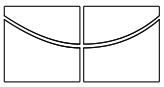

Universidade de Brasília

Instituto de Ciências Exatas

Departamento de Estatística

Dissertação de Mestrado

A Regressão de Touchard e suas aplicações

por

Tallyta Carolyne Martins da Silva

Orientador: Prof. Dr. Raul Yukihiro Matsushita

Brasília

2018

Tallyta Carolyne Martins da Silva

A Regressão de Touchard e suas aplicações

Dissertação apresentada ao Departamento de Estatística do Instituto de Ciências Exatas da Universidade de Brasília como requisito parcial à obtenção do título de Mestre em Estatística.

Banca Examinadora:

- Orientador e presidente da banca examinadora:
  Prof. Dr. Raul Yukihiro Matsushita - EST/UnB
- Examinador interno:
  Prof. Dr. Eduardo Yoshio Nakano - EST/UnB
- Examinador externo:
  Prof. Dr. Luan Carlos de Sena Monteiro Ozelim - ENG/UnB

Agradecimentos

Agradeço primeiramente a Deus, de onde provém todas as coisas, pela saúde e todas as bençãos em minha vida.

Agradeço a minha mãe Eliane e meus irmãos André e Taillyne, pelas orações, pelo apoio e pelo incentivo. Obrigada por compreenderem minha ausência.

A meu querido marido, Eduardo, por estar sempre ao meu lado. Devido a seu companherismo, amizade, apoio, paciência e amor, este trabalho pôde ser concretizado.

Ao meu orientador, Prof. Dr. Raul Matsushita, por seus ensinamentos, pela orientação, paciência, incentivo, sugestões e contribuição na minha formação.

Aos demais professores do mestrado, pelo conhecimento transmitido e por contribuírem na minha formação, em especial ao Antônio Eduardo, Bernardo Borba, Cibele Queiroz, Cira Guevara, Gustavo Gilardoni e Juliana Bettini.

Ao Programa de Pós-Graduação em Estatística - PGEST/UnB - pela oportunidade de realização do meu mestrado.

A todas as pessoas que de alguma forma contribuíram para que esse objetivo fosse alcançado.

Sumário

Lista de Figuras

Lista de Tabelas

Resumo

SILVA, T. C. M. **A Regressão de Touchard e suas aplicações**. 2018. Dissertação (Mestrado) - Departamento de Pós-Graduação em Estatística, Universidade de Brasília, Brasília, 2018.

O presente trabalho apresenta o modelo de regressão de Touchard como uma alternativa para a modelagem de dados de contagens. O modelo basea-se na distribuição de Touchard a qual possui dois parâmetros, λ e δ, ligados a posição e a dispersão, respectivamente, o que possibilita acomodar dados com subdispersão ou superdispersão, e também com excessos de zeros. A estimação dos parâmetros foi feita via máxima verossimilhança. Na análise de ajuste dos dados ao modelo foram utilizados os critérios de seleção de modelo AIC (*Akaike Information Criterion*), o BIC (*Bayesian Information Criterion*), a estatística qui-quadrado e o pseudo-R^2. O trabalho também discute alguns elementos diagnósticos e propõe um procedimento para modelagem e a avaliação da adequabilidade do modelo Touchard. O modelo foi aplicado na distribuição de partos em Goiás no mês de maio de 2017 para avaliar a contribuição dos aspectos socioeconômicos nas contagens de cesáreas. Os dados tem como fonte o Sistema de Informação sobre Nascidos Vivos (SINASC). A aplicação mostra que pelo fato da regressão de Touchard ter dois conjuntos de covariáveis proporciona maior flexibilidade em relação aos demais modelos.

Palavras-chave: Distribuição de Touchard; Regressão de Touchard; técnicas de diagnóstico; estatística qui-quadrado.

Abstract

SILVA, T. C. M. **The Touchard Regression and its applications**. 2018. Dissertação (Mestrado) - Departamento de Pós-Graduação em Estatística, Universidade de Brasília, Brasília, 2018.

The present work presents the Touchard regression model as an alternative for counting data modeling. The model is based on the Touchard distribution which has two parameters, λ and δ, linked to position and dispersion, respectively, which are able to accommodate data with sub-dispersion or over-dispersion, and also with excesses of zeros. The parameters were estimated using maximum likelihood. In the fit analysis of the model we used the AIC (Akaike Information Criterion) model selection criteria, the BIC (Bayesian Information Criterion), the chi-square statistic and pseudo-R^2. The paper also discusses some diagnostic elements and presents a procedure for modeling and evaluating the adequacy of the Touchard model. The model was applied to the distribution of births in Goiás in May 2017 to evaluate the contribution of socioeconomic aspects to cesarean section counts. The data is based on the Information System on Live Births (SINASC). The application shows that because the regression of Touchard has two sets of covariates it provides greater flexibility in relation to the other models.

