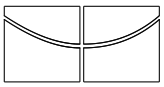

Universidade de Brasília

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 2

Técnicas de diagnósticos em Modelos Lineares Generalizados

2.1 Introdução

Uma etapa importante no ajuste de modelos é a análise de diagnóstico. Ela tem como objetivo avaliar a qualidade do modelo ajustado aos dados, verificar possíveis afastamento das suposições feitas para o modelo e verificar a presença de observações com alguma influência desproporcional nos resultados do ajuste.

Para os Modelos Lineares Generalizados (MLGs), a análise de diagnóstico verifica os seguintes elementos:

- Adequação da distribuição proposta;
- Adequação da parte sistemática (preditor linear) do modelo;
- Adequação da função de ligação;
- Identificação e avaliação de observações mal ajustadas;
- Identificação de observações influentes e pontos de alavanca e a análise do impacto dessas observações no ajuste do modelo.

Alguns métodos para análise de resíduos e diagnóstico em MLGs correspondem a extensões ou adaptações dos procedimentos utilizados no modelo clássico de regres-

são. Contudo, é necessária cautela no uso dessas ferramentas de análise em MLGs visto que alguns resultados dependem fortemente da distribuição proposta.

Na análise de diagnóstico, os resíduos são fundamentais para identificar obervações discrepantes. Conforme Cox e Snell (1968), os resíduos são medidas de afastamento entre a observação y_i e o seu valor ajustado $\hat{\mu}_i$ no modelo, sendo dado por:

$$R_i = h_i(y_i, \hat{\mu}_i) \tag{2.1}$$

sendo h_i uma função adequada, usualmente escolhida para estabilizar a variância ou induzir simetria na distribuição amostral de R_i, a fim de garantir comparabilidade dos resíduos e possibilitar a detecção de resíduos discrepantes.

A matriz de projeção H, nos modelos lineares generalizados é dada por (McCullagh e Nelder, 1989)

$$H = W^{\frac{1}{2}} X (X'WX)^{-1} X' W^{\frac{1}{2}}, \tag{2.2}$$

em que a matriz W é a matriz diagonal, com os elementos da diagonal principal dados por

$$w_{ii} = \frac{1}{v(\mu_i)} \left(\frac{\partial \mu_i}{\partial \eta_i} \right)^2. \tag{2.3}$$

A matriz H tem um papel importante na análise de resíduos nos MLGs e depende das variáveis explicativas, da função de ligação e da função de variância. Possui traço tr(H)=p e $0 \leq h_{ii} \leq 1$.

2.2 Resíduos em Modelos Lineares Generalizados

2.2.1 Resíduo de Pearson

O resíduo de Pearson é o tipo de resíduo mais utilizado nos MLGs, e é expresso por:

$$r_i^p = \frac{y_i - \hat{\mu}_i}{\hat{v}(\mu_i)^{1/2}} \tag{2.4}$$

em que $\hat{\mu}_i$ e $\hat{v}(\mu_i)$ são respectivamente, a média ajustada e a função de variação ajustada de Y_i. Segundo Demétrio e Cordeiro (2007), o resíduo de Pearson corresponde a uma componente da estatística de Pearson generalizada. A desvantagem desse re-

síduo é o fato de ter distribuição fortemente assimétrica para modelos não-normais.

2.2.2 Resíduo de Pearson padronizado

O resíduo de Pearson padronizado é dado por:

$$r_i^p = \frac{y_i - \hat{\mu}_i}{\sqrt{v(\hat{\mu}_i)(1 - h_{ii})}}, \tag{2.5}$$

em que h_{ii} é o i-ésimo elemento da diagonal da matriz de projeção ortogonal dada em (2.2).

2.2.3 Resíduo Componente do Desvio

Para selecionar um modelo dentre um grupo de modelos de regressão é necessário utilizar algumas medidas de discrepância para mensurar o ajuste do modelo. No caso dos modelos lineares generalizados é comum utilizar a função desvio ou deviance (McCullagh e Nelder, 1989), que equivale a distância, para cada observação, entre o logaritmo da função de verossimilhança do modelo saturado (n parâmetros) e do modelo sob investigação (com p parâmetros), avaliado na estimativa de máxima verossimilhança. A função desvio ou deviance proposta por Nelder e Wedderburn (1972) é expressa por:

$$D = 2\left[\hat{l}_{sat} - \hat{l}_{cor}\right], \tag{2.6}$$

em que $\hat{l}_{sat}$ e $\hat{l}_{cor}$ são os máximos da função de verossimilhança para os modelos saturado e corrente (sob investigação), respectivamente.

O resíduo componente do desvio é definido como a raiz quadrada da diferença entre as log-verossimilhanças sob o modelo saturado e o modelo corrente para cada uma das observações, com sinal dado pelo sinal de $y_i - \hat{\mu}_i$, isto é,

$$r_i^D = \text{sinal}(y_i - \hat{\mu}_i)\sqrt{2(\hat{l}_{sat} - \hat{l}_{cor})}. \tag{2.7}$$

Esse resíduo equivale a distância da observação y_i ao seu valor ajustado $\hat{\mu}_i$, medida na escala do logaritmo da função de verossimilhança. De acordo com Demétrio

e Cordeiro (2007) os benefícios de (2.7) são: não necessitar da função normalizadora, implementação simples após o ajuste do MLG e ser definido para todas as observações.

2.2.4 Resíduo Componente do Desvio Padronizado

O resíduo componente do desvio padronizado é uma padronização do resíduo dado em (2.7), assim é definido como

$$r_i^{D'} = \frac{r_i^D}{\sqrt{1-h_{ii}}} \tag{2.8}$$

sendo que h_{ii} representa o i-ésimo elemento da diagonal da matriz de projeção ortogonal dada em (2.2).

