(.) tem uma distribuição usual de Poisson ou Binomial Negativa.

2 OBJETIVOS

2.1 Objetivo Geral

Este trabalho tem como objetivo estudar os modelos inflacionados de zeros para dados de contagem e aplicá-los a dados de um Questionário de Frequência Alimentar (QFA), de modo a verificar os fatores que influenciam no consumo ou não de determinados alimentos por idosos pertencentes a uma cidade de porte médio do interior do Estado de São Paulo, Brasil.

2.2 Objetivos Específicos

1. Estudar os métodos de estimação dos parâmetros utilizados nos modelos inflacionados de zeros;

2. Avaliar a inflação de zeros nos dados do QFA;

3. Utilizar o programa SAS *for Windows*, versão 9.1.3, no ajuste dos modelos usuais para a distribuição de Poisson e Binomial Negativa para os dados de contagem obtidos no QFA para diversos alimentos;

4. Utilizar o programa Stata 9.1 para ajustar modelos inflacionados de zeros aos dados do QFA para os alimentos consumidos esporadicamente e aplicar testes referentes à modelos inflacionados *versus* modelos não inflacionados;

5. Avaliar o uso de modelos inflacionados em relação ao percentual de zeros presente nos dados.

3 METODOLOGIA

3.1 Modelos Inflacionados de Zeros

Os modelos inflacionados de zeros seguem uma estrutura composta por duas partes:

1. A contagem nula, a qual pode ser subdividida em duas partes:

   i) Zeros estruturais: pertencentes a estrutura de zeros dos dados;

   ii) Zeros amostrais: pertencentes a distribuição de Poisson ou Binomial Negativa quando a resposta é nula.

2. A contagem não-nula, na qual o modelo segue uma distribuição de Poisson ou Binomial Negativa.

3.2 Modelo de Poisson Inflacionado de Zeros (ZIP)

Uma variável aleatória Y segue uma distribuição de Poisson Inflacionada de Zeros, com parâmetro μ, probabilidade p para os zeros estruturais e $(1-p)$ para os zeros amostrais (Nagamine et al., 2008; Hall, 2000), se:

$$P(Y=y) = \begin{cases} p + (1-p)e^{-\mu} & \text{para} \quad y = 0 \\ \dfrac{(1-p)e^{-\mu}\mu^y}{y!} & \text{para} \quad y \geq 1. \end{cases}$$

O valor esperado e a variância de Y com distribuição de Poisson Inflacionada de Zeros são dados, respectivamente, por:

$$E(Y) = (1-p)\mu \quad \text{e} \quad Var(Y) = \mu(1-p)(1+p\mu).$$

Nota-se que, nesta distribuição, o valor esperado e a variância contemplam não somente a média, mas também a proporção de zeros (ver Apêndice A).

Na associação com covariáveis, tem-se o Modelo de Poisson Inflacionado de Zeros (*ZIP- Zero Inflated Poisson*), que abrange as contagens nulas respeitando a natureza distinta dos zeros (amostrais e estruturais) e as contagens não nulas que seguem um modelo de Poisson usual. No ajuste de um modelo ZIP, leva-se em conta duas funções de ligação (Ridout et al., 1998):

1. Para a parte inflacionada de zeros, a função de ligação da média ao preditor linear é dada pela ligação logit, definida como:

$$logit(\boldsymbol{p}) = \ln\left(\frac{\boldsymbol{p}}{1-\boldsymbol{p}}\right) = \boldsymbol{G\gamma},$$

   em que $\boldsymbol{G}$ é uma matriz de covariáveis e $\boldsymbol{\gamma}$ o vetor dos parâmetros desconhecidos das covariáveis associadas à parte inflacionada de zeros do modelo.

2. Para a parte não inflacionada de zeros, a função de ligação é a mesma utilizada no modelo usual de Poisson, ou seja,

$$\ln(\boldsymbol{\mu}) = \boldsymbol{B\beta},$$

   em que $\boldsymbol{B}$ é uma matriz de covariáveis associada e $\boldsymbol{\beta}$ é o vetor de parâmetros desconhecidos correspondente a parte não inflacionada do modelo.

3.3 Estimação dos parâmetros no Modelo Poisson Inflacionado de Zeros

3.3.1 Estimação dos parâmetros no modelo ZIP sem covariáveis

A obtenção dos estimadores de p e μ de um Modelo de Poisson Inflacionado de zeros sem as covariáveis pode ser feito pelo Método da Máxima Verossimilhança (Nagamine et al., 2008). Seja $Y_1, Y_2, \ldots, Y_n$ uma amostra de tamanho n, com n_0 observações iguais a zero, e n_y observações diferentes de zero para $y = 1, 2, \ldots,$

tal que n_1 são observações iguais a 1, n_2 são as observações iguais a 2 e assim por diante, de forma que:

$$n = n_0 + \sum_{y=1}^{n} n_y.$$

O logaritmo da função de verossimilhança é dado por:

$$l(p,\mu) = \sum_{i=1}^{n_0} \ln(p + (1-p)e^{-\mu}) + \sum_{i=1}^{n_y} \ln\left((1-p)\frac{e^{-\mu}\mu^{y_i}}{y_i!}\right).$$

As derivadas parciais em relação aos parâmetros p e μ são dadas, respectivamente, por:

$$\frac{\partial l(p,\mu)}{\partial p} = \frac{n_0(1-e^{-\mu})}{p+(1-p)e^{-\mu}} - \frac{n_y}{1-p},$$

$$\frac{\partial l(p,\mu)}{\partial \mu} = \frac{-n_0(1-p)e^{-\mu}}{p+(1-p)e^{-\mu}} + \sum_{i=1}^{n_y} \frac{(y_i-\mu)}{\mu}.$$

Os estimadores de Máxima Verossimilhança de p e μ são obtidos de maneira explícita, dados por:

$$\widehat{p} = \frac{\frac{n_0}{n} - e^{-\mu}}{1-e^{-\mu}} \quad \text{e} \quad \widehat{\mu} = \frac{\overline{y}}{1-\widehat{p}},$$

em que $\overline{y} = \dfrac{1}{n}\sum_{i=1}^{n} y_i$.

3.3.2 Estimação dos parâmetros no modelo ZIP com covariáveis

Seja $Y_1, Y_2, \ldots, Y_n$ uma amostra aleatória seguindo um Modelo de Poisson Inflacionado de Zeros, com parâmetros $\boldsymbol{\mu} = (\mu_1, \mu_2, ..., \mu_n)'$, $\boldsymbol{p} = (p_1, p_2, ..., p_n)'$ e funções de ligação, $\ln(\boldsymbol{\mu}) = \boldsymbol{B\beta}$ para a parte não inflacionada e $logit(\boldsymbol{p}) = \ln\left(\dfrac{\boldsymbol{p}}{1-\boldsymbol{p}}\right) = \boldsymbol{G\gamma}$ para a parte inflacionada, sendo $\boldsymbol{B}$ e $\boldsymbol{G}$ as matrizes de covariáveis.

De acordo com Lambert (1992), as covariáveis podem ou não serem as mesmas para parte não inflacionada e inflacionada. Duas situações distintas podem ser consideradas na estimação dos parâmetros:

1. O vetor de parâmetros $\boldsymbol{p}$ pode não estar relacionado com o vetor de parâmetros $\boldsymbol{\mu}$. Isto resulta na estimação de um número duas vezes maior de parâmetros

presentes numa regressão de Poisson usual, já que é necessária uma estimação dos parâmetros da parte inflacionada e outra estimação da parte não inflacionada;

2. O vetor de parâmetros $\boldsymbol{p}$ pode estar relacionado ao vetor de parâmetros $\boldsymbol{\mu}$. Se isso acontece, o tempo computacional é diminuído e novas parametrizações das funções de ligação são dadas por:

$$\ln(\boldsymbol{\mu}) = \boldsymbol{B\beta} \quad \text{e} \quad logit(\boldsymbol{p}) = -\boldsymbol{\tau B\beta}.$$

em que $\boldsymbol{p} = (1 + \boldsymbol{\mu}^{\boldsymbol{\tau}})^{-1}$, no qual $\boldsymbol{\tau}$ é um vetor de parâmetros de forma desconhecido (ver Apêndice B).

O primeiro modelo é conhecido por ZIP (*Zero-Inflated Poisson*) e o segundo por ZIP($\boldsymbol{\tau}$). Na sequência, apresentam-se os processos de estimação para os dois casos.

1. **$\boldsymbol{\mu}$ e $\boldsymbol{p}$ não são relacionados**

Quando $\boldsymbol{\mu}$ e $\boldsymbol{p}$ não são relacionados, o logaritmo da função de verossimilhança do modelo ZIP com a parametrização padrão é dado por:

$$\begin{aligned} L(\boldsymbol{\gamma}, \boldsymbol{\beta}; \boldsymbol{y}) &= \ln\{\prod_{y=0}[p + (1-p)e^{-\mu}] \prod_{y\geq 1}^{n}[(1-p)\frac{e^{-\mu}\mu^y}{y!}]\} = \\ &= \sum_{y=0} \ln(e^{G_i\boldsymbol{\gamma}} + \exp(-e^{B_i\boldsymbol{\beta}})) + \sum_{y>0}(yB_i\boldsymbol{\beta} - e^{B_i\boldsymbol{\beta}}) - \\ &- \sum_{i=1}^{n} \ln(1 + e^{G_i\boldsymbol{\gamma}}) - \sum_{y>0} \ln y!, \end{aligned}$$

em que G_i e B_i são colunas das matrizes de covariáveis, $i = 1, 2, \ldots, p$. A soma de exponenciais no primeiro termo dificultam a maximização de $L(\boldsymbol{\gamma}, \boldsymbol{\beta}; \boldsymbol{y})$. Mas, supondo que se conhece quais zeros vem da parte estrutural e quais provém da distribuição de Poisson, supõe-se $Z_i = 1$ quando $Y = 0$ e $Z_i = 0$ quando Y provém de uma distribuição de Poisson, para $i = 1, 2, \ldots, n$. O logaritmo da função de verossimilhança será dado por:

$$
\begin{aligned}
L(\boldsymbol{\gamma},\boldsymbol{\beta};\boldsymbol{y},\boldsymbol{z}) &= \ln[\prod_{i=1}^{n} f(y,z,\boldsymbol{\gamma},\boldsymbol{\beta})] = \sum_{i=1}^{n} \ln[f(y,z,\boldsymbol{\gamma},\boldsymbol{\beta})] = \\
&= \sum_{i=1}^{n} \ln[f(y/z,\boldsymbol{\beta}) f(z/\boldsymbol{\gamma})] = \sum_{i=1}^{n} [z_i G_i \boldsymbol{\gamma} - \ln(1+e^{G_i\boldsymbol{\gamma}})] + \\
&+ \sum_{i=1}^{n} (1-z_i)(yB_i\boldsymbol{\beta} - e^{B_i\boldsymbol{\beta}}) - \sum_{i=1}^{n} (1-z_i)\ln y! = \\
&= L_c(\boldsymbol{\gamma};\boldsymbol{y},\boldsymbol{z}) + L_c(\boldsymbol{\beta};\boldsymbol{y},\boldsymbol{z}) - \sum_{i=1}^{n} (1-z_i)\ln y!.
\end{aligned}
$$

Desse modo, o logaritmo da função de verossimilhança é de fácil estimação, pois $L_c(\boldsymbol{\gamma};\boldsymbol{y},\boldsymbol{z})$ e $L_c(\boldsymbol{\beta};\boldsymbol{y},\boldsymbol{z})$ podem ser maximizados separadamente. O método de estimação é feito pelo algoritmo EM (Dempster et al., 1977; Vieira, 1998) que consiste, basicamente, em um processo iterativo de dois passos: E(*Expectation*), em que o valor condicional da variável z é calculado, e M(*Maximization*), que é a etapa de Maximização, na qual utiliza-se os dados observados e os estimados do passo E. Uma vez que os valores esperados das $z's$ convergem, a estimação de $(\boldsymbol{\gamma},\boldsymbol{\beta})$ convergem e as iterações param. A estimação na última iteração é o valor dos parâmetros estimados $(\widehat{\boldsymbol{\gamma}},\widehat{\boldsymbol{\beta}})$.

Em mais detalhes, a iteração $(k+1)$ do algoritmo EM requer, para o modelo ZIP, três etapas:

Passo E: Estima-se o valor esperado de Z sob as estimativas correntes de $\gamma^{(k)}$ e $\beta^{(k)}$. Assim:

$$
\begin{aligned}
Z^{(k)} &= P(\text{zero estrutural}|y;\boldsymbol{\gamma}^{(k)},\boldsymbol{\beta}^{(k)}) = \\
&= \frac{P(y|\text{zero estrutural})P(\text{zero estrutural})}{P(y|\text{zero estrutural})P(\text{zero estrutural}) + P(y|\text{Poisson})P(\text{Poisson})} = \\
&= (1+e^{-\boldsymbol{G}\boldsymbol{\gamma}^{(k)} - \exp(\boldsymbol{B}\boldsymbol{\beta}^{(k)})})^{-1} \quad \text{se} \quad y=0, \\
&= 0 \quad \text{se} \quad y=1,2,\ldots \quad ;
\end{aligned}
$$

Passo M para $\boldsymbol{\beta}$: Encontrar $\boldsymbol{\beta}^{(k+1)}$ para maximização $L_c(\boldsymbol{\beta};\boldsymbol{y},\boldsymbol{Z}^{(k)})$. Note que $\boldsymbol{\beta}^{(k+1)}$ pode ser encontrado proveniente de um modelo de mínimos quadrados

ponderados. A regressão aqui proposta para o Modelo log-linear Poisson tem pesos $(1 - \boldsymbol{Z}^{(k)})$.

Passo M para $\boldsymbol{\gamma}$*:* Maximizar $L_c(\boldsymbol{\gamma}; \boldsymbol{y}, \boldsymbol{Z}^{(k)}) = \sum_{y=0} \boldsymbol{Z}^{(k)} \boldsymbol{G\gamma} - \sum_{y=0} \boldsymbol{Z}^{(k)} \ln(1 + e^{\boldsymbol{G\gamma}}) - \sum_{i=1}^{n} (1 - \boldsymbol{Z}^{(k)}) \ln(1 + e^{\boldsymbol{G\gamma}})$ como uma função de $\boldsymbol{\gamma}$. Esta igualdade ocorre para $Z^{(k)} = 0$ sempre que $y > 0$.

No terceiro passo determina-se a probalidade da ocorrência de todos os zeros, a qual é feita pela combinação entre as probabilidades de zeros encontradas em cada grupo, ponderada pela observação de cada elemento individualmente. Neste passo, usa-se o modelo de regressão logística apresentando como variável resposta Z.

2. ***p* como função de $\boldsymbol{\mu}$**

A verossimilhança para o modelo ZIP($\boldsymbol{\tau}$) é dada por:

$$\sum_{y=0} \ln(e^{-\boldsymbol{\tau} \boldsymbol{B}_i \boldsymbol{\beta}} + \exp(-e^{\boldsymbol{B}_i \boldsymbol{\beta}})) + \sum_{y>0} (y \boldsymbol{B}_i \boldsymbol{\beta} - e^{\boldsymbol{B}_i \boldsymbol{\beta}}) - \sum_{i=1}^{n} \ln(1 + e^{-\boldsymbol{\tau} \boldsymbol{B}_i \boldsymbol{\beta}}).$$

Neste caso, $\boldsymbol{\beta}$ e $\boldsymbol{\gamma}$ são estimados pelo algoritmo de Newton-Raphson com poucas iterações quando se supõe $\boldsymbol{\beta}^{(0)} = \widehat{\boldsymbol{\beta}}_u$ como sendo $\boldsymbol{\tau}^{(0)} = -mediana(\widehat{\boldsymbol{\tau}}_{(u)} / \widehat{\boldsymbol{\beta}}_u)$. Daí, $(\widehat{\boldsymbol{\tau}}_u, \widehat{\boldsymbol{\beta}}_u)$ são os estimadores de máxima verossimilhança encontrados no modelo ZIP $(\boldsymbol{\tau})$. Outra alternativa para iniciar o processo de estimação é encontrar $\boldsymbol{\beta}$ para alguns valores fixos de $\boldsymbol{\tau}_{(0)}$ e então tomar como ponto inicial o valor $(\widehat{\boldsymbol{\beta}}(\boldsymbol{\tau}_0), \boldsymbol{\tau}_0)$, e iniciar o processo iterativo.