Keywords: Touchard distribution; Touchard regression; diagnostic techniques; chi-square statistic.

Introdução

Há muitas situações práticas em que se deseja estudar a relação entre uma variável dependente discreta que representa uma contagem e um conjunto de covariáveis. Genericamente, a análise desse tipo de dados nos remete a uma classe que se denomina *modelos de regressão para dados de contagens*.

Um marco importante no desenvolvimento de modelos de regressão para dados de contagens foi o surgimento dos modelos lineares generalizados, dos quais a regressão de Poisson é um caso particular (Nelder e Wedderburn, 1972; McCullagh e Nelder, 1989). Os trabalhos pioneiros nessa área são de Gourieroux et al. (1984) e Hausman et al. (1984).

A principal razão de o modelo Poisson ser amplamente utilizado para a análise de dados de contagens é a sua simplicidade. Apesar disso, ele requer suposições restritivas que limitam sua aderência aos dados reais. A principal delas é a suposição de equidispersão, em que a variância de uma resposta Poisson y deve ser igual ao seu valor esperado. Na prática, porém, é comum haver contagens cujas distribuições apresentam subdispersão (variância menor que a média), superdispersão (variância maior que a média) e excessos de zeros. Contagens como essas são chamadas genericamente de contagens não-Poisson. Dentre alguns exemplos de aplicações relacionadas com contagens não-Poisson encontram-se: o número de visitas ao consultório médico (Zeileis et al., 2008), número de sílabas de palavras em um dicionário e o número de vendas de roupas no trimestre (Shmueli et al., 2005).

Para resolver os problemas com a análise de dados de contangens não-Poisson foram desenvolvidas uma série de generalizações do modelo de Poisson, tais como: Poisson Generalizada (Chandra et al., 2013), Conway-Maxwel-Poisson (Conway e Maxwell, 1962; Shmueli et al., 2005), Binomial Negativa (Bliss e Fisher, 1953), Nova

Generalização Poisson-Lindley (Bhati et al., 2015), e o modelo de Poisson inflado com zeros (Lambert, 1992).

O lado negativo dessas generalizações propostas é que normalmente apresentam forma analítica complexa, algumas não pertencem a família exponencial e, sobretudo, não descrevem simultaneamente a subdispersão, superdispersão e a concentração de zeros.

Esta dissertação tem como objetivo apresentar um ensaio sobre a regressão de Touchard. Ela se baseia em uma nova extensão do modelo de Poisson, denominada distribuição de Touchard (Matsushita et al., 2018). Essa distribuição possui dois parâmetros, o que possibilita acomodar dados com subdispersão ou superdispersão, e também com excessos de zeros.

Este trabalho inicia-se com uma breve revisão sobre os principais modelos de regressão para dados de contagens. O segundo capítulo trata das técnicas de diagnósticos utilizadas para modelos lineares generalizados. O Capítulo seguinte descreve a distribuição de Touchard e algumas das suas propriedades. O Capítulo 4 apresenta o modelo de regressão de Touchard e aborda acerca dos aspectos principais sobre a estimação dos seus parâmetros.

O Capítulo 5 discute alguns elementos de diagnóstico e propõe um procedimento para modelagem e a avaliação da adequabilidade do modelo Touchard. Uma aplicação é desenvolvida no Capítulo 6, considerando os dados do Sistema de Informação de Nascidos Vivos (SINASC) e do Instituo Brasileiro de Geografia e Estatística (IBGE) para avaliar a distribuição de partos em Goiás em maio de 2017. Finalmente, o Capítulo 7 apresenta uma conclusão deste trabalho.

Capítulo 1

Modelos de regressão para dados de contagens

1.1 Introdução

O objetivo deste capítulo é apresentar sinteticamente os principais modelos de regressão para contagens encontrados na literatura. Aqui, o interesse é descrever uma contagem y_i relativa ao i-ésimo elemento da amostra ($i = 1, \ldots, n$) em função de um conjunto de variáveis explicativas $\{X_{i,1}, \ldots, X_{i,k}\}$. Os modelos expostos neste capítulo pertencem à classe de modelos lineares generalizados (MLG) e suas variações (McCullagh e Nelder, 1989).