2.3 Estatística de Pearson generalizada

Uma medida da discrepância do ajuste de um modelo a um conjunto de dados é a estatística de Pearson generalizada que é definida como a soma dos quadrados dos resíduos de Pearson (2.4) e é expressa por

$$Q = \sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i)^2}{v(\hat{\mu}_i)}, \tag{2.9}$$

sendo $v(\hat{\mu}_i)$ a função de variação estimada sob o modelo ajustado aos dados. Segundo Cordeiro e Demétrio (2008) para dados oriundos das distribuições binomial e Poisson a estatística de Pearson é expressa por

$$Q = \sum_{i=1}^{n} \frac{(o_i - e_i)^2}{e_i}, \tag{2.10}$$

em que o_i corresponde a frequência observada e e_i a frequência esperada. Esta medida pretende testar se as diferenças entre os valores reais e o valores teoricamente esperados são estatisticamente significativas. Uma concordância entre os valores observados e esperados resulta em um pequeno valor de Q e a um grande p-valor.

Por outro lado, a discrepância entre estes valores levará a um valor grande de Q e um p-valor pequeno. Por meio de resultados assintóticos a estatística de Pearson tem distribuição qui-quadrado com $n - p$ graus de liberdade.

2.4 Critérios para seleção de modelos baseados na razão de verossimilhança

A seleção do modelo mais apropriado é uma etapa importante na modelagem de dados. O princípio da parcimônia considera que o modelo mais indicado é aquele que possui o menor número de parâmetros a serem estimados e que explique de forma satisfatória o comportamento da variável resposta.

Os critérios mais utilizados são os baseados na função de verossimilhança: Teste de Razão de Verossimilhança (TRV), Critério de Informação de Akaike (AIC) e o Critério Bayesiano de Schwarz (BIC), sendo que o TRV é apropriado para testar modelos encaixados. Como o foco do trabalho é comparar modelos não aninhados, será utilizado os critérios AIC e BIC, que permitem a comparação da qualidade de ajuste de um modelo estatístico estimado e o seu grau de complexidade. O primeiro critério de informação (AIC) foi proposto Akaike (1974), basea-se no conceito de entropia e é definido por

$$AIC = -2\hat{l}_{cor} + 2p, \tag{2.11}$$

em que p é o número de parâmetros a serem estimados no modelo. Esse critério mensura as informações perdidas, quando um determinado modelo é utilizado para descrever algum fenômeno. O termo adicionado ao máximo da função de verossimilhança é denominado de fator de função de penalidade, visto que tem como objetivo ponderar o viés proveniente da comparação de modelos com diferentes números de parâmetros. Desse modo, o modelo com melhor ajuste aos dados entre os modelos candidatos é aquele que minimizar o AIC (Burnham e Anderson, 2004).

O Critério de Informação Bayesiano (BIC), proposto por Schwarz et al. (1978) é expresso por

$$BIC = -2\hat{l}_{cor} + p\ln(n), \tag{2.12}$$

sendo que p é o número de parâmetros a serem estimados e n é o número de observações da amostra. O critério BIC tem interpretação similar ao AIC, ou seja, quanto menor melhor.

2.5 Pseudo-R^2

O pseudo-R^2 corresponde a uma adapatação do coeficiente de determinação utilizado para medir a qualidade de ajuste de um modelo de regressão linear. Segundo McFadden et al. (1977) o pseudo-R^2 é definido por

$$Pseudo - R^2 = 1 - \frac{l_1}{l_0}, \tag{2.13}$$

em que l_0 representa a log-verossimilhança do modelo nulo, isto é, sem nenhuma variável explicativa e l_1 a log-verossimilhança do modelo ajustado. O seu valor está entre 0 e 1 e quanto mais próximo de 1, melhor o ajuste do modelo. Desse modo, um maior Pseudo-R^2 pode ser utilizado como critério para escolha de um modelo em detrimento de outro.

Capítulo 3

Distribuição de Touchard

3.1 Introdução

Nos últimos anos, uma série de generalizações da distribuição Poisson foram desenvolvidas com o objetivo de se ajustarem as contagens não-Poisson. No entanto, muitas dessas distribuições propostas apresentam forma analítica complexa, algumas não pertencem a família exponencial e, sobretudo, não descrevem simultaneamente a subdispersão, a superdispersão e a concentração de zeros.

Este capítulo descreverá a distribuição de Touchard desenvolvida por Matsushita et al. (2018) que corresponde a uma extensão da distribuição de Poisson. Essa distribuição apresenta a inclusão de um parâmetro, o que proporciona uma maior flexibilidade na modelagem de dados discretos. Uma das vantagens da distribuição de Touchard em relação às outras extensões da Poisson é modelar conjuntamente casos de sub/superdispersão e excesso de zeros.

3.2 Distribuição de Touchard

Uma variável $Y \sim Touchard(\lambda, \delta)$ se sua distribuição de probabilidade é dada por

$$p_y = \mathrm{P}[Y = y] = \frac{\lambda^y (y+1)^\delta}{y! \tau(\lambda, \delta)}, \tag{3.1}$$

para $y = 0, 1, 2, ...$; $\lambda > 0$ e $\delta \in \mathbb{R}$, e

$$\tau(\lambda, \delta) = \sum_{h=1}^{\infty} \frac{\lambda^h (h+1)^\delta}{h!} \tag{3.2}$$

normaliza (3.1) e é conhecido como o polinômio de Touchard (Chrysaphinou, 1985). Recursivamente, a equação 3.1 pode ser escrita da seguinte forma

$$p_{y+1} = \frac{\lambda}{y+1} \left(\frac{y+2}{y+1} \right)^\delta p_y, \tag{3.3}$$

A Figura 3.1 mostra a flexibilidade da distribuição de Touchard que possibilita modelar diferentes formas de distribuição.