Os detalhes das estimações dos parâmetros podem ser vistos no Apêndice C.

3.4 Modelo Binomial Negativo Inflacionado de Zeros (ZINB)

Uma variável aleatória Y tem distribuição Binomial Negativa Inflacionada de Zeros (ZINB), com parâmetros μ, k e probabilidade p para os zeros

estruturais e $1-p$ para os zeros amostrais (Cheung, 2002), se:

$$P(Y=y)=\begin{cases} p+(1-p)\left(\dfrac{k}{k+\mu}\right)^k & \text{para} \quad y=0 \\ (1-p)\dfrac{\Gamma(y+k)}{\Gamma(k)y!}\left(\dfrac{k}{k+\mu}\right)^k\left(\dfrac{\mu}{k+\mu}\right)^y & \text{para} \quad y\geq 1. \end{cases}$$

Analogamente ao modelo ZIP, o seu valor esperado e sua variância consideram, além da média, a porcentagem de zeros, sendo dados por $E(Y)=(1-p)\mu$ e $Var(Y)=(1-p)\mu\left(\mu p+1+\dfrac{\mu}{k}\right)$, respectivamente.

Na associação com covariáveis, tem-se que o Modelo Binomial Negativo Inflacionado de Zeros (ZINB), analogamente ao modelo ZIP, apresenta uma modelagem para contagens nulas, respeitando a natureza distintas dos zeros (amostrais e estruturais) e, para as contagens não nulas, segue um modelo Binomial Negativo usual. As funções de ligação também são análogas ao modelo ZIP, ou seja, para as contagens nulas usa-se a ligação logística e para as não nulas, a ligação logarítmica.

3.5 Estimação dos parâmetros no Modelo Binomial Negativo Inflacionado de Zeros

3.5.1 Estimação dos parâmetros no modelo ZINB sem covariáveis

A estimação dos parâmetros p, μ e k de um Modelo Binomial Negativo Inflacionado de Zeros sem covariáveis, pode ser feito pelo Método da Máxima Verossimilhança. Nagamine et al. (2008) considera $Y_1, Y_2, \ldots, Y_n$ uma amostra de tamanho n, com n_0 observações iguais a zeros e n_y observações diferentes de zero, para $y=1,2,\ldots$, o logaritmo da função de verossimilhança é dado por:

$$\begin{aligned} l(y,p,\mu,k) &= n_0 \ln\left(p+(1-p)\left(\frac{k}{k+\mu}\right)^k\right)+ \\ &+ \sum_{y=1}^{\infty} n_y \ln\left[\frac{(1-p)\Gamma(y+k)}{\Gamma(k)y!}\left(\frac{k}{k+\mu}\right)^k\left(\frac{\mu}{k+\mu}\right)^y\right]. \end{aligned}$$

As derivadas parciais de $l(y, p, \mu, k)$ com p, μ e k são, respectivamente:

$$\frac{\partial l(y,p,\mu,k)}{\partial p} = \frac{n_0\left(1-\left(\frac{k}{k+\mu}\right)^k\right)}{p+(1-p)\left(\frac{k}{k+\mu}\right)^k} - \sum_{y=1}^{\infty}\frac{n_y}{(1-p)},$$

$$\frac{\partial l(y,p,\mu,k)}{\partial \mu} = \frac{n_0(1-p)\left(\frac{k}{k+\mu}\right)^k}{(k+\mu)\left[p+(1-p)\left(\frac{k}{k+\mu}\right)^k\right]} + \sum_{y=1}^{\infty}\frac{n_y}{k+\mu} + \sum_{y=1}^{\infty}\frac{n_y y k}{\mu(k+\mu)},$$

$$\begin{aligned}\frac{\partial l(y,p,\mu,k)}{\partial k} &= \frac{n_0(1-p)\left(\frac{k}{k+\mu}\right)^k}{p+(1-p)\left(\frac{k}{k+\mu}\right)^k}\left[(k+\mu)\ln\left(\frac{k}{k+\mu}\right)+\mu\right] + \\ &+ \sum_{y=1}^{\infty} n_y\left\{\frac{\Gamma(y+k)}{\Gamma(k)y!}[\psi(y+k)-\psi(k)]\right\}\left[\left(\frac{k}{k+\mu}\right)^{k\left(\frac{\mu}{\mu+k}\right)}\right]^y,\end{aligned}$$

em que $\psi(.)$ é a função digama, dada por $\psi(.) = \dfrac{\partial \ln[\Gamma(.)]}{\partial(.)}$.

Igualando as derivadas anteriores a zero, obtém-se as equações que dão origem as estimativas dos parâmetros na ausência das covariáveis. Essas equações são resolvidas por métodos iterativos.

3.5.2 Estimação dos parâmetros no modelo ZINB com covariáveis

Em Minami et al. (2007), a estimação dos parâmetros do Modelo Binomial Negativo Inflacionado de Zeros na presença de covariáveis segue a mesma lógica proposta por Lambert (1992) para o Modelo de Poisson Inflacionado de Zeros (ZIP). Considere $Y_1, Y_2, \ldots, Y_n$ uma amostra aleatória seguindo um Modelo Binomial Negativo Inflacionado de Zeros, com parâmetros $\boldsymbol{\mu}$, $\boldsymbol{p}$ e $\boldsymbol{k}$. As covariáveis relacionadas com a média para as partes não inflacionadas e inflacionadas são dadas, respectivamente, por:

$$\ln(\boldsymbol{\mu}) = \boldsymbol{B\beta} \quad \text{e} \quad logit(\boldsymbol{p}) = \ln\left(\frac{\boldsymbol{p}}{1-\boldsymbol{p}}\right) = \boldsymbol{G\gamma},$$

em que $\boldsymbol{B}$ e $\boldsymbol{G}$ são as matrizes de covariáveis.

As covariáveis em $\boldsymbol{B}$ e $\boldsymbol{G}$ podem ser diferentes dependendo do processo através do qual se obtém os dados. Estimadores para $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$ e $\boldsymbol{k}$ podem

ser obtidos por maximização da função de verossimilhança $L(\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{k}|\boldsymbol{y}, \boldsymbol{B}, \boldsymbol{G}) = \sum_{i=1}^{n} \ln f(\boldsymbol{y}|\boldsymbol{B}_i, \boldsymbol{G}_i, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{k})$. Neste caso o algoritmo proposto para estimação é o mesmo do modelo ZIP, o algoritmo EM.

Dessa forma, para resolução do algoritmo usa-se o artifício da introdução de uma variável aleatória Z que é uma variável indicadora da resposta, a qual assume 1 se a observação pertence a parte inflacionada de zeros, e 0 em caso contrário.

Analogamente ao modelo ZIP, quando a parte inflacionada e não inflacionada apresentam seus parâmetros não relacionados, dada uma estimativa para Z, a iteração do algoritmo EM se reduz para a estimação do vetor de parâmetros $\boldsymbol{\gamma}$ por um modelo de regressão logística com Z como variável resposta, a estimação dos vetores $\boldsymbol{\beta}$ e $\boldsymbol{k}$ se dão pelo modelo de regressão binomial negativo com y como variável resposta e $1 - Z$ como os pesos.

Mais especificamente, na k-ésima iteração, o algoritmo EM apresenta os seguintes passos:

Passo E: Estima-se $z_i^{(k)}$ = esperança condicional de Z_i, assim:

$$\begin{aligned} Z_i &= E[Z_i|y_i, \boldsymbol{\beta}^{(k-1)}, \boldsymbol{\gamma}^{(k-1)}, \boldsymbol{k}^{(k-1)}] = \\ &= \frac{p_i^{(k-1)}}{p_i^{(k-1)} + (1 - p_i^{(k-1)}) q(0|\mu_i^{(k-1)}, k^{(k-1)})} \quad \text{para} \quad y_i = 0 \\ &= 0 \quad \text{para} \quad y_i = 1, 2, \ldots \quad . \end{aligned}$$

Passo M para $\boldsymbol{\beta}$: Obtém-se estimativas $\boldsymbol{\beta}^{(k)}$ por ajuste do modelo de regressão binomial negativo usando pesos $1 - Z_i^{(k)}$ e variável resposta y_i. Atualiza-se a estimação de $\boldsymbol{k}$, $\boldsymbol{k}^{(k)}$, se $\boldsymbol{k}$ não é conhecido.

Passo M para $\boldsymbol{\gamma}$: Obtém-se estimativas $\boldsymbol{\gamma}^{(k)}$ por ajuste do modelo logístico de regressão, usando $Z_i^{(k)}$ como variável resposta.

3.6 O Teste de Vuong

O excesso de zeros apresentado nos dados pode gerar uma superdispersão, mas existem casos em que essa superdispersão não é grande e não afeta o

ajuste do modelo. Vuong (1989) mostra, em detalhes, a construção de um teste para testar modelos não aninhados, o qual baseia-se numa estatística t para comparar modelos usuais e inflacionados de zeros. Em particular, pode-se utilizar esses testes nos modelos de Poisson Inflacionado de Zeros e Poisson, bem como nos modelos Binomial Negativo Inflacionado de Zeros e Binomial Negativo. V é a estatística padrão para testar a hipótese que os modelos são equivalentes, dada por:

$$V = \frac{\sqrt{N}}{s_m}\overline{m},$$

em que cada $m_i = \ln\left(\frac{f_1(y_i)}{f_2(y_i)}\right)$, $i = 1, 2, \ldots, n$, composto por f_1 que é a probabilidade de um modelo inflacionado e f_2 a probabilidade de um modelo usual, $\overline{m}$ é a média e s_m o desvio padrão das $m_i's$ e N é o tamanho da amostra.

Se a hipótese nula é Ho: $E(m_i) = 0$, então Vuong (1989) mostra que assintoticamente V tem uma distribuição normal padrão. Desta forma, se $|V|$ é menor que um valor crítico predeterminado, então o teste resulta não favorável nem a um modelo nem ao outro. Em contrapartida, valores positivos grandes favorecem ao modelo inflacionado enquanto valores negativos grandes favorecem o modelo usual (Greene, 1994).

Para os modelos descritos no estudo, as estatísticas de teste serão compostas para $m_i's$ da forma:

$$m_i = \ln\left(\frac{f_1(\text{ZIP})}{f_2(\text{Poisson})}\right) \quad \text{e} \quad m_i = \ln\left(\frac{f_1(\text{ZINB})}{f_2(\text{binomial negativa})}\right),$$

para comparação entre os modelos que envolvem a distribuição de Poisson e binomial negativa, respectivamente.

3.7 AIC e BIC

Os critérios de *AIC - Akaike Information Criterion* e *BIC - Bayes Information Criterion* são amplamente utilizados na comparação entre modelos aninhados e não aninhados (Kuha, 2004).

Estes critérios são dados por:

$$AIC = 2[l(\widehat{\theta}_2) - l(\widehat{\theta}_1)] - 2(p_2 - p_1),$$

$$BIC = 2[l(\widehat{\theta}_2) - l(\widehat{\theta}_1)] - logn(p_2 - p_1),$$

em que $l(\widehat{\theta}_2)$ e $l(\widehat{\theta}_1)$ são os logaritmos das funções de verossimilhança de dois modelos M_1 e M_2, p_2 e p_1 são os graus de liberdade dos respectivos modelos e n é o número de observações. Nota-se que o AIC leva em consideração o número de parâmetros p do modelo e o BIC o número p de parâmetros e o número n de observações.

Para dois modelos com o mesmo conjunto de dados, o melhor modelo é aquele que possui o menor valor de critério de informação considerado. Neste trabalho são utilizados esses critérios para comparar os modelos inflacionados de zeros.

3.8 Questionário de Frequência Alimentar (QFA)

Para a aplicação dos modelos usuais e inflacionados de Poisson e Binomial Negativo, foram utilizados dados de um Questionário de Frequência Alimentar (QFA) provenientes de um estudo transversal para avaliar a qualidade de vida de idosos do município de Avaré, localizado no interior do Estado de São Paulo, Brasil.

O tamanho amostral inicialmente calculado foi de 365 idosos, sendo considerada uma prevalência de satisfação com a vida de 0,5 (maior valor de prevalência por ser uma característica desconhecida na população), uma margem de erro amostral de 5%, valor z igual a 1,96 relativo a um intervalo de confiança de 95% bilateral. O sorteio dos participantes da amostra foi feito estratificando por faixa etária, de dez em dez anos. Para o presente estudo, da amostra inicialmente calculada foram sorteados 20%, perfazendo um total de 73 indivíduos de ambos os sexos. O sorteio dos participantes foi novamente estratificado por faixa etária. Os dados foram coletados para um estudo de validação sobre o Questionário de Frequência Alimentar (Projeto CNPq nº402533/2007-0), para o qual recomenda-se um tamanho

amostral variando de 50 a 100 pessoas do grupo demográfico em questão (Slater et al., 2003).

A variável resposta utilizada na aplicação dos modelos usuais e inflacionados foi a frequência mensal do consumo de alimentos provenientes do QFA e as covariáveis associadas estão vinculadas às características sociodemográficas, de prevenção e morbidades referidas.

Dos 67 ítens obtidos no QFA, alguns foram selecionados para este estudo segundo consulta a profissionais da área nutricional. Um modelo do questionário aplicado encontra-se no Anexo A.

3.9 Taxa de extra-variação

Os dados provenientes desse questionário de frequência alimentar, por serem dados de contagem, podem apresentar uma extra-variação. Bohning et al. (1997) propõem um cálculo do quanto essa taxa de extra-variação pode ser atribuída ao excesso de zeros. Esta medida baseia-se na variância do modelo de Poisson Inflacionado de Zeros (ZIP) e na diferença entre a média e a variância da variável em questão.

A variância no Modelo ZIP é dada por:

$$\begin{aligned} Var(Y) &= \mu(1-p)(1+p\mu) = (\mu-\mu p)(1+p\mu) = \mu + p\mu^2 - \mu p - \mu^2 p^2 = \\ &= \mu(1-p) + \mu p(\mu - \mu p) = E(Y) + E(Y)(\mu - E(Y)), \end{aligned}$$

em que $E(Y)$ é o valor esperado de uma variável aleatória Y seguindo uma distribuição de Poisson inflacionada de zeros. Os estimadores de máxima verossimilhança de μ e p são dados, respectivamente por, $\widehat{\mu}$ e $\widehat{p}$. Daí, $\widehat{E(Y)} = (1-\widehat{p})\widehat{\mu}$.

No modelo de Poisson, sabe-se que a variância é igual ao valor esperado. No caso de uma superdispersão, a variabilidade é maior do que o valor médio, ou seja, $Var(Y) > E(Y) \Longrightarrow Var(Y) - E(Y) > 0$. Os estimadores da variância e do valor esperado, calculados pelo método da máxima verossimilhança, são dados por s^2 e $\overline{y}$, respectivamente, podendo-se calcular $s^2 - \overline{y}$.

Bohning et al. (1997) propõe uma razão entre o fator adicional da variância do modelo ZIP e a variabilidade da superdispersão. Usando as estimativas dos parâmetros, o cálculo dessa razão é feito da seguinte forma:

$$\frac{\widehat{E(Y)}(\widehat{\mu} - \widehat{E(Y)})}{s^2 - \overline{y}}.$$

Desse modo, essa razão expressa o quanto da variabilidade de um modelo de contagem está sendo atribuída ao excesso de zeros.

3.10 Programas estatísticos

Para o cálculo da taxa de extra-variação foi utilizado o programa Microsoft *Excel 2007*. Os ajustes dos modelos de Poisson e Binomial Negativo usuais foram obtidos através do PROC GENMOD do SAS *for Windows*, versão 9.1.3.

O programa STATA versão 9.0 foi utilizado para ajustar as distribuições inflacionadas de zeros através das rotinas ZIP e ZINB, comparar os modelos usuais e inflacionados pelo teste de Vuong, comparar os modelos inflacionados através dos critérios de AIC e BIC e gerar os gráficos dos modelos ajustados.