No MLG (Seção 1.2), em primeiro lugar, a resposta y_i segue uma distribuição da família exponencial, o que contempla uma série de distribuições discretas conhecidas como a Poisson (Seção 1.3) e a binomial negativa (Seção 1.4). Em segundo lugar, o MLG é constituído por uma combinação linear η_i das variáveis explicativas (denominada preditor linear). E, finalmente, define-se uma função de ligação $g(\cdot)$ entre o preditor linear e a contagem esperada, tal que $g(\mu_i) = g(\mathrm{E}(y_i)) = \eta_i$. A construção de um modelo MLG é facilitada se houver uma forma fechada para a média μ_i, sendo ela parametrizada convenientemente (parâmetro canônico). Dessa forma, a Seção 1.3 apresenta o modelo mais utilizado e simples para dados de contagem, o modelo Poisson. Na seção 1.4 é analisado o modelo binomial negativa e na Seção 1.5 a COM-Poisson, extensão da Poisson, que apesar de não possuir forma fechada

para a média, pode ser considerada como um MLG.

Embora nosso propósito não seja modelar dados com excesso de zeros, será considerado o modelo de barreira (Hurdle) e os modelos ZI (zero inflated), como variações do MLG. O modelo de barreira é constituído por dois MLG's, tendo duas funções de ligação. Veremos no Capítulo 2 que a distribuição de Touchard requer duas funções de ligação, uma vez que ela possui dois parâmetros canônicos.

1.2 Modelos lineares generalizados

Nelder e Wedderburn (1972) propuseram os Modelos Lineares Generalizados como uma extensão dos modelos clássicos de regressão. Essa classe de modelos consiste dos seguintes componentes:

- a distribuição da variável resposta y_i pertence à família exponencial o que abarca para as contagens as distribuições Poisson, Binomial Negativa e Touchard (Capítulo 3);

- um preditor linear (função linear das variáveis preditoras),

$$\eta_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + ... + \beta_k X_{ik}; \tag{1.1}$$

- uma função de ligação $g(.)$ inversível, a qual associa a média da variável resposta μ_i ao preditor linear:

$$\eta_i = g(\mu_i) = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + ... + \beta_k X_{ik}; \tag{1.2}$$

de modo que se tenha

$$\mu_i = g^{-1}(\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + ... + \beta_k X_{ik}). \tag{1.3}$$

Considere que y seja da família exponencial biparamétrica, em que θ é o parâmetro canônico associado à média, e ϕ remete à dispersão. Neste caso, sua função

densidade de probabilidade pode ser escrita como

$$f(y;\theta,\phi) = \exp\{\phi^{-1}[y\theta - b(\theta)] + c(y,\phi)\} \tag{1.4}$$

em que b(.) e c(.) são funções conhecidas. Com base na forma (1.4), tem-se que o valor esperado e a variância de Y com distribuição na família exponencial são

$$\mathrm{E}(Y) = \mu = b'(\theta) \qquad e \qquad \mathrm{Var}(Y) = \phi b"(\theta).$$

Pela expressão da variância, tem-se que ϕ é um parâmetro de dispersão do modelo e que ϕ^{-1} corresponde a uma medida de precisão. A primeira derivada de $b(\theta)$ relaciona o parâmetro canônico com a média μ e a variância como função da média μ pode ser expressa como $b"(\theta) = v(\mu)$, em que $v(\mu)$ é chamada de função de variação. Além disso, o parâmetro canônico pode ser dado pela seguinte expressão

$$\theta = \int v^{-1}(\mu)d\mu,$$

sendo que

$$v(\mu) = \frac{d\mu}{d\theta}.$$

A seguir, serão tratados alguns casos particulares.

1.3 Modelo Poisson

O modelo de regressão Poisson é bastante empregado para a análise estatística de contagens. A literatura apresenta várias aplicações, em especial, a área da saúde tem utilizado esse modelo para estimar a razão de prevalência (Conceição et al., 2001) e para identificar o perfil dos óbitos por acidente de trânsito e fatores associados à morte no trânsito (Paixão et al., 2013). Sua função de probabilidades é dada por:

$$f(y;\mu) = \frac{e^{-\mu}\mu^{y}}{y!}, \tag{1.5}$$

para $y = 0, 1, 2, \ldots$ e $\mu > 0$. Uma das suas propriedades é que seu parâmetro corresponde ao seu valor esperado e sua variância $\mathrm{E}(Y) = \mathrm{Var}(Y) = \mu$.