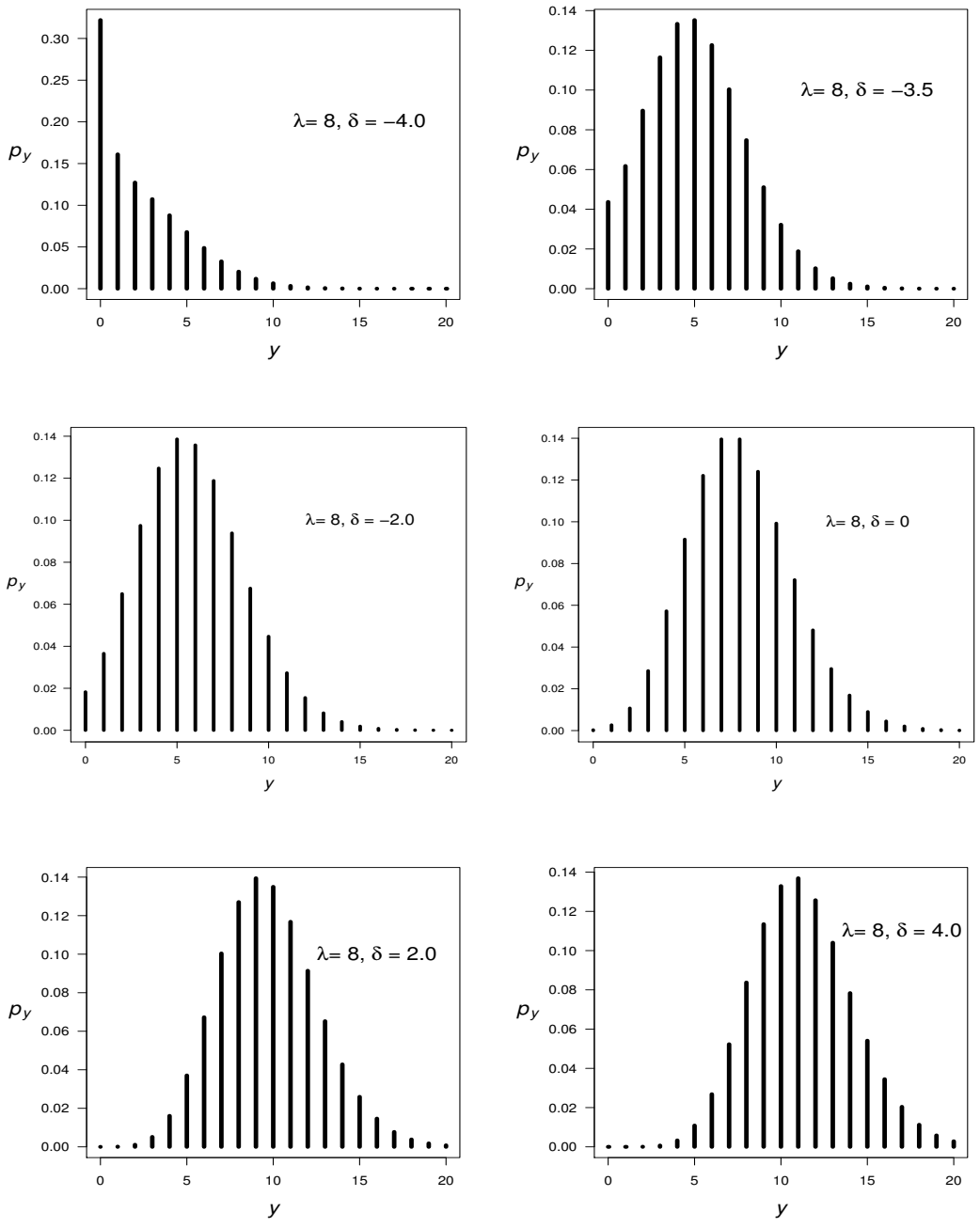

Figura 3.1: Exemplos da distribuição Touchard com $\lambda = 8$ e δ variando entre -4.0 e 4.0. Excessos de zeros aparecem quando $\delta = -4.0$.

3.3 Propriedades da distribuição de Touchard

Seja $Y \sim Touchard(\lambda, \delta)$ o r-ésimo momento de uma variável com distribuição de Touchard é dado por

$$\mathrm{E}[Y^r] = \sum_{j=0}^{r} \binom{r}{j} \frac{(-1)^{r-j}\tau(\lambda, \delta + j)}{\tau(\lambda, \delta)}, \tag{3.4}$$

e sua função geradora de momentos é definida por

$$M_Y(q) = \mathrm{E}[e^{qY}] = \frac{\tau(\lambda e^q, \delta)}{\tau(\lambda, \delta)},$$

em que $q \in \mathbb{R}$.