Para alguns alimentos selecionados foram aplicados modelos usuais e inflacionados de zeros. As saídas dos programas utilizados trazem muitas informações, as quais foram organizadas em tabelas que apresentam:

- a estimativa dos parâmetros de cada covariável e seu respectivo erro padrão;
- a significância do teste (valor de p) dada pelo valor do score z, pertencente a estatística Wald para testar o efeito considerado (Kodde & Palm, 1986);
- a razão de prevalência (RP) com o intervalo de confiança correspondente a 95% (IC95%). A razão de prevalência é uma razão que pode ser estimada em situações nas quais a prevalência é conhecida (Costa & Matos, 2006), como no presente estudo, no qual encontra-se uma alta prevalência de não consumo de determinados alimentos (Fletcher & Fletcher, 2006a);

- no caso da regressão logística presente na parte inflacionada dos modelos ZIP e ZINB, a razão de chances (*odds ratio* - OR) com o seu intervalo de confiança a 95% (IC95%) (Rumel, 1986; Fletcher & Fletcher, 2006b; Vieira, 2004);

- para o modelo binomial negativo inflacionado de zeros, a estimativa do parâmetro de dispersão θ e seu respectivo intervalo de confiança, sendo θ dado por $\theta = \frac{1}{k}$ (Long & Freese, 2001);

- para os modelos inflacionados de zeros, o teste de Vuong com o valor score z e a probabilidade correspondente;

- e os critérios de AIC e BIC para o modelo descrito.

Para os alimentos selecionados na faixa de 10% a 50% de zeros, foram feitos gráficos que comparam os quatro modelos em questão. O principal objetivo dessas figuras foi observar o ajuste para o não consumo dos alimentos. Esses gráficos mostram a diferença entre as proporções observadas e as probabilidades médias provenientes dos quatro modelos. Para gerar os gráficos, a esses modelos foram associadas apenas as covariáveis significativas(Long & Freese, 2001).

Os programas utilizados SAS e STATA, já apresentam rotinas prontas para o ajuste de modelos usuais e inflacionados de zeros, fornecendo medidas referentes a esses ajustes que facilitam a exploração e a inferência (ver Anexo B).

4 RESULTADOS E DISCUSSÃO

4.1 Descrição da amostra

Dos 73 idosos avaliados, a idade média foi de $71,51$ anos (DP=$6,48$ anos) e o consumo médio de água por dia foi de $1130,48$ ml (DP=$588,99$ ml). A Tabela 1 apresenta a descrição das variáveis qualitativas.

Tabela 1: Análise descritiva das variáveis qualitativas. Avaré, 2009.

Variáveis	Categorias	n (%)
Sexo	Masculino	32 (43,84)
	Feminino	41 (56,16)
	Total	73 (100,00)
Estado Nutricional	Desnutrido	9 (12,68)
	Eutrófico	36 (50,70)
	Obeso	26 (36,62)
	Total	71 (100,00)
Medicamentos	Pressão	34 (58,62)
	Outros	24 (41,38)
	Total	58 (100,00)
Intestino	Normal	55 (75,34)
	Constipado/Diarréia (C/D)	18 (24,66)
	Total	73 (100,00)
Atividade Física	Não	47 (64,38)
	Sim	26 (35,62)
	Total	73 (100,00)

4.2 Modelos inflacionados e porcentagem de zeros

A fim de verificar qual o modelo mais adequado para os dados, foi calculado o valor predito para o ajuste do modelo nulo para cada um dos modelos utilizados: Poisson, Binomial Negativo, Poisson Inflacionado de Zeros (ZIP) e Binomial Negativo Inflacionado de Zeros (ZINB). Foi calculada também a taxa de extra-variação que é atribuída aos zeros, dada por:

$$\frac{\widehat{E(Y)}(\widehat{\mu} - \widehat{E(Y)})}{s^2 - \overline{y}},$$

em que $\widehat{E(Y)} = (1 - \widehat{p})\widehat{\mu}$ com $\widehat{\mu} = \overline{y}$, sendo $\widehat{p}$ a estimativa do parâmetro p dada pela proporção de zeros observados na amostra, $\overline{y}$ a estimativa da média e s^2 a estimativa da variância de uma variável resposta, dada pela frequência mensal de consumo dos alimentos do QFA (Bohning et al., 1997). Os resultados são mostrados nas Tabelas 2, 3 e 4 contendo os 67 ítens do Questionário de Frequência Alimentar, ordenados dos mais para os menos consumidos, separados em faixas de não consumo. Alguns alimentos dentro dessas faixas foram selecionados para aplicar os modelos estudados e verificar a influência com as variáveis explanatórias. As covariáveis associadas foram: sexo, estado nutricional, uso de medicamentos, funcionamento do intestino, prática de atividade física, idade e consumo diário de água em ml.

Tabela 2: Porcentagem de zeros observados (0% a 10%), valores preditos nos modelos usuais, nos modelos inflacionados de zeros e cálculo da taxa de extra-variação. Avaré, 2009.

		Preditos				
Alimentos	Observados n (%)	Poisson	BN	ZIP	ZINB	Taxa(%)
Arroz branco ou integral cozido com óleo e temperos	0 (0,00)	0,00	0,20	0,01	0,20	0,00
Carne de boi (bife, cozida, assada), miúdos, vísceras	1 (0,01)	0,00	0,27	0,05	0,33	0,04
Pão francês, pão de forma, integral, pão doce, torrada	2 (0,03)	0,00	0,32	0,12	0,51	0,06
Frango (cozido, frito, grelhado, assado)	2 (0,03)	0,00	0,26	0,13	0,13	0,07
Feijão (carioca, roxo, preto, verde)	4 (0,05)	0,00	0,53	0,10	0,13	0,12
Tomate	4 (0,06)	0,00	0,24	0,11	0,53	0,08
Alface	6 (0,08)	0,00	0,46	0,20	0,38	0,07

Conforme observa-se na Tabela 2, o Modelo de Poisson usual mostrou-

se mais próximo dos valores observados quando a porcentagem de alimentos não consumidos está entre 0% a 10%.

Tabela 3: Porcentagem de zeros observados (10% a 50%), valores preditos nos modelos usuais, nos modelos inflacionados de zeros e cálculo da taxa de extra-variação. Avaré, 2009.

		Preditos				
Alimentos	Observados n (%)	Poisson	BN	ZIP	ZINB	Taxa(%)
Ovo (cozido, frito)	7 (0,10)	0,00	0,20	0,10	0,11	0,15
Banana	8 (0,11)	0,00	0,24	0,12	0,20	0,13
Outros legumes (abobrinha, berinjela, chuchu, pepino)	8 (0,11)	0,00	0,12	0,13	0,41	0,16
Linguiça	8 (0,13)	0,02	0,26	0,14	0,26	0,13
Sal para tempero de salada	10 (0,14)	0,00	0,14	0,14	0,23	0,14
Laranja	12 (0,17)	0,00	0,27	0,17	0,20	0,12
Sopas (de legumes, canja, creme etc)	11 (0,17)	0,02	0,13	0,19	0,46	0,16
Óleo, azeite ou vinagrete para tempero de salada	14 (0,19)	0,00	0,22	0,21	0,42	0,19
Carne de porco (lombo, bisteca)	13 (0,20)	0,02	0,02	0,20	0,21	0,24
Macarrão com molho com carne, lasanha e nhoque	15 (0,22)	0,09	0,00	0,22	0,22	0,51
Café ou chá com açúcar	17 (0,23)	0,00	0,01	0,23	0,23	0,30
Queijo minas, ricota	18 (0,25)	0,00	0,06	0,25	0,34	0,14
Leite integral	19 (0,26)	0,00	0,09	0,27	0,80	0,19
Açúcar, mel, geléia	19 (0,26)	0,00	0,10	0,26	0,48	0,23
Batata, mandioca, inhame (cozida ou assada), purê	18 (0,26)	0,01	0,09	0,26	0,43	0,25
Embutidos (presunto, mortadela, salsicha)	19 (0,26)	0,01	0,25	0,27	0,31	0,19
Peixe (cozido, frito) e frutos do mar	16 (0,27)	0,05	0,05	0,27	0,27	0,10
Cenoura	20 (0,28)	0,00	0,11	0,30	0,46	0,16
Suco natural	20 (0,29)	0,00	0,13	0,31	0,83	0,06
Brócolis, couve-flor, repolho	20 (0,30)	0,02	0,12	0,30	0,30	0,25
Verduras cozidas	21 (0,31)	0,02	0,18	0,31	0,39	0,14
Salgados fritos (pastel, coxinha, risólis, bolinho)	19 (0,32)	0,10	0,17	0,32	0,35	0,18
Bolo (simples, recheado)	21 (0,32)	0,10	0,16	0,34	0,73	0,21
Batata ou mandioca frita	26 (0,37)	0,05	0,03	0,37	0,53	0,18
Refrigerante comum	29 (0,40)	0,00	0,08	0,40	0,43	0,08
Polenta cozida ou frita	24 (0,40)	0,24	0,13	0,40	0,41	0,28
Farinha de mandioca, farofa, cuscuz, aveia, tapioca	29 (0,41)	0,00	0,08	0,41	0,46	0,09
Biscoito sem recheio (doce, salgado)	32 (0,45)	0,00	0,20	0,45	0,49	0,07
Maçã, pêra	33 (0,46)	0,02	0,09	0,46	0,79	0,13
Suco industrializado	34 (0,47)	0,00	0,15	0,47	0,54	0,06
Pizza, panqueca	28 (0,49)	0,21	0,16	0,49	0,53	0,06
Salgados assados (esfirra, bauruzinho, torta)	30 (0,49)	0,27	0,05	0,49	0,50	0,14

De acordo com a Tabela 3, os valores observados foram próximos dos valores preditos no Modelo de Poisson Inflacionado de Zeros (ZIP) e, na maioria dos casos, se aproximaram também do Modelo Binomial Negativo Inflacionado de Zeros (ZINB). Observou-se também que, a taxa de extra-variação, que mede o quanto da superdispersão foi atribuída as respostas nulas, foi considerável. Desse modo, para

essa porcentagem de respostas nulas os modelos inflacionados apresentaram-se os mais adequados para o conjunto de dados.

Tabela 4: Porcentagem de zeros observados (50% a 100%), valores preditos nos modelos usuais, nos modelos inflacionados de zeros e cálculo da taxa de extra-variação. Avaré, 2009.

		Preditos				
Alimentos	Observados n (%)	Poisson	BN	ZIP	ZINB	Taxa(%)
Mamão	37 (0,52)	0,04	0,02	0,52	-	0,09
Queijo mussarela, prato, parmesão, provolone	40 (0,56)	0,10	0,05	0,56	0,57	0,07
Melão, melância	39 (0,57)	0,32	0,08	0,58	0,88	0,07
Salada de maionese com legumes	39 (0,59)	0,24	0,05	0,59	0,63	0,03
Outras verduras cruas (acelga, rúcula, agrião)	41 (0,59)	0,11	0,06	0,59	0,61	0,02
Sobremesas, doces, tortas, pudins	41 (0,60)	0,14	0,07	0,60	0,64	0,06
Manteiga ou margarina comum	45 (0,62)	0,00	0,26	0,65	0,84	0,07
Chocolate, bombom, brigadeiro	44 (0,63)	0,17	0,07	0,63	0,78	0,04
Carne seca, carne de sol, bacon	47 (0,70)	0,29	0,20	0,72	0,84	0,05
Iogurte com frutas	52 (0,73)	0,28	0,17	0,75	0,96	0,03
Macarrão com molho sem carne	54 (0,75)	0,53	0,05	0,75	0,82	0,08
Feijoada, feijão tropeiro	39 (0,75)	0,43	0,01	0,75	0,77	0,01
Cerveja	54 (0,76)	0,12	0,00	0,76	0,76	0,03
Achocolatado em pó (adicionado ao leite)	55 (0,76)	0,12	0,01	0,76	0,76	0,03
Goiaba	52 (0,78)	0,53	0,04	0,78	0,82	0,04
Leite desnatado	58 (0,81)	0,00	0,04	0,81	0,86	0,03
Hambúrguer, nuggets, almôndegas	56 (0,81)	0,49	0,04	0,81	0,85	0,02
Café ou chá sem açúcar	61 (0,84)	0,00	0,06	0,84	0,84	0,02
Biscoito recheado, waffer, amanteigado	61 (0,85)	0,62	0,04	0,85	0,88	0,02
Abacate	59 (0,87)	0,67	0,17	0,87	0,89	0,01
Sanduíche (cachorro quente, hambúrguer)	59 (0,88)	0,85	0,29	0,88	0,91	0,04
Refrigerante diet/light	65 (0,89)	0,39	0,01	0,89	0,90	0,01
Manteiga ou margarina light	66 (0,90)	0,06	0,01	0,90	0,90	0,01
Lentilha, ervilha seca, grão de bico, soja	58 (0,91)	0,70	0,63	0,95	-	0,01
Iogurte natural	70 (0,96)	0,79	0,10	0,96	0,99	0,00
Condimentos	71 (0,97)	0,88	0,04	0,97	0,98	0,00
Leite semi-desnatado	72 (0,99)	0,44	0,10	0,99	0,99	0,00
Maionese, molho para salada, patê, chantilly	72 (0,99)	0,85	0,01	0,99	0,99	0,00

A Tabela 4 apresenta os alimentos que possuem as maiores taxas de não consumo. Observou-se novamente, que os modelos inflacionados de zeros apresentaram um ajuste mais próximo dos observados. Nota-se, porém, que, a taxa de extra-variação que explicita o quanto da superdispersão foi atribuída ao excesso de zeros, diminui na medida em que se aumenta a porcentagem de respostas nulas. Este fato pode ser explicado pela amplitude dos dados, os quais apresentaram de contagens nulas até cento e vinte vezes de consumo por mês. Essa alta amplitude

também comprometeu o ajuste dos modelos associados às covariáveis, os quais, na maioria, não convergiram para porcentagem acima de 50% de zeros. As observações faltantes na Tabela 4 referem-se à falta de convergência do processo iterativo no ajuste do Modelo Binomial Negativo Inflacionado de Zeros (ZINB).

4.2.1 Ajuste dos modelos para os alimentos com porcentagem de não consumo de 0% a 10%

Para os alimentos com porcentagem de não consumo entre 0% e 10%, o modelo Poisson apresentou o melhor ajuste, de acordo com a Tabela 2. Porém, ao analisar a *deviance*, foi encontrada uma superdispersão, e como alternativa, ajustou-se o modelo binomial negativo.

Para o consumo de frango, carne de boi, tomate e alface não foram encontradas covariáveis significativamente associadas ao consumo desses alimentos. Já para o arroz branco, feijão e pão francês, as associações ocorreram com as covariáveis sexo e atividade física, conforme apresenta a Tabela 5.

Tabela 5: Ajuste de um modelo Binomial Negativo para o consumo de arroz branco, feijão e pão francês. Avaré, 2009.

Alimentos	Variáveis	Categorias	Estimativa	Erro Padrão	valor p	RP (IC95%)
Arroz	Sexo	Masculino	0,341	0,120	0,006	1,406 (1,112-1,777)
		Feminino	-	-	-	1,000
	AIC=498,710		BIC=519,140			
Feijão	Sexo	Masculino	0,581	0,224	0,011	1,788 (1,153-2,775)
		Feminino	-	-	-	1,000
	AIC=514,215		BIC=534,646			
Pão Francês	Atividade	Não	-0,610	0,230	0,008	0,544 (0,347-0,853)
	Física	Sim	-	-	-	1,000
	AIC=504,254		BIC=524,685			

Conforme mostra a Tabela 5, para os alimentos mais consumidos como arroz branco (ou integral cozido com óleo e temperos) e feijão (carioca, roxo, preto, verde), o modelo binomial negativo caracterizou uma associação significativa que

apresentou os idosos do sexo masculino com uma probabilidade de consumo maior do que os idosos do sexo feminino (Arroz: RP=1,406 (IC(95%=1,112-1,777); Feijão: RP=1,788 (IC95%=1,153-2,775)). Para o consumo de pão francês, pão de forma, integral, pão doce e torrada, foi observada associação com a realização de atividade física. Idosos que não praticavam este tipo de atividade mostraram um fator menor de consumo (RP=0,544 (IC95%=0,347-0,853)) em relação aqueles que as praticavam.