A distribuição Poisson pertence à família (1.4), com os seguintes elementos: $\phi = 1$, $\theta = \ln(\mu)$, $b(\theta) = e^{\theta}$, $c(y, \phi) = -\ln(y!)$, $\mu(\theta) = e^{\theta}$ e $V(\mu) = \mu$. Dos elementos anteriores, considera-se que o parâmetro canônico se relaciona com a média mediante a transformação logarítmica. Dessa forma,

$$\ln(\mu_i) = \mathbf{x_i}'\boldsymbol{\beta}$$

em que

$$E(Y_i|x_i) = \exp\{\mathbf{x_i}'\boldsymbol{\beta}\}$$

sendo $\mathbf{x_i} = (1, x_{i,1}, ..., x_{i,k})'$ o vetor de covariáveis e $\boldsymbol{\beta}$ o vetor (k+1) de parâmetros de regressão. A log-verossimilhança é

$$l(\boldsymbol{\beta}) = \sum_{1}^{n} \left[y_i x_i' \boldsymbol{\beta} - \exp(x_i' \boldsymbol{\beta}) - \ln(y_i!) \right]. \tag{1.6}$$

A modelagem (estimação e diagnósticos) pode ser feita com o pacote "glm"do software R (https://cran.r-project.org/).

Na prática, é comum encontrar dados de contagens que apresentem variância menor/maior que a média, o que pode limitar a aplicação do modelo Poisson. Quando o modelo é utilizado para dados não equidispersos resulta em erros padrões não-confiáveis o que acarreta em inferências incorretas. Por isso, outros modelos devem ser considerados.

1.4 Modelo Binomial Negativo

A distribuição binomial negativa é uma generalização da Poisson e é amplamente utilizada para dados com superdispersão. Sua função de probabilidade é expressa por

$$f(y; \mu, k) = \frac{\Gamma(k + y)}{\Gamma(k) y!} \frac{\mu^y k^k}{(\mu + k)^{k+y}}, \tag{1.7}$$

em que $k > 0$, $\mu > 0$ e $y = 0, 1, \ldots$. A função pode ser escrita como

$$f(y; \mu, k) = \exp\left[\ln\left(\frac{\Gamma(k+y)}{\Gamma(k)y!}\right) + y\ln(\mu) + k\ln(k) - (k+y)\ln(\mu+k)\right],$$

$$= \exp\left[y(\ln(\mu) - \ln(\mu+k)) + k(\ln(k) - \ln(\mu+k)) + \ln\left(\frac{\Gamma(k+y)}{\Gamma(k)y!}\right)\right].$$

Utilizando a notação da família exponencial (1.4): $\phi = 1$, $\theta = \ln\left(\frac{\mu}{\mu+k}\right)$, $b(\theta) = -k\ln(1-e^{\theta})$ e $c(y,\phi) = \ln\left(\frac{\Gamma(k+y)}{\Gamma(k)y!}\right)$.

A esperança e a variância são dadas por

$$E(Y) = \frac{ke^{\theta}}{1-e^{\theta}},$$

$$Var(Y) = \frac{ke^{\theta}}{(1-e^{\theta})^2}. \tag{1.8}$$

O modelo de regressão com resposta binomial negativa pode ser epecificado da seguinte forma:

$$E(Y_i|x_i) = \exp\{\mathbf{x_i}'\boldsymbol{\beta}\},$$

sendo $\mathbf{x_i} = (1, x_{i,1}, ..., x_{i,k})'$ o vetor de covariáveis e $\boldsymbol{\beta}$ o vetor (k+1) de parâmetros de regressão.

A literatura mostra aplicações do modelo binomial negativo na análise da produtividade científica dos antropólogos brasileiros (Alvarado e Oliveira, 2001) e várias aplicações em dados biológicos com superdispersão (Bliss e Fisher, 1953). A modelagem (estimação e diagnósticos) pode ser feita com o pacote "glm.nb"do software R (https://cran.r-project.org/).