O valor esperado de Y é

$$\begin{aligned} \mu = \mathrm{E}[Y] &= \frac{\tau(\lambda, \delta + 1)}{\tau(\lambda, \delta)} - 1 \\ &= \lambda E\left[\left(\frac{Y+2}{Y+1}\right)^{\delta}\right], \end{aligned} \tag{3.5}$$

e a sua variância é expressa por

$$\begin{aligned} \sigma^2 = Var(Y) &= \frac{\tau(\lambda, \delta + 2)}{\tau(\lambda, \delta)} - \left[\frac{\tau(\lambda, \delta + 1)}{\tau(\lambda, \delta)}\right]^2 \\ &= \lambda E\left[(Y+1)\left(\frac{Y+2}{Y+1}\right)^{\delta}\right] - \mu^2. \end{aligned} \tag{3.6}$$

A média de Y (3.5) mostra que $\mu > \lambda$ se $\delta > 0$ e $\mu < \lambda$ se $\delta < 0$. Para $\delta = 0$, $Y \sim Poisson(\lambda)$. Para avaliar a dispersão, utiliza-se a razão $d = \frac{\sigma^2}{\mu}$, que pode ser definida como

$$d = \frac{E\left[(Y+1)\left(\frac{Y+2}{Y+1}\right)^{\delta}\right]}{E\left[\left(\frac{Y+2}{Y+1}\right)^{\delta}\right]} - \mu. \tag{3.7}$$

Considerando a distribuição Poisson ($\delta = 0$) tem-se que $d = 1$. Para $\delta > 0$, como Y+1 e $([Y+2]/[Y+1])^{\delta}$ estão inversamente relacionados, logo $d < 1$ (subdispersão). Por outro lado, se $\delta < 0$, então $d > 1$ (superdispersão).

Capítulo 4

Regressão de Touchard

4.1 Introdução

O objetivo deste capítulo é apresentar o modelo de regressão de Touchard que corresponde a uma opção de análise para dados de contagens que violam as suposições do modelo Poisson. Este modelo assume que a variável resposta Y tem distribuição de Touchard com parâmetros λ e δ, ligados a posição e a dispersão, respectivamente. Uma particularidade da distribuição de Touchard é utilizar covariáveis para descrever seus dois parâmetros. De modo geral, a modelagem do parâmetro λ é mais simples, no entanto, em alguns casos pode-se observar que uma determinada variável explicativa pode estar mais ligada com a variabilidade dos dados do que com a média.

4.2 Modelo de regressão de Touchard

Seja $y_1, y_2, \ldots, y_n$ uma amostra aleatória da distribuição de Touchard. O modelo de regressão Touchard é definido assumindo que

$$\lambda_i = \exp\{\mathbf{x}_i'\boldsymbol{\beta}\} \tag{4.1}$$

e

$$\delta_i = \mathbf{z}_i'\boldsymbol{\alpha}, \tag{4.2}$$

em que $\mathbf{x}_i' = (1, x_{i1}, \ldots, x_{ip})$ e $\mathbf{z}_i' = (1, z_{i1}, \ldots, z_{iq})$ são vetores de covariáveis não exclusivos e os termos unitários remetem aos interceptos correspondentes em λ e δ; e $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_p)' \in \mathbb{R}^{p+1}$ e $\boldsymbol{\alpha} = (\alpha_0, \alpha_1, \ldots, \alpha_q)' \in \mathbb{R}^{q+1}$ são vetores de coeficientes desconhecidos.

4.3 Estatísticas suficientes

A distribuição de Touchard é uma integrante da família exponencial biparamétrica, uma vez que pode ser escrita da seguinte forma:

$$p_y = \mathrm{P}[Y = y] = \frac{1}{y!}\exp\{y\ln(\lambda) + \delta\ln(y+1) - \ln[\tau(\lambda,\delta)]\}, \tag{4.3}$$

sendo que $\theta = (\theta_1, \theta_2)$, $\theta_1 = \ln(\lambda)$, $\theta_2 = \delta$ e $b(\theta_1, \theta_2) = \ln[\tau(\lambda, \delta)]$. Desse modo, a distribuição de Touchard não segue a forma usual de um MLG com um único parâmetro canônico, considera-se o par (θ_1, θ_2) os parâmetros de ligação (McCullagh e Nelder, 1989). Assim, tem-se que $\frac{\partial b(\theta_1,\theta_2)}{\partial \theta_1} = \mu$, $\frac{\partial b(\theta_1,\theta_2)}{\partial \theta_1^2} = \sigma^2$, $\frac{\partial b(\theta_1,\theta_2)}{\partial \theta_2} = \mu^*$ e $\frac{\partial b(\theta_1,\theta_2)}{\partial \theta_2^2} = \sigma^{2*}$, em que $\mu^* = \mathrm{E}[\ln(Y+1)]$ e $\mathrm{Var}[\ln(Y+1)]$.

Seja $y_1, y_2, \ldots, y_n$ uma amostra aleatória da distribuição de Touchard a função de verossimilhança pode ser definida por

$$L(\lambda, \delta|y) = \left(\prod_i^n y_i\right)^{-1} \lambda^{S1} e^{\delta S2}[\tau(\lambda,\delta)]^{-n}, \tag{4.4}$$

em que $S_1 = \sum y_i$ e $S_2 = \sum \ln(y_i + 1)$ são estatísticas suficientes pelo Teorema da Fatoração.

A função de log-verossimilhança incluindo as covariáveis é dada por

$$l(\beta, \alpha) = \sum_{i=1}^{n} \ln p_{yi}, \tag{4.5}$$

em que

$$\ln p_{yi} = y_i \ln \lambda_i + \delta_i \ln(y_i + 1) - \ln y_i! - \ln \tau(\lambda_i, \delta_i), \tag{4.6}$$

sendo que λ_i e δ_i são funções de β e α conforme mostra as equações (4.1) e (4.2).