Navarro et al. (2001), estudando os fatores associados ao número de internações recorrentes de pacientes idosos, afirma que na comparação entre os resíduos de *deviance* dos modelos de Poisson e binomial negativo, 67,9% das observações mal ajustadas pelo modelo de Poisson são ajustadas pelo modelo binomial negativo. O mesmo pode ser verificado por Bulsara et al. (2004) ao avaliar fatores de riscos associados a hipoglicemia, no qual sugere o ajuste do modelo de regressão binomial negativa como alternativa ao modelo de Poisson, visto que este último, no caso de uma superdispersão, superestima os parâmetros envolvidos.

Tem-se assim que, para uma porcentagem pequena de não consumo, o modelo binomial negativo se mostrou adequado para o ajuste dos dados como alternativa ao modelo de Poisson.

4.2.2 Ajuste dos modelos para alimentos com porcentagem de não consumo de 10% a 50%

A Tabela 3 mostrou que os modelos inflacionados foram os mais adequados quando o não consumo varia entre 10% e 50%, em especial devido à taxa de extra-variação que foi atribuída aos zeros.

Para aplicação dos modelos inflacionados foram escolhidos os alimentos: laranja, leite integral, açúcar, mel e geléia, batata, mandioca, inhame (cozida ou assada), purê, embutidos (presunto, mortadela e salsicha), peixe (cozido ou frito) e frutos do mar, maçã (pêra), bolo (simples, recheado) e macarrão com molho (com carne, lasanha, nhoque), mostrados nas Tabelas de 6 a 14. A partir dos critérios de seleção de AIC e BIC, foi selecionado o Modelo de Poisson Inflacionado de Ze-

ros para o ajuste das frequências de consumo dos alimentos citados, com exceção da frequência de consumo de leite integral, açúcar, mel e geléia, no qual o Modelo Binomial Negativo Inflacionado de Zeros apresentou-se mais adequado.

Tabela 6: Ajuste de um modelo de Poisson Inflacionado de Zeros para o consumo de laranja. Avaré, 2009.

Parte	Variáveis	Categorias	Estimativa	Erro Padrão	valor p	RP (IC95%)
	Idade		-0,019	0,008	0,027	0,982 (0,965-0,998)
	Estado	Desnutrido	0,538	0,166	0,001	1,712 (1,237-2,369)
	Nutricional	Eutrófico	-0,018	0,097	0,855	0,982 (0,813-1,187)
Não		Obeso	-	-	-	1,000
Inflacionada	Medicamentos	Pressão	-0,212	0,093	0,022	0,809 (0,674-0,970)
		Outros	-	-	-	1,000
	Intestino	Normal	-0,342	0,111	0,002	0,710 (0,572-0,882)
		C/D	-	-	-	1,000
	Atividade	Não	-0,321	0,094	0,001	0,725 (0,604-0,872)
	Física	Sim	-	-	-	1,000
Teste Vuong	ZIP vs Poisson		z=2,94	P(z)=0,002		
	AIC=625,343		BIC=661,475			

A Tabela 6 apresenta o ajuste do modelo ZIP com covariáveis significativas associadas ao consumo de laranja (parte não inflacionada de zeros). Verificou-se que, quanto mais avançada a idade, menor o consumo dessa fruta (RP=0,982 (IC95%=0,965-0,998)). Idosos com baixo peso apresentaram um consumo maior em relação aos obesos (RP=1,712 (IC95%=1,237-2,369)). Idosos que referiram tomar medicamentos para pressão, que apresentavam um regular funcionamento do intestino e não praticavam atividade física, apresentaram um consumo menor dessa fruta em relação às demais categorias (RP=0,809 (IC95%=0,674-0,970); RP=0,710 (IC95%=0,572-0,882); RP=0,725 (IC95%=0,604-0,872), respectivamente.).

A Figura 1 mostra os quatro modelos ajustados para frequência do consumo de laranja associado às covariáveis significativas. A probabilidade do modelo de Poisson Inflacionado de Zeros ficou próximo da proporção observada, em especial quando referiu-se ao não consumo dessa fruta.

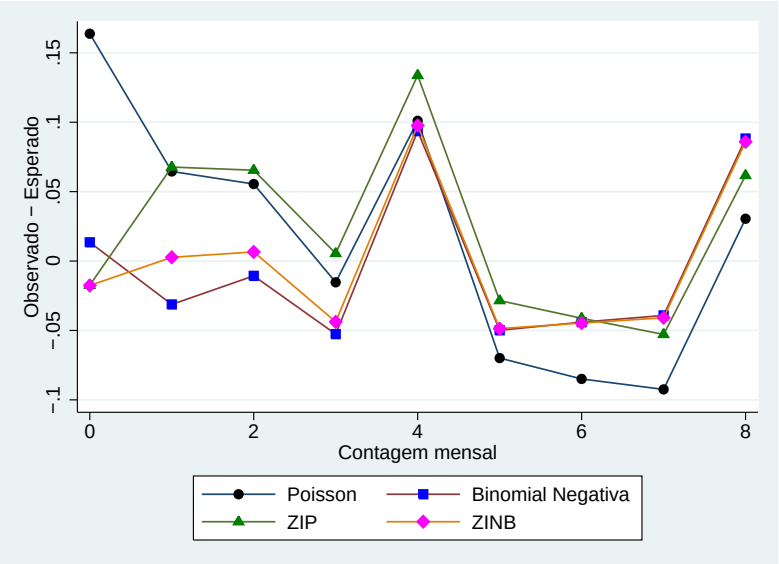

Figura 1: Diferença entre a proporção observada e as probabilidades médias provenientes dos quatro modelos ajustados para o consumo de laranja. Avaré, 2009.

Tabela 7: Ajuste de um modelo Binomial Negativo Inflacionado de Zeros para o consumo de leite integral. Avaré, 2009.

Parte	Variáveis	Estimativa	Erro Padrão	valor p	OR (IC95%)
Inflacionada	Água ml	0,002	0,001	0,043	1,002 (1,001-1,003)
θ		0,585	0,144		0,585 (0,361-0,948)
Teste Vuong	ZINB vs BN	z=3,09	P(z)=0,001		
	AIC=447,989	BIC=486,807			

De acordo com a Tabela 7, o modelo ZINB mostrou uma única associação significativa, dada pela probabilidade de não consumo de leite integral (parte inflacionada) associada ao consumo de água em ml por dia. Idosos que consumiram mais água diariamente, apresentaram uma probabilidade maior de não consumo de leite integral (OR=1,002 (IC95%=1,001-1,003)).

A Figura 2 mostra os quatro modelos ajustados para frequência do consumo de leite integral associado à covariável significativa. As probabilidades médias dos modelos inflacionados apresentaram-se próximas das proporções observadas, re-

velando o excelente ajuste dos modelos inflacionados para caracterizar o consumo de leite integral.

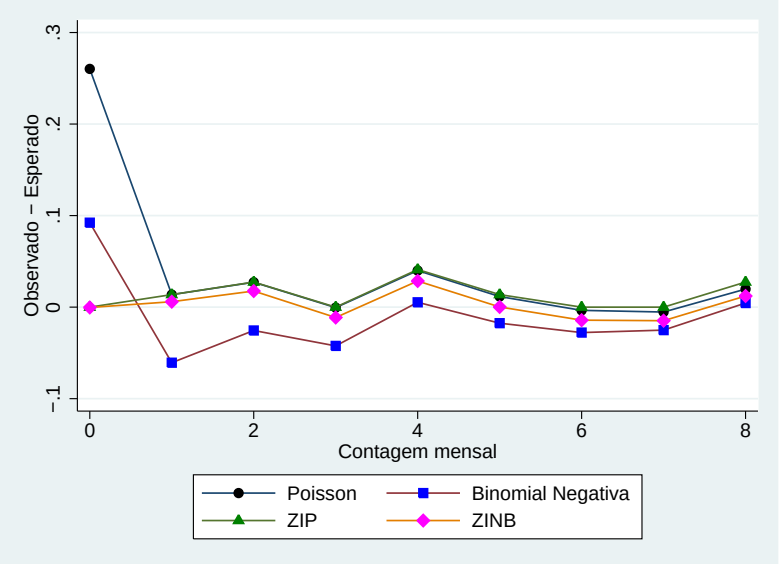

Figura 2: Diferença entre a proporção observada e as probabilidades médias provenientes dos quatro modelos ajustados para o consumo de leite integral. Avaré, 2009.

Tabela 8: Ajuste de um modelo Binomial Negativo Inflacionado de Zeros para o consumo de açúcar, mel e geléia. Avaré, 2009.

Parte	Variáveis	Categorias	Estimativa	Erro Padrão	valor p	RP (IC95%)
	Sexo	Masculino	0,490	0,188	0,009	1,633 (1,130-2,359)
		Feminino				1,000
Não	Medicamentos	Pressão	0,486	0,194	0,012	1,627 (1,111-2,380)
Inflacionada		Outros	-	-	-	1,000
	Intestino	Normal	0,455	0,218	0,037	1,575 (1,027-2,417)
		C/D	-	-	-	1,000
Parte	Variáveis	Categorias	Estimativa	Erro Padrão	valor p	OR (IC95%)
	Estado	Eutrófico	-2,152	1,068	0,044	0,116 (0,014-0,943)
Inflacionada	Nutricional	Obeso	-1,354	1,022	0,185	0,258 (0,035-1,914)
		Desnutrido	-	-	-	1,000
θ			0,310	0,078		0,310 (0,190-0,507)
Teste Vuong	ZINB vs BN		z=3,09	P(z)=0,001		
	AIC=459,034		BIC=483,550			

A Tabela 8 apresenta o ajuste do modelo ZINB com as covariáveis significativas associadas. A probabilidade de maior consumo (parte não inflacionada) foi relacionada à idosos que eram do sexo masculino (RP=1,633 (IC95%=1,130-2,359)), que referiram tomar medicamentos para pressão arterial (RP=1,627 (IC95%=1,111-2,380)) e declararam ter um funcionamento normal do intestino (RP=1,575 (IC95%=1,027-2,417)). A probabilidade de não consumo (parte inflacionada) mostrou-se associada ao estado nutricional. Idosos eutróficos apresentaram uma tendência menor ao não consumo de açúcar, mel e geléia (OR=0,116 (IC95%=0,014-0,943)) em relação aos desnutridos.

A Figura 3 mostra os quatro modelos ajustados para frequência do consumo de açúcar, mel e geléia associado às covariáveis significativas. As probabilidades médias dos modelos inflacionados apresentaram-se próximas das proporções observadas, revelando o excelente ajuste dos modelos inflacionados para caracterizar o consumo de açúcar, mel e geléia.

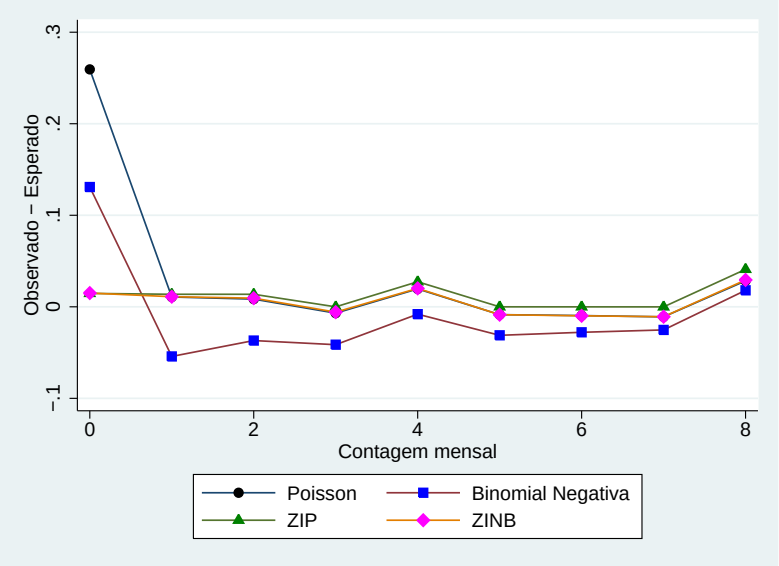

Figura 3: Diferença entre a proporção observada e as probabilidades médias provenientes dos quatro modelos ajustados para o consumo de açúcar, mel e geléia. Avaré, 2009.

Tabela 9: Ajuste de um modelo de Poisson Inflacionado de Zeros para o consumo de batata, mandioca, inhame (cozida ou assada), purê. Avaré, 2009.

Parte	Variáveis	Categorias	Estimativa	Erro Padrão	valor p	RP (IC95%)
Não	Água ml		0,001	0,000	0,000	1,001 (1,000-1,001)
Inflacionada	Estado	Desnutrido	0,638	0,285	0,025	1,893 (1,084-3,307)
	Nutricional	Eutrófico	-0,346	0,184	0,061	0,708 (0,493-1,015)
		Obeso	-	-	-	1,000
Teste Vuong	ZIP vs Poisson		z=3,29	P(z)=0,001		
	AIC=288,065		BIC=323,530			

A Tabela 9 apresenta o ajuste do modelo ZIP com covariáveis para o consumo de batata, mandioca, inhame (cozida ou assada), purê associado significativamente à covariável estado nutricional. Idosos desnutridos apresentaram um consumo maior (parte não inflacionada) em relação aos idosos obesos (RP=1,893 (IC95%=1,084-3,307)).

A Figura 4 mostra os quatro modelos ajustados para frequência do consumo de batata, mandioca, inhame (cozida ou assada), purê associado a covariável significativa.

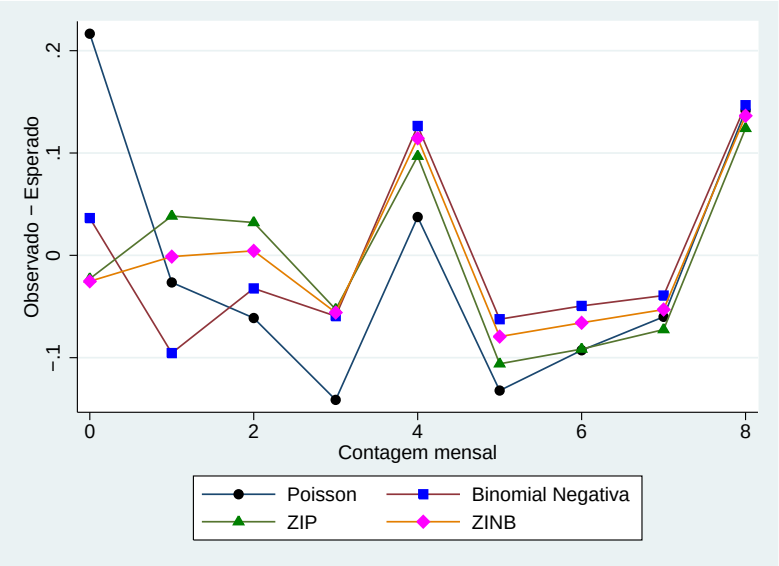

Figura 4: Diferença entre a proporção observada e as probabilidades médias provenientes dos quatro modelos ajustados para o consumo de batata, mandioca, inhame (cozida ou assada), purê. Avaré, 2009.

Nota-se na Figura 4 que as probabilidades médias dos modelos inflacionados apresentam-se próximas às proporções observadas para o não consumo de batata, mandioca, inhame (cozida ou assada) e purê.

Tabela 10: Ajuste de um modelo de Poisson Inflacionado de Zeros para o consumo de embutidos (presunto, mortadela e salsicha). Avaré, 2009.

Parte	Variáveis	Categorias	Estimativa	Erro Padrão	valor p	RP (IC95%)
Não	Estado	Desnutrido	0,715	0,204	0,000	2,045 (1,372-3,048)
Inflacionada	Nutricional	Eutrófico	0,240	0,132	0,068	1,271 (0,982-1,645)
		Obeso	-	-	-	1,000
	Medicamentos	Pressão	-0,450	0,122	0,000	0,638 (0,502-0,811)
		Outros	-	-	-	1,000
Parte	Variáveis	Categorias	Estimativa	Erro Padrão	valor p	OR (IC95%)
Inflacionada	Sexo	Masculino	-1,271	0,772	0,100	0,280 (0,006-1,270)
		Feminino	-	-	-	1,000
	Intestino	Normal	-1,420	0,696	0,041	0,242 (0,062-0,945)
		C/D	-	-	-	1,000
Teste Vuong	ZIP vs Poisson		z=3,49	P(z)=0,000		
	AIC=331,795		BIC=345,972			

A Tabela 10 refere-se ao ajuste do modelo ZIP na associação das covariáveis significativas para o consumo de embutidos (presunto, mortadela e salsicha). Para o não consumo (parte inflacionada) tem-se que idosos que referiram funcionamento normal do intestino apresentaram uma probabilidade menor de não consumo (OR=0,242 (IC95%=0,062-0,945)) em relação aqueles que referiram problemas de constipação e diarréia. Para o consumo (parte não inflacionada), destacaram-se idosos desnutridos os quais apresentaram maior probabilidade (RP=2,045 (IC95%=1,372-3,048)) em relação aos idosos obesos. O uso de medicamentos para pressão em relação as demais categorias também foi um fator significativo (RP=0,638 (IC95%=0,502-0,811)).