1.5 Modelo COM-Poisson

A distribuição de probabilidades Conway-Maxwell-Poisson (COM-Poisson) foi proposta por Conway e Maxwell (1962) e modifica a distribuição Poisson com a

adição de um parâmetro que permite modelar sub/superdispersão. Seja Y uma variável aleatória COM-Poisson sua distribuição de probabilidades é

$$f(y;\lambda,\upsilon) = \frac{\lambda^y}{(y!)^\upsilon Z(\lambda,\upsilon)}, \tag{1.9}$$

para o qual, $y = 0, 1, 2, \ldots$. Nesta distribuição, $\lambda > 0$ corresponde ao parâmetro de forma, $\upsilon \geq 0$ corresponde ao parâmetro de dispersão e $Z(\lambda,\upsilon)$ é uma constante de normalização definida por

$$Z(\lambda,\upsilon) = \sum_{s=0}^{\infty} \frac{\lambda^s}{(s!)^\upsilon}. \tag{1.10}$$

No caso de $\upsilon = 1$, a distribuição COM-Poisson equivale a distribuição Poisson. Além disso, $\upsilon > 1$ caracteriza subdispersão e $\upsilon < 1$ caracteriza superdispersão. A distribuição pertence à família exponencial, pois pode ser escrita como $\exp[y\ln(\lambda) - \upsilon\ln(y!)]Z^{-1}(\lambda,\upsilon)$. Como caso limitante a distribuição COM-Poisson inclui a distribuição Benoulli ($\upsilon = \infty$) e com um caso especial, a distribuição geométrica ($\upsilon = 0$ e $\lambda < 1$).

O modelo de regressão é definido, segundo a notação de MLG's

$$E(Y_i|x_i) = \mathbf{x}_i{}'\boldsymbol{\beta} \tag{1.11}$$

em que $\mathbf{x_i} = (1, x_{i,1}, ..., x_{i,k})'$ é o vetor de covariáveis do i-ésimo indivíduo e $\boldsymbol{\beta}$ o vetor (k+1) de parâmetros. A função de verossimilhança é dada por

$$L(\boldsymbol{\beta},\upsilon;y) = \lambda_i^{\sum_{i=1}^n y_i} \prod_{i=1}^{n} \frac{Z(\lambda_i,\upsilon)^{-1}}{(y_i!)^\upsilon} \tag{1.12}$$

e a função de log-verossimilhança é definida como

$$l(\boldsymbol{\beta},\upsilon;y) = \sum_{i=1}^{n} y_i \ln(\lambda_i) - \upsilon \sum_{i=1}^{n} \ln(y!) - \sum_{i=1}^{n} \ln(Z(\lambda_i,\upsilon)). \tag{1.13}$$

Desse modo as estimativas de máxima verossimilhança são dadas por

$$(\hat{\upsilon}, \hat{\beta}) = argmax_{(\upsilon,\beta)} l(\upsilon, \beta; y). \tag{1.14}$$

O modelo COM-Poisson possui várias aplicações na literatura, como por exemplo, na análise do sistema de compartilhamento de bicicletas (Babu Chatla e Shmueli, 2016) e em dados de melanoma cutâneo (Rodrigues et al., 2009). A modelagem (estimação e diagnósticos) pode ser feita com o pacote "COMPoissonReg"do software R (https://cran.r-project.org/).

1.6 Modelos para excesso de zeros

É comum encontrar dados de contagens com concentração de zeros, isto é, com quantidade de valores nulos maior do que seria esperada pelo modelo ajustado.

O excesso de zeros pode ser explicado por dois processos geradores de dados em uma variável aleatória de contagem. O primeiro denomina-se zeros amostrais, ocorrem segundo um processo gerador de contagens e o segundo são os zeros estruturais, que são relacionados à ausência de determinado atributo da população. Para modelar contagens com excessos de zeros empregam-se, por exemplo, modelos de barreira e de mistura. Os modelos de barreira (Hurdle) modelam apenas os zeros estruturais e as contagens positivas. Já o modelo de mistura (Zero Inflated Model) considera os dois tipos de zeros, além das contagens positivas. A partir desses modelos, conforme mostra Zeileis et al. (2008) constroem-se as regressões direcionadas para os casos de excessos de zeros que são o modelo Hurdle e o Zero Inflated.

1.6.1 Modelos de barreira

Nos modelos de barreira a variável de interesse é dividida em contagens nulas e não nulas, sendo considerados apenas os zeros estruturais. Nesta abordagem um modelo de contagem truncado à esquerda do ponto $y = 1$ é combinado com um modelo censurado à direita no mesmo ponto.

A distribuição de probabilidade é

$$f(y) = \begin{cases} f_z(0), & se \quad y = 0, \\ (1 - f_z(0)) \frac{f_c(Y=y)}{1-f_c(Y=0)}, & se \quad y = 1, 2, ... \end{cases} \tag{1.15}$$

em que f_z é uma função de probabilidade degenerada no ponto $y = 0$, isto é, tem toda massa no ponto 0 e f_c uma função de probabilidades de um variavel Y^* truncada em $y = 1$.