4.4 O escore

Seja $y^* = \ln[y_i + 1]$, $\mu = \mu_i$ e $\mu_i^* = \mathrm{E}[\ln(Y_i + 1)], y = 1, 2, ..., n$.

A função escore, é obtida pela derivação da expressão (4.5) em relação aos coeficientes desconhecidos β e α.

A partir de (4.6), denota-se a log-verossimilhança para a i-ésima observação como $l_i = \ln p_{y_i}$, assim

$$\begin{aligned}
\frac{\partial l_i}{\partial \beta_j} &= y_i x_{ij} - \frac{\partial \ln \tau(\lambda_i, \delta_i)}{\partial \beta_j} \\
&= y_i x_{ij} - \frac{1}{\tau(\lambda_i, \delta_i)} \frac{\partial \tau(\lambda_i, \delta_i)}{\partial \beta_j} \\
&= y_i x_{ij} - \frac{1}{\tau(\lambda_i, \delta_i)} \frac{\partial \tau(\lambda_i, \delta_i)}{\partial \lambda_i} \frac{\partial \lambda_i}{\partial \beta_j} \\
&= y_i x_{ij} - \frac{1}{\tau(\lambda_i, \delta_i)} \frac{\tau(\lambda_i, \delta_i)}{\lambda_i} \mu_i x_{ij} \lambda_i \\
&= x_{ij}(y_i - \mu_i)
\end{aligned} \tag{4.7}$$

para $j = 1, 2, ..., p$.

De maneira similar, tem-se que

$$\begin{aligned}\frac{\partial l_i}{\partial \alpha_k} &= y_i^* z_{ik} - \frac{\partial \ln \tau(\lambda_i, \delta_i)}{\partial \alpha_k} \\ &= y_i^* z_{ik} - \frac{1}{\tau(\lambda_i, \delta_i)} \frac{\partial \tau(\lambda_i, \delta_i)}{\partial \alpha_k} \\ &= y_i^* z_{ik} - \frac{1}{\tau(\lambda_i, \delta_i)} \frac{\partial \tau(\lambda_i, \delta_i)}{\partial \delta_i} \frac{\partial \delta_i}{\partial \alpha_k} \\ &= y_i^* z_{ik} - \frac{1}{\tau(\lambda_i, \delta_i)} \tau(\lambda_i, \delta_i) \mu_i^* z_{ik} \\ &= z_{ik}(y_i^* - \mu_i^*) \end{aligned} \tag{4.8}$$

para $k = 1, 2, ..., q$.

Desse modo, as primeiras derivadas de (4.5) em relação ao β_j e α_k são, respectivamente

$$U_{\beta_j} = \frac{\partial l(\boldsymbol{\beta}, \boldsymbol{\alpha})}{\partial \beta_j} = \sum_{i=1}^{n} x_{ij}(y_i - \mu_i) \tag{4.9}$$

e

$$U_{\alpha_k} = \frac{\partial l(\boldsymbol{\beta}, \boldsymbol{\alpha})}{\partial \alpha_k} = \sum_{i=1}^{n} z_{ik}(y_i^* - \mu_i^*). \tag{4.10}$$

Por fim, chega-se a expressão da matriz para a função escore, em que (4.9) e (4.10) são os elementos correspondentes de $\mathbf{U}_{\boldsymbol{\beta}}$ e $\mathbf{U}_{\boldsymbol{\alpha}}$.

4.5 A hessiana

Seja $\sigma_i^{*2} = \text{Var}[\ln(Y_i + 1)]$ e $\gamma_i = \text{Cov}[Y_i, \ln(Y_i + 1)]$. A matriz hessiana é dada por

$$\mathbf{H} = \begin{pmatrix} \mathbf{H}_{11} & \mathbf{H}_{12} \\ \mathbf{H}_{12} & \mathbf{H}_{22} \end{pmatrix}, \tag{4.11}$$

com

$$\begin{cases} \mathbf{H}_{11} = -\mathbf{X}'\mathbf{D}\mathbf{X}; \\ \mathbf{H}_{22} = -\mathbf{Z}'\mathbf{D}^*\mathbf{Z}; \\ \mathbf{H}_{12} = \mathbf{H}'_{21} = -\mathbf{X}'\mathbf{C}\mathbf{Z}, \end{cases}$$

em que $\mathbf{D} = diag(\sigma_i^2)$, $\mathbf{D}^* = diag(\sigma_i^{*2})$ e $\mathbf{C} = diag(\gamma_i)$. Como $\mathbf{H}$ é uma matriz particionada (Mardia et al., 1980), a inversa é

$$\mathbf{H}^{-1} = \begin{pmatrix} \mathbf{H}^{11}\mathbf{H}^{12} \\ \mathbf{H}^{12}\mathbf{H}^{22} \end{pmatrix}, \tag{4.12}$$

$$\begin{cases} \mathbf{H}^{11} = \left(\mathbf{H}_{11} - \mathbf{H}_{12}{\mathbf{H}_{22}}^{-1}\mathbf{H}_{21}\right)^{-1}; \\ \mathbf{H}^{22} = \left(\mathbf{H}_{22} - \mathbf{H}_{21}{\mathbf{H}_{11}}^{-1}\mathbf{H}_{12}\right)^{-1}; \\ \mathbf{H}^{12} = \mathbf{H}^{12'} = -\mathbf{H}^{11}\mathbf{H}_{12}{\mathbf{H}_{22}}^{-1}. \end{cases}$$