A Figura 5 mostra os quatro modelos ajustados para frequência do consumo de embutidos (presunto, mortadela e salsicha) associado às covariáveis significativas. Destacaram-se, novamente, os modelos inflacionados, os quais apresentaram probabilidades médias próximas das proporções observadas para modelagem da frequência de não consumo de embutidos.

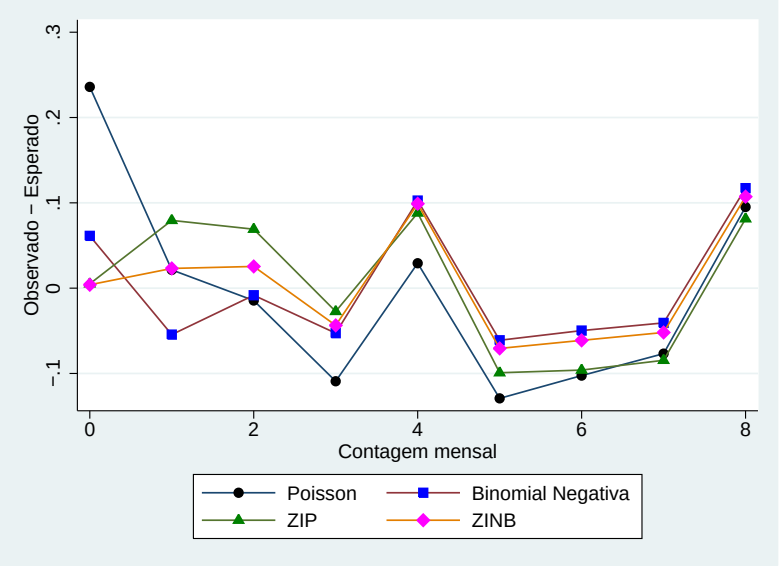

Figura 5: Diferença entre a proporção observada e as probabilidades médias provenientes dos quatro modelos ajustados para o consumo de embutidos (presunto, mortadela e salsicha). Avaré, 2009.

Tabela 11: Ajuste de um modelo de Poisson Inflacionado de Zeros para o consumo de peixe (cozido, frito) e frutos do mar. Avaré, 2009.

Parte	Variáveis	Categorias	Estimativa	Erro Padrão	valor p	RP (IC95%)
	Idade		-0,092	0,017	0,000	0,912 (0,882-0,943)
Não	Estado	Desnutrido	0,178	0,448	0,691	1,195 (0,497-2,877)
Inflacionada	Nutricional	Eutrófico	0,891	0,248	0,000	2,437 (1,499-3,963)
		Obeso	-	-	-	1,000
	Atividade	Não	-0,866	0,206	0,000	0,420 (0,281-0,630)
	Física	Sim	-	-	-	1,000
Teste Vuong	ZIP vs Poisson		z=1,90	P(z)=0,029		
	AIC=234,525		BIC=267,441			

A Tabela 11 apresenta o ajuste do modelo ZIP com a associação das covariáveis para o consumo (parte não inflacionada) de peixe (cozido, frito) e frutos do mar. Idosos eutróficos apresentaram uma maior probabilidade de consumo (RP=2,437 (IC95%=1,499-3,963)) em relação aos obesos. Idosos que não praticavam atividade física tenderam a consumir menos este alimento em relação àqueles que praticavam atividade física (RP=0,420 (IC95%=0,281-0,630)). Além disso, quanto mais a idade era avançada, menor o consumo desse alimento (RP=0,912 (IC95%=0,882-0,943)).

A Figura 6 mostra os quatro modelos ajustados para frequência do consumo de peixe (cozido, frito) e frutos do mar associado as covariáveis significativas. Nota-se que as probabilidades médias apresentaram-se altas quando comparadas com as proporções observadas.

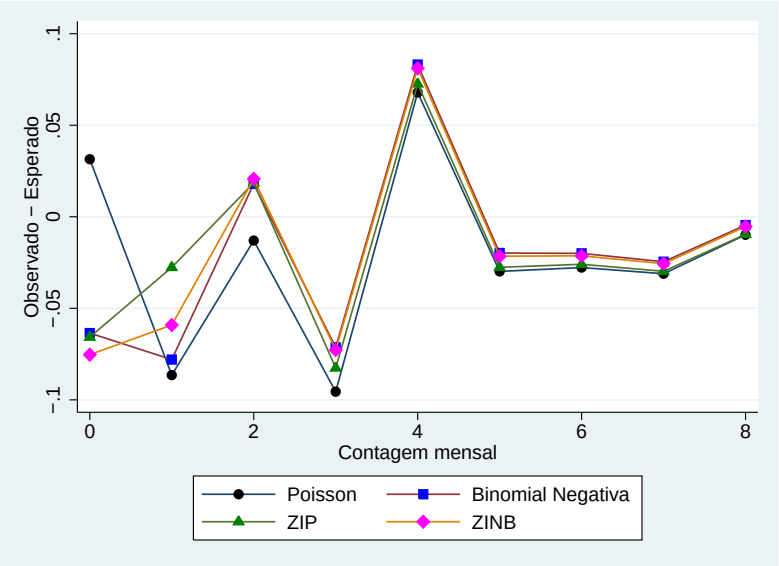

Figura 6: Diferença entre a proporção observada e as probabilidades médias provenientes dos quatro modelos ajustados para o consumo de peixe (cozido, frito) e frutos do mar. Avaré, 2009.

Tabela 12: Ajuste de um modelo de Poisson Inflacionado de Zeros para o consumo de maçã e pêra. Avaré, 2009.

Parte	Variáveis	Categorias	Estimativa	Erro Padrão	valor p	RP (IC95%)
	Água		-0,000	0,000	0,042	1,000 (0,999-1,000)
Não	Sexo	Masculino	0,360	0,163	0,027	1,433 (1,041-1,973)
Inflacionada		Feminino	-	-	-	1,000
Parte	Variáveis	Categorias	Estimativa	Erro Padrão	valor p	OR (IC95%)
Inflacionada	Atividade	Não	1,305	0,652	0,045	3,687 (1,026-13,245)
	Física	Sim	-	-	-	1,000
Teste Vuong	ZIP vs Poisson		z=3,88	P(z)=0,000		
	AIC=331,547		BIC=368,004			

A Tabela 12 apresenta o modelo ZIP com as covariáveis significativas associadas ao consumo de maçã e pêra. Idosos do sexo masculino apresentaram um consumo maior (parte não inflacionada) em relação as mulheres (RP=1,433 (IC95%=1,041-1,973)). O não consumo de maçã (parte inflacionada) apareceu associado a atividade física. Idosos que não praticavam atividade física apresentaram uma probabilidade maior de não consumo em relação aos que praticavam (OR=3,687 (IC95%=1,026-13,245)).

A Figura 7 apresenta os quatro modelos ajustados para frequência do consumo de maçã e pêra associado às covariáveis significativas. As probabilidades médias dos modelos inflacionados apresentaram-se próximas das proporções observadas, revelando o excelente ajuste dos modelos inflacionados para caracterizar o consumo de maçã e pêra.

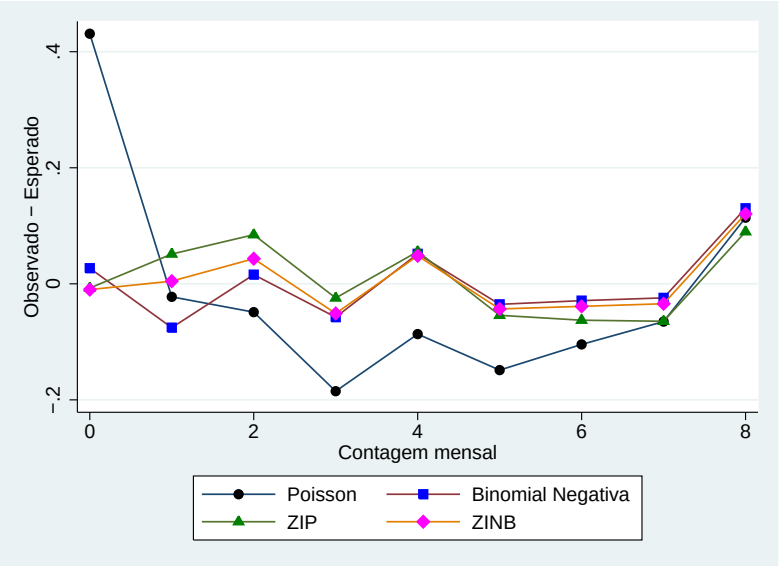

Figura 7: Diferença entre a proporção observada e as probabilidades médias provenientes dos quatro modelos ajustados para o consumo de maçã e pêra. Avaré, 2009.

Tabela 13: Ajuste de um modelo de Poisson Inflacionado de Zeros para o consumo de bolo (simples, recheado). Avaré, 2009.

Parte	Variáveis	Categorias	Estimativa	Erro Padrão	valor p	RP (IC95%)
	Água		-0,001	0,000	0,000	0,998 (0,998-0,999)
	Idade		-0,060	0,017	0,001	0,942 (0,911-0,974)
Não	Estado	Desnutrido	1,337	0,354	0,000	3,808 (1,901-7,628)
Inflacionada	Nutricional	Eutrófico	0,591	0,224	0,008	1,806 (1,165-2,799)
		Obeso	-	-	-	1,000
	Intestino	Normal	-1,050	0,290	0,000	0,350 (0,198-0,618)
		C/D	-	-	-	1,000
Parte	Variáveis	Categorias	Estimativa	Erro Padrão	valor p	OR (IC95%)
	Intestino	Normal	-2,653	1,257	0,035	0,070 (0,000-0,827)
Inflacionada		C/D	-	-	-	1,000
	Atividade	Não	2,256	1,100	0,040	9,538 (1,105-82,296)
	Física	Sim	-	-	-	1,000
Teste Vuong	ZIP vs Poisson		z=2,63	P(z)=0,004		
	AIC=210,855		BIC=245,628			

A Tabela 13 apresenta o ajuste do modelo ZIP com as associações significativas para o consumo ou não da variável bolo (simples, recheado). O consumo de bolo (parte não inflacionada) foi associado a idade, consumo de água, estado nutricional e atividade do intestino. Idosos com idade mais avançada e com maior consumo de água diária apresentaram um consumo menor desse alimento (RP=0,998 (IC95%=0,998-0,999); RP=0,942 (IC95%=0,911-0,974), respectivamente). Idosos desnutridos e eutróficos apresentaram uma maior chance de consumo em relação aos idosos obesos (RP=3,808 (IC95%=1,901-7,628); RP=1,806 (IC95%=1,165-2,799), respectivamente). Aqueles que referiram apresentar um funcionamento normal do intestino apresentaram um fator protetor ao consumo desse alimento (RP=0,350 (IC95%=0,198-0,618)). Para o não consumo (parte inflacionada), idosos que referiram não praticar atividade física apresentaram uma probabilidade maior de não consumo em relação aqueles que praticavam atividade física (OR=9,538 (IC95%=1,105-82,296)). Idosos que referiram ter funcionamento normal do intestino apresentaram uma probabilidade menor de não consumo em relação aqueles que referiram ter pro-

blemas com diarréia e constipação intestinal (OR=0,070 (IC95%=0,000-0,827)).

A Figura 8 apresenta os quatro modelos ajustados para frequência do consumo de bolo (simples, recheado) com as covariáveis significativas.

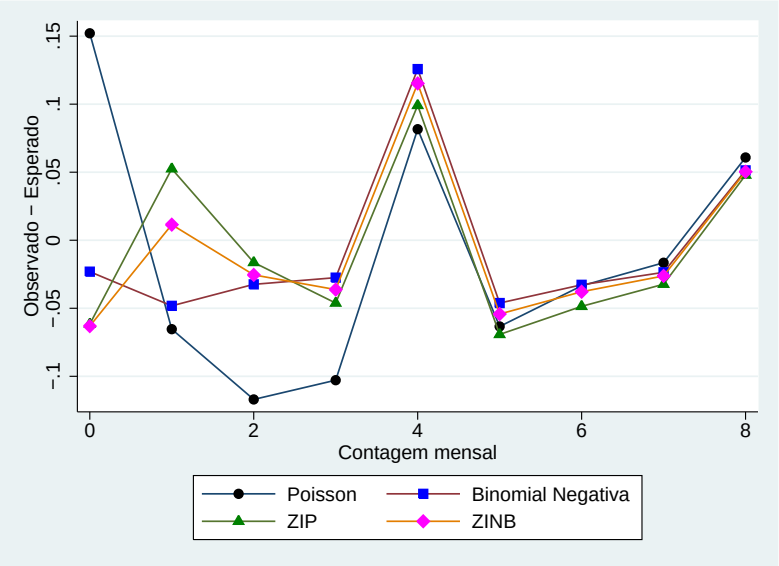

Figura 8: Diferença entre a proporção observada e as probabilidades médias provenientes dos quatro modelos ajustados para o consumo de bolo (simples, recheado). Avaré, 2009.

Tabela 14: Ajuste de um modelo de Poisson Inflacionado de Zeros para o consumo de macarrão com molho com carne, lasanha e nhoque. Avaré, 2009.

Parte	Variáveis	Categorias	Estimativa	Erro Padrão	valor p	RP (IC95%)
Não	Medicamentos	Pressão	0,542	0,190	0,004	1,720 (1,184-2,497)
Inflacionada		Outros	-	-	-	1,000
Parte	Variáveis	Categorias	Estimativa	Erro Padrão	valor p	OR (IC95%)
	Água		-0,007	0,004	0,045	0,992 (0,986-0,999)
Inflacionada	Sexo	Masculino	4,557	2,174	0,036	95,285 (1,345-6751,264)
		Feminino	-	-	-	1,000
Teste Vuong	ZIP vs Poisson		z=2,58	P(z)=0,005		
	AIC=220,085		BIC=255,550			

A Tabela 14 mostra o ajuste do modelo ZIP com as covariáveis significativas associadas ao consumo (parte não inflacionada) ou não (parte inflacionada) de macarrão com molho com carne, lasanha e nhoque. Idosos que referiram tomar medicamentos para a pressão arterial apresentaram um consumo maior em relação aqueles que tomam vários medicamentos (RP=1,720 (IC95%=1,184-2,497)). Para o não consumo, idosos que possuíram um consumo maior de água diariamente apresentaram uma probabilidade menor de não consumo (OR=0,992 (IC95%=0,986-0,999)) e idosos do sexo masculino apresentaram uma probabilidade maior de não consumo em relação a idosos do sexo feminino (OR=95,285 (1,345-6751,264)).

A Figura 9 apresenta os quatro modelos ajustados para frequência do consumo de macarrão com molho com carne, lasanha e nhoque com as covariáveis significativas. Em especial para o não consumo de macarrão, os modelos inflacionados mostraram probabilidades médias mais próximas das proporções observadas.

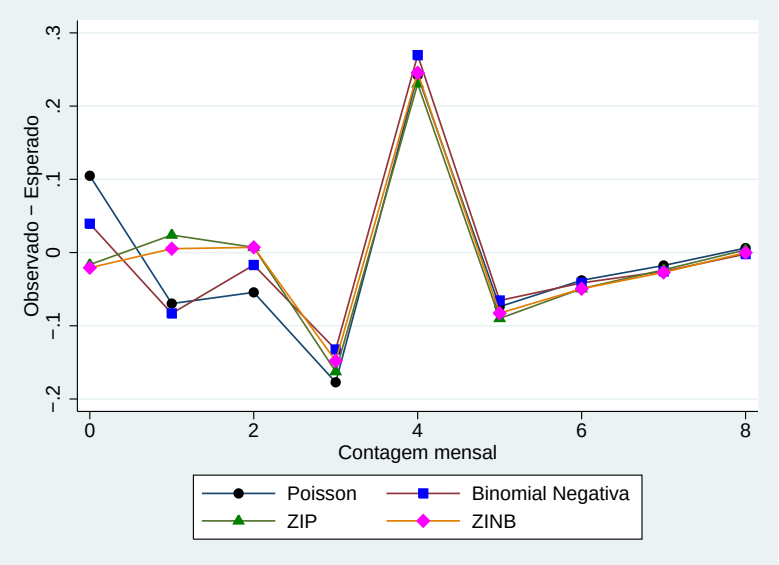

Figura 9: Diferença entre a proporção observada e as probabilidades médias provenientes dos quatro modelos ajustados para o consumo de macarrão com molho com carne, lasanha e nhoque. Avaré, 2009.