O valor esperado da distribuição é dado por

$$\mathrm{E}(Y) = \frac{(1 - f_z(0))\,\mathrm{E}(Y^*)}{1 - f_c(Y = 0)} \tag{1.16}$$

e a variância

$$\mathrm{Var}(Y) = \frac{1 - f_z(0)}{1 - f_c(Y = 0)} \left[\mathrm{E}(Y^*) \frac{(1 - f_z(0))}{1 - f_c(Y = 0)} \right]. \tag{1.17}$$

Diferentes distribuições podem ser propostas para f_z e f_c, mas uma combinação comum considera Bernoulli para f_z e Poisson para f_c.

Os modelos de regressão Hurdle são construídos incorporando-se covariáveis em f_z e f_c na forma $h(Z\gamma)$ e $g(X\beta)$, sendo que as funções $h(.)$ e $g(.)$ são as funções de ligação escolhidas segundo os modelos f_z e f_c.

A função de log-verossimilhança é dada por

$$l(\theta; y) = \sum_{i=1}^{n} (1 - I)(\ln(f_{zi}(0))) + \sum_{i=1}^{n} I(\ln(1 - f_{zi}(0)) + \ln(f_{ci}(y_i)) - \ln(1 - f_{ci}(0))). \tag{1.18}$$

sendo I a função indicadora que assume o valor 1, se $y > 0$, 0 se $y = 0$, e θ o vetor de parâmetros do modelo.

Uma aplicação do modelo Hurdle é descrita por Zeileis et al. (2008), na qual modela-se o número de consultas médicas em função de algumas covariáveis, como por exemplo, gênero e escolaridade. Este modelo pode ser ajustado a um conjunto de dados com a utilização do pacote "pscl"do R.

1.6.2 Modelo *Zero Inflate*

O modelo *Zero Inflate* (Lambert, 1992), também chamado de modelo de mistura, considera a contribuição de duas funções de probabilidades para a estimação da probabilidade em zero. Esta abordagem une um modelo de contagem sem restrição e um modelo censurado à direita no ponto $y = 1$.

A distribuição de probabilidade é

$$f(y) = \begin{cases} f_z(0) + (1 - f_z(0))f_c(Y = y), & se \quad y = 0, \\ (1 - f_z(0))f_c(Y = y), & se \quad y = 1, 2, ... \end{cases} \tag{1.19}$$

em que f_z é uma função de probabilidades degenerada no ponto $y = 0$ e f_c é uma função de probabilidades para dados de contagens e assim o modelo mistura as duas funções para descrever Y. O valor esperado da distribuição é dado por

$$\mathrm{E}(Y) = (1 - f_z(0))\,\mathrm{E}(Y^*) \tag{1.20}$$

e a variância

$$\mathrm{Var}(Y) = (1 - f_z(0))\,\mathrm{E}(Y^*)[\mathrm{E}(Y^{*2}) - (1 - f_z(0))\,\mathrm{E}(Y^{*2})]. \tag{1.21}$$

Uma combinação comum é considerar a distribuição Bernoulli para f_z e Poisson para f_c.

A função de log-verossimilhança é dada por

$$l(\theta; y) = \sum_{i=1}^{n} I(\ln(1-f_{zi}(0))+\ln(f_{ci})+\sum_{i=1}^{n}(1-I)(\ln(f_{zi}(0))+(1-f_{zi}(0))f_{ci}(0) \tag{1.22}$$

sendo que I é a função indicadora que assume o valor 1 se $y > 0$, 0 se $y = 0$ e θ é o vetor de parâmetros do modelo.

Para modelagem de dados de contagens com excesso de zeros, Lambert (1992) foi pioneiro aplicando esse tipo de modelo na análise do número de defeitos em equipamentos manufaturados. Este modelo pode ser ajustado por meio do pacote "pscl"no software R.

1.7 Considerações

Este capítulo apresentou alguns modelos utilizados para a análise de dados de contagens, como os modelos Poisson, binomial negativa e COM-Poisson que pertencem a família exponencial. Além disso, abordou-se também dois modelos para modelar conjunto de dados com concentração de zeros: modelo de barreira e *Zero Inflate*.

O próximo capítulo pretende explorar os métodos de diagnósticos utilizados principalmente para os modelos lineares generalizados. Serão abordados alguns tipos de resíduos, a estatística generalizada de Pearson e os critérios de seleção de modelos.