Segundo Casella e Berger (2002) para amostras de tamanho grande, sob as condições de regularidade, $\hat{\boldsymbol{\beta}}$ e $\hat{\boldsymbol{\alpha}}$, estimadores de máxima verossimilhança de $\boldsymbol{\beta}$ e $\boldsymbol{\alpha}$, seguem a distribuição

$$\begin{pmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\boldsymbol{\alpha}} \end{pmatrix} \sim \mathrm{N}\left(\begin{pmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\boldsymbol{\alpha}} \end{pmatrix}, \mathbf{H}^{-1}\right). \tag{4.13}$$

Para encontrar a matriz hessiana realizou-se os seguintes procedimentos: primeiro, obteve-se as derivadas da média μ_i com respeito a λ_i e δ_i que são, respectivamente,

$$\begin{aligned} \frac{\partial \mu_i}{\partial \lambda_i} &= \frac{\partial}{\partial \lambda_i} \sum_{y=0}^{\infty} \frac{y\lambda_i^y (y+1)^{\delta_i}}{y!\tau(\lambda_i, \delta_i)} \\ &= \sum_{y=0}^{\infty} \left[\frac{y^2 \lambda_i^{y-1} (y+1)^{\delta_i}}{y!\tau(\lambda_i, \delta_i)} - \frac{y\lambda_i^y (y+1)^{\delta_i}}{y!\tau^2(\lambda_i, \delta_i)} \frac{\partial \tau(\lambda_i, \delta_i)}{\partial \lambda_i} \right] \\ &= \frac{1}{\lambda_i} \mathrm{E}[Y_i^2] - \sum_{y=0}^{\infty} \left[\frac{y\lambda_i^y (y+1)^{\delta_i}}{y!\tau^2(\lambda_i, \delta_i)} \frac{\tau(\lambda_i, \delta_i)\mu_i}{\lambda_i} \right] \\ &= \frac{\sigma_i^2}{\lambda_i}, \end{aligned} \tag{4.14}$$

e

$$\begin{aligned}
\frac{\partial \mu_i}{\partial \delta_i} &= \frac{\partial}{\partial \delta_i} \sum_{y=0}^{\infty} \frac{y\lambda_i^y (y+1)^{\delta_i}}{y!\tau(\lambda_i, \delta_i)} \\
&= \sum_{y=0}^{\infty} \left[\frac{y \ln(y+1)\lambda_i^y (y+1)^{\delta_i}}{y!\tau(\lambda_i, \delta_i)} - \frac{y\lambda_i^y (y+1)^{\delta_i}}{y!\tau^2(\lambda_i, \delta_i)} \frac{\partial \tau(\lambda_i, \delta_i)}{\partial \delta_i} \right] \\
&= \mathrm{E}[Y_i \ln(Y_i + 1)] - \sum_{y=0}^{\infty} \frac{y\lambda_i^y (y+1)^{\delta_i}}{y!\tau^2(\lambda_i, \delta_i)} \tau(\lambda_i, \delta_i) \, \mathrm{E}[\ln(Y_i + 1)] \\
&= \mathrm{E}[Y_i Y_i^*] - \mu_i \mu_i^* \\
&= \mathrm{Cov}[Y_i, Y_i^*] = \gamma_i. \qquad (4.15)
\end{aligned}$$

Posteriormente, calculou-se a derivada de $\mu_i^* = \mathrm{E}[\ln(Y_i + 1)]$ em relação a δ_i

$$\begin{aligned}
\frac{\partial \mu_i^*}{\partial \delta_i} &= \frac{\partial}{\partial \delta_i} \sum_{y=0}^{\infty} \frac{\ln(y+1)\lambda_i^y (y+1)^{\delta_i}}{y!\tau(\lambda_i, \delta_i)} \\
&= \sum_{y=0}^{\infty} \left[\frac{\ln^2(y+1)\lambda_i^y (y+1)^{\delta_i}}{y!\tau(\lambda_i, \delta_i)} - \frac{\ln(y+1)\lambda_i^y (y+1)^{\delta_i}}{y!\tau^2(\lambda_i, \delta_i)} \frac{\partial \tau(\lambda_i, \delta_i)}{\partial \delta_i} \right] \\
&= \mathrm{E}[(Y_i^*)^2] - \mathrm{E}^2(Y_i^*) \\
&= \mathrm{Var}[Y_i^*] = \sigma^{*2}_i. \qquad (4.16)
\end{aligned}$$

A derivada de (4.7) em relação a $\beta_{j'}$ que corresponde a

$$\begin{aligned}
\frac{\partial^2 l_i}{\partial \beta_j \partial \beta_{j'}} &= \frac{\partial x_{ij}(y_i - \mu_i)}{\partial \beta_{j'}} \\
&= -x_{ij} \frac{\partial \mu_i}{\partial \beta_{j'}} \\
&= -x_{ij} \frac{\partial \mu_i}{\partial \lambda_i} \frac{\partial \lambda_i}{\partial \beta_{j'}} \\
&= -x_{ij} \left(\frac{\sigma_i^2}{\lambda_i} \right) x_{ij'} \lambda_i \\
&= -x_{ij} x_{ij'} \sigma_i^2. \qquad (4.17)
\end{aligned}$$

Também obteve-se a derivada de (4.7) em relação a α_k que é dada por

$$\begin{aligned}\frac{\partial^2 l_i}{\partial\beta_j\partial\alpha_k} &= \frac{\partial x_{ij}(y_i-\mu_i)}{\partial\alpha_k}\\ &= -x_{ij}\frac{\partial\mu_i}{\partial\alpha_k}\\ &= -x_{ij}\frac{\partial\mu_i}{\partial\delta_i}\frac{\partial\delta_i}{\partial\alpha_k}\\ &= -x_{ij}z_{ik}\delta_i \end{aligned} \tag{4.18}$$