Os modelos inflacionados de zeros ajustaram bem as variáveis que possuíam de 10% a 50% de zeros. Horton et al. (2007) apresentam a limitação da distribuição de Poisson no ajuste dos casos com superdispersão quando esta é ocasionada pela presença dos zeros. Slymen et al. (2006) comparam cinco modelos: Poisson, Poisson superdisperso, binomial negativo, ZIP e ZINB para dados de atividade física praticada por mulheres latinas, relatando que o Modelo ZIP apresenta o melhor ajuste para avaliar fatores associados à prática ou não de atividade física. Poston & McKibben (2003) trabalham com modelagem de número de crianças nascidas numa população de mulheres com baixa fertilidade onde há uma predominância de mulheres sem filhos, explicitando através dos modelos inflacionados os fatores associados a esse baixo nascimento de crianças. Cheung (2002) aplica os modelos inflacionados para verificar fatores que influenciam no desenvolvimento motor desde a gestação.

Desse modo, assim como na literatura, através dos modelos inflacionados este trabalho apresenta uma caracterização dos idosos para o consumo ou não dos alimentos referidos.

4.2.3 Ajuste dos modelos para alimentos com porcentagem de não consumo acima de 50%

Para os alimentos com mais de 50% de zeros, os modelos inflacionados não foram ajustados, apresentando em muitos casos problemas na convergência. Kipnis et al. (2009) propõem outros modelos mais adequados para caracterização e consumo envolvendo o Questionário de Frequência Alimentar. Os autores mostram o uso de modelos com erro de medida, nos quais a informação obtida pelo QFA é utilizada como uma variável instrumental para a correção do consumo alimentar feito através de outro instrumento utilizado na área nutricional, como por exemplo, o recordatório 24 horas. Bohning (1998) afirma que, o fato de um modelo ter uma porcentagem de zeros relativamente alta, não implica necessariamente que o melhor ajuste seja dado por um modelo inflacionado de zeros. Deste modo, para o conjunto

de dados em questão estes modelos não apresentaram-se adequados.

Um fator que pode ser considerado como explicativo para o não ajuste dos modelos é a amplitude da frequência dos dados gerados por esse Questionário de Frequência Alimentar, a qual variou da ausência para uma contagem de até cento e vinte vezes de consumo do alimento por mês, dificultando o ajuste dos modelos propostos.

Sheu et al. (2004) ajustam o modelo ZINB para um estudo sobre os fatores associados aos fumantes, no qual a variável resposta é o número de cigarros fumados por dia com variação de consumo diário de zero a 20 cigarros, numa amostra com uma porcentagem de não fumantes de 82,1%. Em Chin & Quddus (2003), o modelo ZIP é utilizado para ajustar dados sobre acidentes envolvendo pedestres, no qual a porcentagem de não acidentes é superior a 80%, com uma variação de zero a 6 acidentes. Bohning et al. (1997) expõem em seu trabalho dados modelados pelo modelo ZIP, no qual o percentual de não acidentes de carro por motorista é de 82,98%, com uma amplitude de zero a 3 acidentes por motorista. Bohning et al. (1999) apresentam um trabalho para comparação de tratamentos distintos para saúde bucal de crianças em seis escolas distintas, nos quais comparou-se o número de problemas odontológicos, registrando uma amplitude de nenhuma até 8 ocorrências, com uma porcentagem considerável de não ocorrência. Martins et al. (2005) aplicam os Modelos ZIP e ZINB para dados referentes a contagem de 31 espécies de pássaros, os quais apresentaram mais de 70% de ausência das espécies nos habitats pesquisados, variando de nenhuma espécie até no máximo 25, em quatro florestas distintas relatadas no estudo.

Desse modo, faz-se necessário estudos mais amplos para se inferir sobre a adequação de tais modelos e as porcentagens de zeros. A priori, observa-se que a distribuição dos dados é um fator importante para a decisão sobre o modelo a ser ajustado.

A amplitude dos dados pode também ter influenciado nas taxas de extra-variação encontradas nesse trabalho, as quais apresentaram valores menores

que a literatura, como no trabalho de Bohning et al. (1997), no qual a taxa de extra-variação é de 77% e, em Bohning et al. (1999), a taxa atinge 90% de explicação da superdispersão devido aos zeros, ambos para o ajuste do modelo ZIP.

5 CONCLUSÕES

Para os dados do Questionário de Frequência Alimentar, os modelos inflacionados de zeros resolveram parcialmente a problemática do excesso de zeros, mostrando-se adequados para o ajuste na porcentagem de 10% a 50% de não consumo. Os modelos usuais de Poisson e Binomial Negativo apresentaram-se coerentes para uma porcentagem com até 10% de zeros. Acima desta porcentagem, outros modelos podem ser testados, como modelos com erros de medidas e variáveis instrumentais. Todavia, outros estudos devem ser feitos para inferir sobre os modelos inflacionados de zeros e as faixas de zeros estabelecidas, através de simulação de dados, levando-se em conta a variação na porcentagem de zeros e a amplitude dos dados.

Este trabalho forneceu uma caracterização integrada do consumo e do não consumo de alimentos nesta população de idosos através do uso de modelos inflacionados de zero. Com isto, pode-se evidenciar que a probabilidade de consumo ou não dos alimentos selecionados está associada ao sexo, ao estado nutricional, ao funcionamento do intestino, ao consumo diário de água e a prática de atividade física. A idade e uso de medicamentos apareceram associadas somente ao consumo dos alimentos selecionados.

ANEXOS

ANEXO A

Modelo do Questionário de Frequência Alimentar

Setor: __ ____ Nº. do quest _____ __ __

QUESTIONÁRIO DE FREQÜÊNCIA ALIMENTAR
ADULTO

PARA TODAS AS PESSOAS COM 20 ANOS OU MAIS

Data da entrevista _____/____/_____ **Hora de início:** _____
Nome do entrevistador: _____
Nº de identificação: _____
Nome: _____ **Sexo** ( )F ( )M
Idade atual: _____ **Data de nascimento:** ____/____/_____

1. Você mudou seus hábitos alimentares recentemente ou está fazendo dieta para emagrecer ou por qualquer outro motivo?

(1) Não
(2) Sim, para perda de peso
(3) Sim, por orientação médica
(4) Sim, para dieta vegetariana ou redução do consumo de carne
(5) Sim, para redução de sal
(6) Sim, para redução de colesterol
(7) Sim, para ganho de peso
Outro motivo: _____

2 . Você está tomando algo para suplementar sua dieta (vitaminas, minerais e outros produtos)?

(1) não (2) sim, regularmente (3) sim, mas não regularmente

3. Se a resposta da pergunta anterior for sim, favor preencher o quadro abaixo:

SUPLEMENTO	MARCA COMERCIAL	DOSE	FREQUÊNCIA

4. As questões seguintes relacionam-se ao seu hábito alimentar usual no PERÍODO DE UM ANO. Para cada quadro responda, por favor, a frequência que melhor descreva QUANTAS VEZES você costuma comer cada item e a respectiva UNIDADE DE TEMPO (se por dia, por semana, por mês ou no ano). Depois responda qual a sua PORÇÃO INDIVIDUAL USUAL em relação à porção média indicada. ESCOLHA SOMENTE UM CÍRCULO PARA CADA COLUNA. Muitos grupos de alimentos incluem exemplos. Eles são sugestões e você pode pode consumir todos os itens indicados. Se você não come ou raramente come um determinado item, preencha o círculo da primeira coluna (N=nunca come). NÃO DEIXE ITENS EM BRANCO.

GRUPO DE ALIMENTOS	Com que frequência você costuma comer?		Qual o tamanho de sua porção em relação à porção média?	
	QUANTAS VEZES VOCÊ COME:	UNIDADE	PORÇÃO MÉDIA (M)	SUA PORÇÃO
Alimentos e preparações	Número de vezes: 1, 2, 3, etc. (N = nunca ou raramente comeu no último ano)	Dpor dia Spor semana Mpor mês Apor ano	Porção média de referência	P = menor que a porção média M = igual à porção média G = maior que a porção média E = bem maior que a porção média

SOPAS E MASSAS	QUANTAS VEZES VOCÊ COME	UNIDADE	PORÇÃO MÉDIA (M)	SUA PORÇÃO
Sopas (de legumes, canja, creme, etc)	N 1 2 3 4 5 6 7 8 9 10 O O O O O O O O O O O	D S M A O O O O	1 concha média (150g)	P M G E O O O O
Salgados fritos (pastel, coxinha, risssóis, bolinho)	N 1 2 3 4 5 6 7 8 9 10 O O O O O O O O O O O	D S M A O O O O	1 unidade grande (80g)	P M G E O O O O
Salgados assados (esfiha, bauruzinho, torta)	N 1 2 3 4 5 6 7 8 9 10 O O O O O O O O O O O	D S M A O O O O	2 unidades ou 2 pedaços médios (140g)	P M G E O O O O
Macarrão com molho sem carne	N 1 2 3 4 5 6 7 8 9 10 O O O O O O O O O O O	D S M A O O O O	1 prato raso (200g)	P M G E O O O O
Macarrão com molho com carne, lasanha, nhoque	N 1 2 3 4 5 6 7 8 9 10 O O O O O O O O O O O	D S M A O O O O	1 escumadeira ou 1 pedaço pequeno (110g)	P M G E O O O O
Pizza, panqueca	N 1 2 3 4 5 6 7 8 9 10 O O O O O O O O O O O	D S M A O O O O	2 fatias pequenas ou 2 unidades (180g)	P M G E O O O O
Polenta cozida ou frita	N 1 2 3 4 5 6 7 8 9 10 O O O O O O O O O O O	D S M A O O O O	2 colheres de sopa ou 2 fatias pequenas (70g)	P M G E O O O O

CARNES E PEIXES	QUANTAS VEZES VOCÊ COME	UNIDADE	PORÇÃO MÉDIA (M)	SUA PORÇÃO
Carne de boi (bife, cozida, assada), miúdos, vísceras	N 1 2 3 4 5 6 7 8 9 10 O O O O O O O O O O O	D S M A O O O O	1 bife médio ou 2 pedaços (100g)	P M G E O O O O
Carne de porco (lombo, bisteca)	N 1 2 3 4 5 6 7 8 9 10 O O O O O O O O O O O	D S M A O O O O	1 fatia média (100g)	P M G E O O O O
Carne seca, carne de sol, bacon	N 1 2 3 4 5 6 7 8 9 10 O O O O O O O O O O O	D S M A O O O O	2 pedaços pequenos (40g)	P M G E O O O O
Linguiça	N 1 2 3 4 5 6 7 8 9 10 O O O O O O O O O O O	D S M A O O O O	1 gomo médio (60g)	P M G E O O O O
Embutidos (presunto, mortadela, salsicha)	N 1 2 3 4 5 6 7 8 9 10 O O O O O O O O O O O	D S M A O O O O	2 fatias médias (30g)	P M G E O O O O
Frango (cozido, frito, grelhado, assado)	N 1 2 3 4 5 6 7 8 9 10 O O O O O O O O O O O	D S M A O O O O	1 pedaço ou 1 filé pequeno (60g)	P M G E O O O O
Hambúrguer, nuggets, almôndega	N 1 2 3 4 5 6 7 8 9 10 O O O O O O O O O O O	D S M A O O O O	1 unidade média (60g)	P M G E O O O O
Peixe (cozido, frito, assado) e frutos do mar	N 1 2 3 4 5 6 7 8 9 10 O O O O O O O O O O O	D S M A O O O O	1 filé pequeno ou 1 posta pequena (100g)	P M G E O O O O

LEITE E DERIVADOS	QUANTAS VEZES VOCÊ COME	UNIDADE	PORÇÃO MÉDIA (M)	SUA PORÇÃO
Leite - tipo: ( ) integral ( ) desnatado ( ) semi-desnatado	N 1 2 3 4 5 6 7 8 9 10 O O O O O O O O O O O	D S M A O O O O	1/2 copo requeijão (125ml)	P M G E O O O O
Iogurte - tipo: ( ) natural ( ) com frutas	N 1 2 3 4 5 6 7 8 9 10 O O O O O O O O O O O	D S M A O O O O	1 unidade pequena (140g)	P M G E O O O O

VERDURAS E LEGUMES	QUANTAS VEZES VOCÊ COME	UNIDADE	PORÇÃO MÉDIA (M)	SUA PORÇÃO
Outras verduras cozidas (acelga, espinafre, escarola, couve)	N 1 2 3 4 5 6 7 8 9 10 O O O O O O O O O O O	D S M A O O O O	1 colher de servir (30g)	P M G E O O O O
Brócolis, couve-flor, repolho	N 1 2 3 4 5 6 7 8 9 10 O O O O O O O O O O O	D S M A O O O O	1 ramo ou 2 colheres de sopa (30g)	P M G E O O O O

MOLHOS E TEMPEROS	QUANTAS VEZES VOCÊ COME	UNIDADE	PORÇÃO MÉDIA (M)	SUA PORÇÃO
Óleo, azeite ou vinagrete para tempero de salada	N 1 2 3 4 5 6 7 8 9 10 O O O O O O O O O O O	D S M A O O O O	1 fio (5ml)	P M G E O O O O
Maionese, molho para salada, paté, chantilly	N 1 2 3 4 5 6 7 8 9 10 O O O O O O O O O O O	D S M A O O O O	1 colher de chá (4g)	P M G E O O O O
Sal para tempero de salada	N 1 2 3 4 5 6 7 8 9 10 O O O O O O O O O O O	D S M A O O O O	1 pitada (0,35g)	P M G E O O O O
Condimentos	N 1 2 3 4 5 6 7 8 9 10 O O O O O O O O O O O	D S M A O O O O	1 pitada (0,35g)	P M G E O O O O

FRUTAS	QUANTAS VEZES VOCÊ COME	UNIDADE	PORÇÃO MÉDIA (M)	SUA PORÇÃO
Laranja, mexerica, abacaxi	N 1 2 3 4 5 6 7 8 9 10 O O O O O O O O O O O	D S M A O O O O	1 unidade média ou 1 fatia grande (180g)	P M G E O O O O
Banana	N 1 2 3 4 5 6 7 8 9 10 O O O O O O O O O O O	D S M A O O O O	1 unidade média (86g)	P M G E O O O O
Maçã, pêra	N 1 2 3 4 5 6 7 8 9 10 O O O O O O O O O O O	D S M A O O O O	1 unidade média (110g)	P M G E O O O O
Melão, melancia	N 1 2 3 4 5 6 7 8 9 10 O O O O O O O O O O O	D S M A O O O O	1 fatia média (150g)	P M G E O O O O
Mamão	N 1 2 3 4 5 6 7 8 9 10 O O O O O O O O O O O	D S M A O O O O	1 fatia média ou **½ unidade média (168g)**	P M G E O O O O
Goiaba	N 1 2 3 4 5 6 7 8 9 10 O O O O O O O O O O O	D S M A O O O O	1 unidade grande **(225g)**	P M G E O O O O
Abacate	N 1 2 3 4 5 6 7 8 9 10 O O O O O O O O O O O	D S M A O O O O	2 colheres de sopa cheias (30g)	P M G E O O O O

BEBIDAS	QUANTAS VEZES VOCÊ COME	UNIDADE	PORÇÃO MÉDIA (M)	SUA PORÇÃO
Suco natural	N 1 2 3 4 5 6 7 8 9 10 O O O O O O O O O O O	D S M A O O O O	1/2 copo americano (80ml)	P M G E O O O O
Suco industrializado	N 1 2 3 4 5 6 7 8 9 10 O O O O O O O O O O O	D S M A O O O O	1 copo de requeijão (240ml)	P M G E O O O O
Café ou chá sem açúcar	N 1 2 3 4 5 6 7 8 9 10	D S M A	2 xícaras de café	P M G E