Capítulo 7

Conclusão

Este trabalho apresentou o modelo de regressão de Touchard que corresponde a uma ferramenta de análise para dados não-Poisson. Esse modelo assume que a variável resposta Y tem distribuição Touchard com parâmetros λ e δ, ligados a posição e a dispersão, respectivamente, o que possibilita acomodar dados com subdispersão ou superdispersão, e também com excessos de zeros.

Para ilustrar a aplicação do modelo foram utilizados dados do Sistema de Informação sobre Nascidos Vivos (SINASC) e do Instituto Brasileiro de Geografia e Estatística (IBGE). O intuito foi modelar a variável dependente número de partos e avaliar a contribuição das variáveis explicativas (características socioeconômicas e perfil do estabelecimento) na distribuição dos partos em Goiás.

A aplicação mostra que a existência de dois conjuntos de covariáveis, um para a matriz $\mathbf{X}$ e outro para $\mathbf{Z}$, resulta em maior flexibilidade em relação às outras regressões, como a de Poisson. No entanto, o processo de seleção de variáveis se tornou mais demorado, embora as estatísticas gerais AIC (*Akaike Information Criterion*), o BIC (*Bayesian Information Criterion*) e o coeficiente χ^2 se mantiveram úteis na busca de um modelo adequado. Todavia o pseudo-R^2 foi pouco informativo.

Conforme os resultados da aplicação, a regressão de Touchard se mostrou competitiva à frente dos modelos mais utilizados para dados de contagens (Poisson, Binomial Negativa e COM-Poisson).

O trabalho propôs um procedimento que consiste em dividir a base de dados em duas partes, sendo que a primeira parte dos dados é utilizada para simulação

e a segunda para validação. Esse método é analógo a validação cruzada e resolve o problema de encontrar a distribuição da estatística generalizada de Pearson, no entanto, reduz (divide) o tamanho amostral. Em trabalhos futuros pretende-se usar mínimos quadrados ponderados para estimar o modelo e desse modo, obter uma estatística de Pearson mais apropriada.

Há outros elementos da análise de diagnósticos que não foram considerados aqui, como a análise de pontos influentes e os de alavanca (*leverage*). Esses assuntos serão desenvolvidos em oportunidades futuras.

O ponto importante foi demonstrar a viabilidade, a flexibilidade e o potencial da regressão de Touchard para a modelagem de dados de contagens, incrementando nosso *portfolio* de modelos.

Além disso, embora a Touchard pertença à família exponencial, aparentemente ela não se encaixa naturalmente à classe dos modelos lineares generalizados. Ela representa um caso peculiar, com dois conjuntos de covariáveis, com parâmetros λ_i e δ_i estimados linearmente em função dessas covariáveis. Esse assunto poderá ser melhor abordado posteriormente.

Referências Bibliográficas

Akaike, H. (1974). A new look at the statistical model identification. *IEEE transactions on automatic control*, 19(6):716–723.

Akaike, H. (1992). Information theory and an extension of the maximum likelihood principle. In: *Breakthroughs in statistics*, pages 610–624. Springer.

Alvarado, R. U. e Oliveira, M. (2001). A produtividade dos autores na antropologia brasileira. *Data Grama Zero-Revista de Ciência da Informação*, 2(6).

Babu Chatla, S. e Shmueli, G. (2016). An efficient estimation of Conway-Maxwell Poisson regression and additive model with an application to bike sharing.

Bhati, D., Sastry, D., e Qadri, P. M. (2015). A new generalized Poisson-Lindley distribution: Applications and properties. *Austrian Journal of Statistics*, 44(4):35–51.

Bliss, C. I. e Fisher, R. A. (1953). Fitting the negative binomial distribution to biological data. *Biometrics*, 9(2):176–200.

Burnham, K. P. e Anderson, D. R. (2004). Multimodel inference: understanding AIC and BIC in model selection. *Sociological methods & research*, 33(2):261–304.

Casella, G. e Berger, R. L. (2002). *Statistical inference*, volume 2. Duxbury Pacific Grove, CA.

Chandra, N. K., Roy, D., e Ghosh, T. (2013). A generalized Poisson distribution. *Communications in Statistics-Theory and Methods*, 42(15):2786–2797.

Chrysaphinou, O. (1985). On Touchard polynomials. *Discrete mathematics*, 54(2):143–152.