A derivada de (4.8) em relação a α'_k é

$$\begin{aligned}\frac{\partial^2 l_i}{\partial\alpha_k\partial\alpha'_k} &= \frac{\partial z_{ik}(y_i^*-\mu_i^*)}{\partial\alpha'_k}\\ &= -z_{ik}\frac{\partial\mu_i^*}{\partial\alpha'_k}\\ &= -z_{ik}\frac{\partial\mu_i^*}{\partial\delta_i}\frac{\partial\delta_i}{\partial\alpha'_k}\\ &= -z_{ik}z_{ik'}\sigma^{*2}. \end{aligned} \tag{4.19}$$

Portanto, a derivada de segunda ordem de (4.5) em relação a β_j and α_k é dada por

$$\frac{\partial^2 l(\boldsymbol{\beta},\boldsymbol{\alpha})}{\partial\beta_j\partial\beta'_j} = \sum_{i=1}^{n} x_{ij}x_{ij'}\sigma_i^2, \tag{4.20}$$

$$\frac{\partial^2 l(\boldsymbol{\beta},\boldsymbol{\alpha})}{\partial\beta_j\partial\alpha_k} = \sum_{i=1}^{n} x_{ij}z_{ik}\gamma_i \tag{4.21}$$

e

$$\frac{\partial^2 l(\boldsymbol{\beta},\boldsymbol{\alpha})}{\partial\alpha_k\partial\alpha'_k} = \sum_{i=1}^{n} z_{ik}z_{ik'}\sigma^{*2}_i, \tag{4.22}$$

que fornecem a matriz hessiana apresentada na expressão (4.11).

4.6 Implementação computacional

O modelo Touchard foi ajustado no software R e as estimativas de máxima verossimilhança para os parâmetros λ e δ da distribuição e para os parâmetros da regressão β e α foram obtidas por meio das funções (dtouchard, mle.touchard e touch.reg) desenvolvidas no trabalho de Oliveira (2017). Neste estudo foram feitas algumas alterações nessas funções e aplicou-se a versão que utiliza a função de otimização interna do R, nlminb [1].

4.7 Considerações

Este capítulo apresentou o modelo de regressão de Touchard, sua estimação e detalhou os procedimentos para encontrar a função escore e a matriz hessiana. Além disso, foram abordadas duas particularidades da distribuição Touchard: a distribuição pertence a família exponencial biparamétrica, no entanto, não segue a forma usual de um MLG, pois apresenta um par de parâmetros de ligação e, ainda, a distribuição usa covariáveis tanto para descrever o parâmetro λ, quanto o parâmetro δ.

[1] https://stat.ethz.ch/R-manual/R-devel/library/stats/html/nlminb.htmlzx

Capítulo 5

Alguns elementos de diagnóstico para a Regressão de Touchard

5.1 Introdução

A análise de diagnósticos agrega várias ferramentas estatísticas para avaliar a adequabilidade de um modelo, entretanto, este capítulo se limita a desenvolver discussões pertinentes à estatística χ^2 generalizada de Pearson (Q) e a deviance residual (D). Tais estatísticas são medidas frequentemente utilizadas para os diagnósticos de modelos lineares generalizados (MLG's).

Em razão das particularidades da regressão de Touchard, as medidas Q, e D se comportam de forma distinta daquelas proporcionadas pelos modelos MLG's tradicionais.

A estatística Q, é definida como uma junção dos quadrados dos resíduos de Pearson $(y_i - \hat{\mu}_i)^2/v(\hat{\mu}_i)$, em que y_i representa a variável resposta ligada ao i-ésimo elemento da amostra, $\hat{\mu}_i$ representa a estimativa da sua resposta esperada segundo o modelo ajustado, e $v(\hat{\mu}_i)$ denota a função de variação. Para o caso em que a função de variação dependa de um segundo parâmetro, Smyth (2003) desenvolveu um resultado, desde que esse parâmetro adicional não dirija a média. Na regressão de Touchard, porém, tem-se, $\hat{\mu}_i = \hat{\mu}_i(\lambda_i, \delta_i)$ e $v(\lambda_i, \delta_i)$, ou seja, tanto a média como a função de variação depende de dois parâmetros. Por meio de simulação computacional, mostra-se que a estatística Q não segue a distribuição χ^2 da forma

esperada.

5.2 Deviance

Seja Θ o vetor de parâmetros, a deviance corresponde ao logaritmo da estatística da razão de verossimilhança entre o modelo ajustado $l(\hat{\Theta})$ e o modelo saturado $l(\hat{\Theta}_{max})$, sendo que o modelo saturado refere-se ao modelo que ajusta um coeficiente para cada observação. O problema na regressão de Touchard é que além de não haver solução única para tais coeficientes, tem-se $l(\hat{\Theta}_{max}) \approx 0$. Isto é, se y_i for o valor modal com máxima probabilidade, tal probabilidade seria praticamente igual a 1 para pelo menos um par (λ_i, δ_i).

A deviance do modelo Touchard pode ser obtida utilizando a forma recurssiva

$$p_{y+1} = \frac{\lambda}{y+1}\left(\frac{y+2}{y+1}\right)^{\delta} p_y. \tag{5.1}$$

Seja y^* a moda de Y, implica que $p_{y^*-1} < p_{y^*}$ e $p_{y^*+1} < p_{y^*}$, isto é, o parâmetro λ da Touchard poderia assumir os seguintes valores: $y^*\left(\frac{y^*}{y^*+1}\right)^{\delta} < \lambda < (y^*+1)\left(\frac{y^*+1}{y^*+2}\right)^{\delta}$. Além disso, para um δ suficientemente grande, $\frac{\lambda^{y^*}(y^*+1)^{\delta}}{y^*!}$ torna-se o termo dominante da equação (3.2) e $p_{y^*} \approx 1$.