BEBIDAS	QUANTAS VEZES VOCÊ COME	UNIDADE	PORÇÃO MÉDIA (M)	SUA PORÇÃO
	O O O O O O O O O O	O O O O	(90ml)	P M G E O O O O
Café ou chá com açúcar	N 1 2 3 4 5 6 7 8 9 10 O O O O O O O O O O	D S M A O O O O	2 xícaras de café (90ml)	P M G E O O O O
Refrigerante ( ) comum ( ) diet/light	N 1 2 3 4 5 6 7 8 9 10 O O O O O O O O O O	D S M A O O O O	1 copo de requeijão (240ml)	P M G E O O O O
Cerveja	N 1 2 3 4 5 6 7 8 9 10 O O O O O O O O O O	D S M A O O O O	2 latas (700ml)	P M G E O O O O

PÃES E BISCOITOS	QUANTAS VEZES VOCÊ COME	UNIDADE	PORÇÃO MÉDIA (M)	SUA PORÇÃO
Pão francês, pão de forma, integral, pão doce, torrada	N 1 2 3 4 5 6 7 8 9 10 O O O O O O O O O O	D S M A O O O O	1 unidade ou 2 fatias (50g)	P M G E O O O O
Biscoito sem recheio (doce, salgado)	N 1 2 3 4 5 6 7 8 9 10 O O O O O O O O O O	D S M A O O O O	4 unidades (24g)	P M G E O O O O
Biscoito recheado, wafer, amanteigado	N 1 2 3 4 5 6 7 8 9 10 O O O O O O O O O O	D S M A O O O O	3 unidades (41g)	P M G E O O O O
Bolo (simples, recheado)	N 1 2 3 4 5 6 7 8 9 10 O O O O O O O O O O	D S M A O O O O	1 fatia média (60g)	P M G E O O O O
Manteiga ou margarina passada no pão ( ) comum ( ) light	N 1 2 3 4 5 6 7 8 9 10 O O O O O O O O O O	D S M A O O O O	3 pontas de faca (15g)	P M G E O O O O
Sanduíche (cachorro-quente, hambúrguer)	N 1 2 3 4 5 6 7 8 9 10 O O O O O O O O O O	D S M A O O O O	2 unidades simples (220g)	P M G E O O O O

DOCES E SOBREMESAS	QUANTAS VEZES VOCÊ COME	UNIDADE	PORÇÃO MÉDIA (M)	SUA PORÇÃO
Chocolate, bombom, brigadeiro	N 1 2 3 4 5 6 7 8 9 10 O O O O O O O O O O	D S M A O O O O	1 barra pequena (25g)	P M G E O O O O
Achocolatado em pó (adicionado ao leite)	N 1 2 3 4 5 6 7 8 9 10 O O O O O O O O O O	D S M A O O O O	2 colheres de sopa (25g)	P M G E O O O O
Sobremesas, doces, tortas e pudins	N 1 2 3 4 5 6 7 8 9 10 O O O O O O O O O O	D S M A O O O O	1 pedaço ou 1 fatia média (60g)	P M G E O O O O
Açúcar, mel, geléia	N 1 2 3 4 5 6 7 8 9 10 O O O O O O O O O O	D S M A O O O O	1/2 colher de sopa (6g)	P M G E O O O O

5 . Por favor, liste qualquer outro alimento ou preparação importante que você costuma comer ou beber pelo menos UMA VEZ POR SEMANA que não foram citados aqui (por exemplo: leite-de-coco, outros tipos de carnes, receitas caseiras, creme de leite, leite condensado, gelatina e outros doces etc.).

ALIMENTO	FREQUÊNCIA POR SEMANA	QUANTIDADE CONSUMIDA

6 . Quando você come carne bovina ou suína, você costuma comer a gordura visível?

(1) nunca ou raramente (2) algumas vezes (3) sempre (9) não sabe ☐

7 . Quando você come frango ou peru, você costuma comer a pele?

(1) nunca ou raramente (2) algumas vezes (3) sempre (9) não sabe ☐

Hora do Término da entrevista _____

ANEXO B

Comandos do programa SAS, versão 9.1.3, para o ajuste dos modelos usuais de contagem.

1. Ajuste de um Modelo de Poisson para a variável feijao considerando as covariáveis idade, sexo, estado nutricional, uso de medicamentos, consumo de água em ml, funcionamento do intestino, prática de atividade física, através do $PROC \quad GENMOD$.

   proc genmod;

   title 'feijao Poisson';

   class sexo en medicamentos atividadefisica intestino;

   model feijao= idade sexo en medicamentos aguaml intestino atividade-fisica/dist=poisson type3;

   estimate 'orsexo' sexo 1 -1 /e exp;

   estimate 'oren' en 1 -2 1/e exp;

   estimate 'medicamentos' medicamentos 1 -1/e exp;

   estimate 'intestino' intestino 1 -1/e exp;

   estimate 'atividadefisica' atividadefisica 1 -1/e exp;

   run;

2. Ajuste de um Modelo Binomial para a variável feijao considerando as covariáveis idade, sexo, estado nutricional, uso de medicamentos, consumo de água em ml, funcionamento do intestino, prática de atividade física, através do $PROC \quad GENMOD$.

```
proc genmod;

title 'feijao BN';

class sexo en medicamentos atividadefisica intestino;

model feijao= idade sexo en medicamentos aguaml intestino atividade-
fisica/dist=nb type3;

estimate 'orsexo' sexo 1 -1 /e exp;

estimate 'oren' en 1 -2 1/e exp;

estimate 'medicamentos' medicamentos 1 -1/e exp;

estimate 'intestino' intestino 1 -1/e exp;

estimate 'atividadefisica' atividadefisica 1 -1/e exp;

run;
```

Comandos do programa STATA, versão 9.0, para o ajuste dos Modelos Inflacionados de Zeros.

1. Sintaxe do comando ZIP para o ajuste do Modelo de Poisson Inflacionado de Zeros para a variável resposta bolo considerando as covariáveis sexo, estado nutricional, uso de medicamentos, funcionamento do intestino, prática de atividade física, idade e consumo de água em ml. Na parte inflacionada (inf) inclui-se as variáveis consideradas possíveis geradoras de uma inflação de zeros. $Vuong$ executa o teste de Vuong que compara os modelos ZIP e Poisson.

   zip bolo sexo1 en1 en2 medicame1 intestin1 atividad1 idade aguaml, inf(sexo1 en1 en2 medicame1 intestin1 atividad1 idade aguaml) vuong

2. Sintaxe do comando $ZINB$ para o ajuste do Modelo Binomial Negativo Inflacionado de Zeros (ZINB) para a variável resposta bolo considerando as covariáveis sexo, estado nutricional, uso de medicamentos, funcionamento do intestino, prática de atividade física, idade e consumo de água em ml. Na parte inflacionada (inf) inclui-se as variáveis consideradas possíveis geradoras de uma inflação de zeros. $Vuong$ executa o teste de Vuong que compara os modelos ZINB e Binomial Negativo.

   zinb bolo sexo1 en1 en2 medicame1 intestin1 atividad1 idade aguaml, inf(sexo1 en1 en2 medicame1 intestin1 atividad1 idade aguaml) vuong

3. O comando *estat* a seguir exibe os critérios de AIC e BIC.

   estat ic

4. Para a composição do gráfico foram ajustados os modelos usuais de contagem e os inflacionados.
   Sintaxe do comando *poisson* para o ajuste do Modelo de Poisson para a

variável resposta bolo considerando as covariáveis significativas consumo de água em ml, idade, estado nutricional, funcionamento do intestino e prática de atividade física.

poisson bolo aguaml idade en1 en2 intestin1 atividad1

O comando *prcounts* calcula os valores preditos na regressão de Poisson, para as contagens de zero a sete (8 valores) que serão utilizados depois na geração do gráfico.

prcounts pois, max(8) plot

Sintaxe do comando *nbreg* para o ajuste do Modelo Binomial Negativo para a variável resposta bolo considerando as covariáveis significativas consumo de água em ml, idade, estado nutricional, funcionamento do intestino e prática de atividade física.

nbreg bolo aguaml idade en1 en2 intestin1 atividad1

O comando *prcounts* calcula os valores preditos na regressão Binomial Negativa, para as contagens de zero a sete (8 valores) que serão utilizados depois na geração do gráfico.

prcounts nbreg, max(8) plot

Sintaxe do comando ZIP para o ajuste do Modelo de Poisson Inflacionado de Zeros para a variável resposta bolo considerando as covariáveis significativas associadas: consumo de água em ml, idade, estado nutricional, funcionamento do intestino e prática de atividade física. Na parte inflacionada (inf) inclui-se as variáveis consideradas geradoras de uma inflação de zeros: funcionamento do intestino e prática de atividade física.

zip bolo aguaml idade en1 en2 intestin1, inf(intestin1 atividad1)

O comando *prcounts* calcula os valores preditos na regressão de Poisson Inflacionada de Zeros, para as contagens de zero a sete (8 valores) que serão utilizados depois na geração do gráfico.

prcounts zip, max(8) plot

Sintaxe do comando $ZINB$ para o ajuste do Modelo Binomial Negativo Inflacionado de Zeros para a variável resposta bolo considerando as covariáveis significativas associadas: consumo de água em ml, idade, estado nutricional, funcionamento do intestino e prática de atividade física. Na parte inflacionada (inf) inclui-se as variáveis consideradas geradoras de uma inflação de zeros: funcionamento do intestino e prática de atividade física.

zinb bolo aguaml idade en1 en2 intestin1, inf(intestin1 atividad1)

O comando *prcounts* calcula os valores preditos na regressão Binomial Negativa Inflacionada de Zeros, para as contagens de zero a sete (8 valores) que serão utilizados depois na geração do gráfico.

prcounts zinb, max(8) plot

Cálculo da diferença entre a probabilidade observada para cada contagem e o valor predito para cada um dos quatro modelos.

generate devpois=poisobeq-poispreq

generate devnbreg=poisobeq-nbregpreq

generate devzip=poisobeq-zippreq

generate devzinb=poisobeq-zinbpreq

Legendas

label var devpois "Poisson"

label var devnbreg "Binomial Negativa"

label var devzip "ZIP"

label var devzinb "ZINB"

Geração do gráfico para os quatro modelos.

twoway (connected devpois poisval, msymbol(circle) mcolor(black)) (connected devnbreg poisval, msymbol(square) mcolor(blue)) (connected devzip poisval, msymbol(triangle) mcolor(green)) (connected devzinb poisval, msymbol(diamond) mcolor(magenta)), ytitle(Observado - Esperado) xtitle(Contagem mensal)

REFERÊNCIAS

BICKEL, P. J.; DOKSUM, K. A. **Mathematical statistics: basic ideas and selected topics**. Oakland: Holden-Day, 1977. 553p.

BOHNING, D. Zero-inflated Poisson models and C. A. MAN: A tutorial collection of evidence. **Biometrical Journal**, v.40, n.7, p.833–843, 1998.

BOHNING, D.; DIETZ, E.; SCHLATTMANN, P. Zero-Inflated Count Models and their Applications in Public Health and Social Science. In: ROST, J.; LANGE-HEINE, R. (Ed.). **Applications of latent trait and latent class models in the social sciences**. Munster: Waxmman, 1997. p.333–344.

BOHNING, D.; DIETZ, E.; SCHLATTMANN, P.; MENDONCA, L.; KIRCHNER, U. The Zero-inflated Poisson model and the decayed, missing and filled teeth index in dental epidemiology. **Journal of the Royal Statistical Society. Series A (Statistics in Society)**, v.162, n.2, p.195–209, 1999.

BORGATTO, A. F. Modelos para proporções com superdispersão e excesso de zeros - Um procedimento Bayesiano. Piracicaba, 2004. 90p. Tese (Doutorado) - Escola Superior de Agricultura Luiz de Queiroz, Universidade de São Paulo.

BULSARA, M. K.; HOLMAN, C. D.; DAVIST, E. A.; JONEST, T. W. Evaluating risk factors associated with severe hypoglycaemia in epidemiology studies - what method should we use? **Diabetic Medicine**, v.21, n.8, p.914–919, 2004.

CHEUNG, Y. B. Zero-inflated models for regression analysis of count data: a study of growth and development. **Statistics in Medicine**, v.21, n.10, p.1461–1469, 2002.

CHIN, H. C.; QUDDUS, M. A. Modeling count data with excess zeros: an empirical application to traffic accidents. **Sociological Methods and Research**, v.32, n.1, p.90–116, 2003.

COSTA, M. F. L.; MATOS, D. L. Carta referente ao artigo intitulado auto-avaliação da saúde bucal entre adultos e idosos residentes na região sudeste: resultados do projeto SB-Brasil, 2003. **Cadernos de Saúde Pública**, v.22, n.11, p.1, 2006.

DEMPSTER, A. P.; LAIRD, N. M.; RUBIN, D. B. Maximum likelihood from incomplete data via the EM algoritm. **Journal of the Royal Statistical Society**, v.39, n.1, p.1–38, 1977.

DEMÉTRIO, C. G. B. **Modelos lineares generalizados em experimentação agronômica**. Piracicaba: ESALQ/USP, 2001. 113p.

FLETCHER, R. H.; FLETCHER, S. W. Frequência. In: DUNCAN, M. S. (Ed.). **Epidemiologia clínica: elementos essenciais**. Porto Alegre: Artmed, 2006a. p.82–97.

FLETCHER, R. H.; FLETCHER, S. W. Risco: Um olhar sobre o Passado. In: DUNCAN, M. S. (Ed.). **Epidemiologia clínica: elementos essenciais**. Porto Alegre: Artmed, 2006b. p.98–130.

GREENE, W. H. **Accounting for excess zeros and sample selection in Poisson and negative binomial regression models**. New York: Working paper, Stern School of Business, New York University, Departament of Econometrics, 1994. 36p.

HALL, D. B. Zero-inflated Poisson and binomial regression with random effects: a case study. **Biometrics**, v.56, n.1, p.1030–1039, 2000.

HORTON, N. J.; KIM, E.; SAITZ, R. A cautionary note regarding count models of alcohol consumption in randomized controlled trials. **BMC Medical Research Methodology**, v.7, n.9, p.1–9, 2007.

JONHSON, N. L.; KOTZ, S. **Distributions in statistics: discrete distributions**. Boston: Houghton Miffin, 1969. 646p.

KIPNIS, V.; MIDTHUNE, D.; BUCKMAN, D. W.; DODD, K. W.; GUENTHER, P. M.; KREBS-SMITH, S. M.; SUBAR, A. F.; TOOZE, J. A.; CARROLL, R. J.; FREEDMAN, L. S. Modeling data with excess zeros and measurement error: application to evaluating relationships between episodically consumed foods and health outcomes. **Biometrics**, p.1–8, 2009, doi:10.1111/J.1541-0420.01223-X.

KODDE, D. A.; PALM, F. C. Wald criteria for jointly testing equality and inequality restrictions. **Econometrica**, v.54, n.5, p.1243–1248, 1986.

KUHA, J. AIC e BIC: comparisions of assumptions and performance. **Sociological Methods Research**, v.33, n.2, p.188–229, 2004.

LAMBERT, D. Zero-inflated Poisson regression, with an application to defects in manufacturing. **Technometrics**, v.34, n.1, p.1–14, 1992.

LONG, J. S.; FREESE, J. Predicted probabilities for count models. **The Stata Journal**, v.1, n.1, p.51–57, 2001.

MARTINS, T. G.; WINTLE, B. A.; RHODES, J. R.; KUHNERT, P. M.; FIELD, S. A.; LOW-CHOY, S. J.; TYRE, A. J.; POSSINGHAM, H. P. Zero tolerance ecology: improving ecological inference by modelling the source of zero observations. **Ecology Letters**, v.8, n.11, p.1235–1246, 2005.

MCCULLAGH, P.; NELDER, J. A. **Generalized linear models**. Londres: Chapman e Hall, 1989. 511p.

MEYER, P. L. **Probabilidade: aplicações à estatística**. Rio de Janeiro: Livros Técnicos e Científicos, 1974. 426p.