Conceição, G. M. d. S., Saldiva, P. H. N., e Singer, J. d. M. (2001). Modelos MLG e MAG para análise da associação entre poluição atmosférica e marcadores de

morbi-mortalidade: uma introdução baseada em dados da cidade de São Paulo. *Revista Brasileira de Epidemiologia*, 4:206–219.

Consul, P. C. e Jain, G. C. (1973). A generalization of the Poisson distribution. *Technometrics*, 15(4):791–799.

Conway, R. W. e Maxwell, W. L. (1962). A queuing model with state dependent service rates. *Journal of Industrial Engineering*, 12(2):132–136.

Cordeiro, G. M. e Demétrio, C. G. (2008). Modelos lineares generalizados e extensões. *São Paulo*.

Cox, D. R. e Snell, E. J. (1968). A general definition of residuals. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 248–275.

Demétrio, C. e Cordeiro, G. (2007). Modelos lineares generalizados. *Simpósio de Estatística Aplicada à Experimentação Agronômica*, 12.

Dobson, A. J. (2002). *An introduction to generalized linear models*. CRC press.

Gourieroux, C., Monfort, A., e Trognon, A. (1984). Pseudo maximum likelihood methods: applications to Poisson models. *Econometrica*, 52(3):701–720.

Hausman, J. A., Hall, B. H., e Griliches, Z. (1984). Econometric models for count data with an application to the patents-R&D relationship.

IMB (2013). Estado de Goiás: caracterísitcas socioeconômicas e tendendências recentes. Technical report, Instituto Mauro Borges, Secretaria de Estado de Gestão e Planejamento, Governo de Goiás.

Lambert, D. (1992). Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics*, 34(1):1–14.

Mardia, K. V., Kent, J. T., e Bibby, J. M. (1980). Multivariate analysis (probability and mathematical statistics).

Matsushita, R., Pianto, D., De Andrade, B. B., Cançado, A., e Da Silva, S. (2018). The Touchard distribution. *Communications in Statistics-Theory and Methods*, pages 1–11.

McCullagh, P. e Nelder, J. A. (1989). *Generalized linear models*, volume 37. CRC press.

McFadden, D. et al. (1977). *Quantitative methods for analyzing travel behavior of individuals: some recent developments*. Institute of Transportation Studies, University of California.

Nelder, J. e Wedderburn, R. (1972). Generalized linear models. *Journal of the Royal Statistical Society A*, 1972:370–384.

Oliveira, S. B. d. (2017). A distribuição Touchard e suas aplicações.

Paixão, L. M. M. M., da Silva Costa, D. A., Caiaffa, W. T., Gontijo, E. D., e de Lima Friche, A. A. (2013). Aplicação do modelo de regressão de Poisson: Identificação do perfil dos óbitos por acidente de trânsito (AT) e fatores associados à morte no trânsito em belo horizonte (MG). *Matemática e Estatística em Foco*, 1(2).

Pearson, K. (1900). On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 50(302):157–175.

RDevelopment, C. Team. 2008. r: A language and environment for statistical computing. vienna: R foundation for statistical computing.

Rodrigues, J., de Castro, M., Cancho, V. G., e Balakrishnan, N. (2009). COM-Poisson cure rate survival models and an application to a cutaneous melanoma data. *Journal of Statistical Planning and Inference*, 139(10):3605–3611.

Sankaran, M. (1970). 275. note: The discrete Poisson-Lindley distribution. *Biometrics*, pages 145–149.

Schmidt, C. (2003). Modelo de regressão de Poisson aplicado à área da saúde.

Schwarz, G. et al. (1978). Estimating the dimension of a model. *The annals of statistics*, 6(2):461–464.

Shmueli, G., Minka, T. P., Kadane, J. B., Borle, S., e Boatwright, P. (2005). A useful distribution for fitting discrete data: revival of the Conway–Maxwell–Poisson distribution. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 54(1):127–142.

Smyth, G. K. (2003). Pearson's goodness of fit statistic as a score test statistic. *Lecture Notes-Monograph Series*, pages 115–126.

UNICEF (2008). Iniciativa hospital da criança: revista, atualizada e ampliada para o cuidado integrado. Technical report, Fundo das Nações Unidas para a Infância, Organização das Nações Unidas (ONU).

Velasque, L. d. S. (2011). Aplicação dos modelos de Cox e Poisson para obter medidas de efeito em um estudo de coorte.

WHO (2018). Who recommendations: intrapartum care for a positive childbirth experience. Technical report, World Health Organization.

Zeileis, A., Kleiber, C., e Jackman, S. (2008). Regression models for count data in R. *Journal of statistical software*, 27(8):1–25.