MINAMI, M.; LENNERT-CODY, C. E.; GAO, W.; ROMÁN-VERDESOTO, M. Modeling shark bycatch: the zero-inflated negative binomial regression model with smoothing. **Fisheries Research**, v.84, n.2, p.210–221, 2007.

NAGAMINE, C. M. L.; CANDOLO, C.; MOURA, M. S. A. Uma aplicação de modelos para dados de contagem inflacionados de zeros na modelagem do número de ovos do mosquito *Aedes Aegypti*. **Revista Brasileira de Biometria**, v.26, n.1, p.99–114, 2008.

NAVARRO, A.; UTZET, F.; CAMINAL, J.; MARTIN, M. La distributión binomial negativa frente a la de Poisson en el análisis de fenómenos recurrentes. **Gaceta Sanitaria**, v.15, n.5, p.447–452, 2001.

NELDER, J. A.; WEDDERBURN, R. W. M. Generalized Linear Models. **Journal of the Royal Statistical Society A**, v.135, n.3, p.370–384, 1972.

POSTON, D. L.; MCKIBBEN, S. L. Using zero-inflated count regression models to estimated the fertility of U. S. women. **Journal of Modern Applied Statistical Methods**, v.2, n.2, p.371–379, 2003.

RIDOUT, M.; DEMÉTRIO, C. G. B.; HINDE, J. Models for count data with many zeros, 1998. In: INTERNATIONAL BIOMETRIC CONFERENCE, 19. South Africa. **Resumos**: Cape Town. p.13.

ROSS, G. J. S.; PREECE, D. A. The negative binomial distribution. **Statistican**, v.34, n.3, p.323–335, 1985.

RUMEL, D. Odds ratio: algumas considerações. **Revista de Saúde Pública**, v.20, n.3, p.253–258, 1986.

SHEU, M.; HU, T.; KEELER, T. E.; ONG, M.; SUNG, H. Y. The effect of a major cigarette price change on smoking behavior in California: a zero-inflated negative binomial model. **Health Economics**, v.8, n.13, p.781–791, 2004.

SLATER, B.; MARCHIONI, S. T.; FISBERG, R. M. Validação de questionários de frequência alimentar - QFA - considerações metodológicas. **Revista Brasileira de Epidemiologia**, v.6, n.3, p.200–2008, 2003.

SLYMEN, D. J.; AYALA, G. A.; ARREDONDO, E. M.; ELDER, J. P. A demonstration of modeling count data with an application to physical activity. **Epidemiologic Perspectives and Innovations**, v.3, n.3, p.1–9, 2006.

VIEIRA, A. M. Modelos para dados de proporções com superdispersão aplicados ao controle biológico. Piracicaba, 1998. 61p. Dissertação (Mestrado) - Escola Superior de Agricultura "Luiz de Queiroz", Universidade de São Paulo.

VIEIRA, S. **Bioestatística: Tópicos avançados**. Rio de Janeiro: Elsevier, 2004. 216p.

VUONG, Q. H. Likelihoood ratio tests for model selection and non-nested hypotheses. **Econometrica**, v.57, n.2, p.307–333, 1989.

WHITAKER, T. B.; DICKENS, J. W. Comparison of the observed distribution of aflatoxin in shelled peanuts to the negative binomial distribution. **Journal of American Oil Chemists' Society**, v.49, n.10, p.590–593, 1972.

APÊNDICES

APÊNDICE A

Valor Esperado e Variância do Modelo de Poisson Inflacionado de Zeros (ZIP)

1. Valor Esperado do Modelo de Poisson Inflacionado de Zeros

$$\begin{aligned} E(Y) &= 0.[p+(1-p)e^{-\mu}] + \sum_{y=1}^{n} y(1-p)\frac{e^{-\mu}\mu^y}{y!} = \\ &= (1-p)\sum_{y=1}^{n} y\frac{e^{-\mu}\mu^y}{y!} = (1-p)\sum_{y=0}^{n} y\frac{e^{-\mu}\mu^y}{y!} \end{aligned}$$

Considerando $Y \sim P(\mu)$,

$$E(Y) = \sum_{y=0}^{n} y\frac{e^{-\mu}\mu^y}{y!} = \mu,$$

Tem-se,

$$E(Y) = (1-p)\mu.$$

2. Variância do Modelo de Poisson Inflacionado de Zeros

$$\begin{aligned} Var(Y) &= E(Y^2) - [E(Y)]^2 = (1-p)(\mu+\mu^2) - (1-p)^2\mu^2 = \\ &= \mu + \mu^2 - p\mu - p\mu^2 - (1-2p+p^2)\mu^2 = \\ &= \mu + \mu^2 - p\mu - p\mu^2 - \mu^2 + 2p\mu^2 - p^2\mu^2 = \\ &= \mu - p\mu + p\mu^2 - p^2\mu^2 = \mu(1-p+p\mu-p^2\mu) = \\ &= \mu(1(1-p)+p\mu(1-p)) = \mu(1-p)(1+p\mu). \end{aligned}$$

Cálculos auxiliares:

$$\begin{aligned} E(Y^2) &= 0^2.[p+(1-p)e^{-\mu}] + \sum_{y=1}^{n} y^2(1-p)\frac{e^{-\mu}\mu^y}{y!} = \\ &= (1-p)\sum_{y=1}^{n} y^2(1-p)\frac{e^{-\mu}\mu^y}{y!} = \\ &= (1-p)\sum_{y=0}^{n} y^2(1-p)\frac{e^{-\mu}\mu^y}{y!}, \end{aligned}$$

Considerando $Y \sim P(\mu)$,

$$\begin{aligned} Var(Y) &= E(Y^2) - [E(Y)]^2 = \\ \mu &= E(Y^2) - \mu^2 \\ E(Y^2) &= \mu + \mu^2, \end{aligned}$$

Tem-se,

$$E(Y^2) = (1-p)(\mu + \mu^2).$$

APÊNDICE B

Estimação dos parâmetros do Modelo de Poisson Inflacionado de Zeros

Segundo caso: O vetor de parâmetros p como função do vetor de parâmetros μ

Um Modelo de Poisson Inflacionado de Zeros apresenta os parâmetros $\boldsymbol{\mu} = (\mu_1, \mu_2, ..., \mu_n)'$ e $\boldsymbol{p} = (p_1, p_2, ..., p_n)'$ provenientes das funções $\ln(\boldsymbol{\mu}) = \boldsymbol{B\beta}$ e $logit(\boldsymbol{p}) = \ln\left(\dfrac{\boldsymbol{p}}{1-\boldsymbol{p}}\right) = \boldsymbol{G\gamma}$, sendo $\boldsymbol{B}$ e $\boldsymbol{G}$ as matrizes de covariáveis.

O vetor de parâmetros $\boldsymbol{\mu}$ pode ser escrito como função do vetor de parâmetros $\boldsymbol{p}$ tomando-se:

$$\boldsymbol{G\gamma} = -\boldsymbol{\tau B\beta}.$$

Assim,

$$\ln\left(\frac{\boldsymbol{p}}{1-\boldsymbol{p}}\right) = -\boldsymbol{\tau B\beta} \implies \boldsymbol{p} = \frac{e^{-\boldsymbol{\tau B\beta}}}{1+e^{-\boldsymbol{\tau B\beta}}},$$

Considerando,

$$\ln(\boldsymbol{\mu}) = \boldsymbol{B\beta} \Longrightarrow \boldsymbol{\mu} = e^{\boldsymbol{B\beta}},$$

Tem-se,

$$\boldsymbol{p} = (1+\boldsymbol{\mu}^{\boldsymbol{\tau}})^{-1}.$$

APÊNDICE C

Estimação dos parâmetros do Modelo de Poisson Inflacionado de Zeros

1. *O vetor de parâmetros p não relacionado com o vetor de parâmetros μ*

   Quando $\boldsymbol{\mu}$ e $\boldsymbol{p}$ não são relacionados, o logaritmo da função de verossimilhança da regressão ZIP com a parametrização padrão é dado por:

$$\begin{aligned}
L(\boldsymbol{\gamma}, \boldsymbol{\beta}; \boldsymbol{y}) &= \ln\{\prod_{y=0}[p + (1-p)e^{-\mu}] \prod_{y\geq 1}^{n}[(1-p)\frac{e^{-\mu}\mu^{y}}{y!}]\} = \\
&= \sum_{y=0} \ln[p + (1-p)e^{-\mu}] + \sum_{y\geq 1}^{n} \ln\left[(1-p)\frac{e^{-\mu}\mu^{y}}{y!}\right] = \\
&= \sum_{y=0} \ln\left[\left(\frac{p}{1-p} + e^{-\mu}\right)(1-p)\right] + \sum_{y\geq 1}^{n}[\ln(1-p) + \\
&+ \ln e^{-\mu} + \ln \mu^{y} - \ln y!] = \\
&= \sum_{y=0} \ln\left[\left(\frac{p}{1-p} + e^{-\mu}\right)(1-p)\right] + \\
&+ \sum_{y\geq 1}^{n}[\ln(1-p) - \mu + y\ln\mu - \ln y!].
\end{aligned}$$

   Considerando-se que:

$$\begin{aligned}
\ln\left(\frac{p}{1-p}\right) = G\gamma &\implies e^{G\gamma} = \frac{p}{1-p}, \\
\ln\left(\frac{p}{1-p}\right) = G\gamma &\implies p = \frac{e^{G\gamma}}{1+e^{G\gamma}}, \\
\ln\left(\frac{p}{1-p}\right) = G\gamma &\implies 1 - p = (1+e^{G\gamma})^{-1}, \\
\ln(\mu) = B\beta &\implies \mu = e^{B\beta},
\end{aligned}$$

   tem-se que:

$$
\begin{aligned}
L(\boldsymbol{\gamma},\boldsymbol{\beta};\boldsymbol{y}) &= \sum_{y=0}\ln(e^{G_i\gamma}+\exp(e^{-B_i\beta}))+\sum_{y=0}\ln(1-p)+\\
&+ \sum_{y\geq 1}^{n}\ln(1-p)+\sum_{y\geq 1}^{n}(yB_i\beta-e^{B_i\beta})-\sum_{y>0}\ln y!=\\
&= \sum_{y=0}\ln(e^{G_i\gamma}+\exp(e^{-B_i\beta}))-\sum_{i=1}^{n}\ln(1+e^{G_i\gamma})+\\
&+ \sum_{y\geq 1}^{n}(yB_i\beta-e^{B_i\beta})-\sum_{y>0}\ln y!=\\
&= \sum_{y=0}\ln(e^{G_i\gamma}+\exp(-e^{B_i\beta}))+\sum_{y>0}(yB_i\beta-e^{B_i\beta})-\\
&- \sum_{i=1}^{n}\ln(1+e^{G_i\gamma})-\sum_{y>0}\ln y!,
\end{aligned}
$$

em que G_i e B_i são colunas das matrizes de covariáveis. A soma de exponenciais no primeiro termo complicam a maximização do $L(\gamma,\beta;y)$. Inserindo a variável indicadora proposta por Lambert (1992), tem-se em detalhes:

$$
\begin{aligned}
L(\boldsymbol{\gamma},\boldsymbol{\beta};\boldsymbol{y},\boldsymbol{z}) &= \ln[\prod_{i=1}^{n}f(y,z,\boldsymbol{\gamma},\boldsymbol{\beta})]=\sum_{i=1}^{n}\ln[f(y,z,\boldsymbol{\gamma},\boldsymbol{\beta})]=\\
&= \sum_{i=1}^{n}\ln[f(y/z,\boldsymbol{\beta})f(z/\boldsymbol{\gamma})]=\\
&= \sum_{i=1}^{n}\ln\left[\frac{e^{-\mu}\mu^{y}}{y!}\right]^{1-z_i}+\\
&+ \sum_{i=1}^{n}\ln[p^{z_i}(1-p)^{1-z_i}]=\\
&= \sum_{i=1}^{n}(1-z_i)[\ln e^{-\mu}+\ln\mu^{y}-\ln y!]+\\
&+ \sum_{i=1}^{n}[\ln p^{z_i}+\ln(1-p)^{1-z_i}]=\\
&= \sum_{i=1}^{n}(1-z_i)[-\mu+y\ln\mu-\ln y!]+\\
&+ \sum_{i=1}^{n}[z_i\ln p+(1-z_i)\ln(1-p)]=\\
&= \sum_{i=1}^{n}(1-z_i)[-e^{B_i\beta}+yB_i\beta-\ln y!]+\\
&+ \sum_{i=1}^{n}[z_i\ln p+\ln(1-p)-z_i\ln(1-p)]=
\end{aligned}
$$

$$
\begin{aligned}
&= \sum_{i=1}^{n}(1-z_i)(yB_i\beta - e^{B_i\beta}) - \sum_{i=1}^{n}(1-z_i)\ln y! + \\
&+ \sum_{i=1}^{n}\left[\ln(1-p) + z_i \ln\left(\frac{p}{1-p}\right)\right] = \\
&= \sum_{i=1}^{n}[z_i G_i\gamma - \ln(1+e^{G_i\gamma})] + \\
&+ \sum_{i=1}^{n}(1-z_i)(yB_i\beta - e^{B_i\beta}) - \sum_{i=1}^{n}(1-z_i)\ln y! = \\
&= L_c(\boldsymbol{\gamma}; \boldsymbol{y}, \boldsymbol{z}) + L_c(\boldsymbol{\beta}; \boldsymbol{y}, \boldsymbol{z}) - \sum_{i=1}^{n}(1-z_i)\ln y!.
\end{aligned}
$$

2. *O vetor de parâmetros p relacionado com o vetor de parâmetros μ*

   Quando $\boldsymbol{\mu}$ e $\boldsymbol{p}$ são relacionados, o logaritmo da função de verossimilhança da regressão ZIP com a parametrização padrão é o mesmo dado anteriormente:

$$
\begin{aligned}
L(\boldsymbol{\gamma}, \boldsymbol{\beta}; \boldsymbol{y}) &= \ln\{\prod_{y=0}[p + (1-p)e^{-\mu}] \prod_{y\geq 1}^{n}[(1-p)\frac{e^{-\mu}\mu^y}{y!}]\} = \\
&= \sum_{y=0}\ln[p + (1-p)e^{-\mu}] + \sum_{y\geq 1}^{n}\ln\left[(1-p)\frac{e^{-\mu}\mu^y}{y!}\right] = \\
&= \sum_{y=0}\ln\left[\left(\frac{p}{1-p} + e^{-\mu}\right)(1-p)\right] + \\
&+ \sum_{y\geq 1}^{n}[\ln(1-p) + \ln e^{-\mu} + \ln\mu^y - \ln y!] = \\
&= \sum_{y=0}\ln\left[\left(\frac{p}{1-p} + e^{-\mu}\right)(1-p)\right] + \\
&+ \sum_{y\geq 1}^{n}[\ln(1-p) - \mu + y\ln\mu - \ln y!] = \\
&= \sum_{y=0}\ln\left(\frac{p}{1-p} + e^{-\mu}\right) + \sum_{y=0}\ln(1-p) + \\
&+ \sum_{y\geq 1}^{n}\ln(1-p) + \sum_{y\geq 1}^{n}(y\ln\mu - \mu) - \sum_{y\geq 1}^{n}\ln y!,
\end{aligned}
$$

   Os vetores de parâmetros serão relacionados da forma:

$$
\ln\left(\frac{p}{1-p}\right) = G\gamma = -\tau B\beta \quad \Longrightarrow \quad \frac{p}{1-p} = e^{G\gamma} = e^{-\tau B\beta},
$$

$$1 - p = (1 + e^{G\gamma})^{-1} = (1 + e^{-\tau B\beta})^{-1},$$

$$\ln \mu = B\beta \quad \Longrightarrow \quad \mu = e^{B\beta},$$

Tem-se:

$$\sum_{y=0} \ln(e^{-\boldsymbol{\tau B}_i\boldsymbol{\beta}} + \exp(-e^{\boldsymbol{B}_i\boldsymbol{\beta}})) + \sum_{y>0}(y\boldsymbol{B}_i\boldsymbol{\beta} - e^{\boldsymbol{B}_i\boldsymbol{\beta}}) - \sum_{i=1}^{n} \ln(1 + e^{-\boldsymbol{\tau B}_i\boldsymbol{\beta}}).$$