Para ilustrar, considere o caso em que $y^* = 2$, $\lambda = \exp(-97)$ e $\delta = 282$, tem-se $p_2 \approx 1$. Neste caso, a distribuição apresenta $\mu \approx y^*$ e $\sigma^2 \approx 0$.

Assim, a log-verossimilhança do modelo saturado é aproximadamente nula. Em razão dessa característica, a definição usual da deviance torna-se sem utilidade.

5.3 Estatísticas Qui-quadrado

5.3.1 Distribuição amostral dos escores

Seja $\mathbf{U}_{\boldsymbol{\beta}} = \mathbf{X}' \cdot (\mathbf{y} - \boldsymbol{\mu})$ e que $\mathbf{H}_{\beta\beta}^{-1} = (\mathbf{X}'\mathbf{D}\mathbf{X})^{-1}$, com $\mathbf{D} = diag(\sigma_i^2)$, temos $\mathrm{E}[\mathbf{U}_{\boldsymbol{\beta}}] = \mathbf{X}' \cdot (\mathrm{E}[\mathbf{y}] - \boldsymbol{\mu}) = 0$, pois $\mathrm{E}(y) = \mu$ e $\mathrm{Var}(\mathbf{U}_{\boldsymbol{\beta}}) = \mathbf{H}_{\beta\beta}$. Para o escore aproximadamente gaussiano, tem-se a seguinte forma quadrática

$$\mathbf{U}'_{\boldsymbol{\beta}}\mathbf{H}^{-1}_{\beta\beta}\mathbf{U}_{\boldsymbol{\beta}} = (\mathbf{y}-\boldsymbol{\mu})' \cdot \mathbf{X}(\mathbf{X}'\mathbf{D}\mathbf{X})^{-1}\mathbf{X}' \cdot (\mathbf{y}-\boldsymbol{\mu})$$

$$= (\mathbf{y}-\boldsymbol{\mu})'\mathbb{J}_{\boldsymbol{\beta}}(\mathbf{y}-\boldsymbol{\mu}) \sim \chi^2_{(p)} \tag{5.2}$$

em que $\mathbb{J}_{\boldsymbol{\beta}} = \mathbf{X}(\mathbf{X}'\mathbf{D}\mathbf{X})^{-1}\mathbf{X}'$ é uma matriz $n \times n$.

Tomando o resíduo com média zero e variância unitária como

$$r_i = \frac{(y_i - \mu_i)}{\sigma_i}, \tag{5.3}$$

ou em notação vetorial,

$$\mathbf{r} = \mathbf{D}^{-\frac{1}{2}} \cdot (\mathbf{y}-\boldsymbol{\mu}). \tag{5.4}$$

Assim, a equação (5.2) pode ser expressa por

$$\mathbf{r}' \cdot \mathbf{D}^{\frac{1}{2}}\mathbf{X}(\mathbf{X}'\mathbf{D}\mathbf{X})^{-1}\mathbf{X}'\mathbf{D}^{\frac{1}{2}} \cdot \mathbf{r} = \mathbf{r}' \cdot \mathbb{P}_{\boldsymbol{\beta}} \cdot \mathbf{r} \sim \chi^2_{(p)}, \tag{5.5}$$

em que $\mathbb{P}_{\boldsymbol{\beta}} = \mathbf{D}^{\frac{1}{2}}\mathbf{X}(\mathbf{X}'\mathbf{D}\mathbf{X})^{-1}\mathbf{X}'\mathbf{D}^{\frac{1}{2}}$é uma matriz de projeção, pois ela é simétrica e idempotente. Nota-se que $\mathbb{P}_{\boldsymbol{\beta}} \cdot \mathbf{D}^{\frac{1}{2}}\mathbf{X} = \mathbf{D}^{\frac{1}{2}}\mathbf{X}$, e $\mathbf{X}'\mathbf{D}^{\frac{1}{2}} \cdot \mathbb{P}_{\boldsymbol{\beta}} = \mathbf{X}'\mathbf{D}^{\frac{1}{2}}$. No caso em que $\hat{\mu}$ é o estimador de máxima verossimilhança, $\mathbf{X}'(\mathbf{y}-\hat{\boldsymbol{\mu}}) = 0$, então $\mathbb{P}_{\boldsymbol{\beta}} \cdot \hat{\mathbf{r}} = 0$ e $\hat{\mathbf{r}} = \mathbf{D}^{-\frac{1}{2}}(\mathbf{y}-\hat{\boldsymbol{\mu}})$ denomina-se resíduo de Pearson.

Cada elemento da diagonal de $\mathbb{P}_{\boldsymbol{\beta}}$, $L_i = [\mathbb{P}_{\boldsymbol{\beta}}]_{ii}$ denomina-se leverage, e com base nele define-se o resíduo padronizado como

$$rp_i = \frac{(y_i - \mu_i)}{\sigma_i\sqrt{1-L_i}}. \tag{5.6}$$

5.3.2 Caso saturado

Considerando o caso saturado em que $\mathbf{X} = \mathbf{I}_{n\times n}$, na equação (5.2) tomaria-se $\mathbb{J}_{\boldsymbol{\beta}} = \mathbf{I}(\mathbf{I}'\mathbf{D}\mathbf{I})^{-1}\mathbf{I}' = \mathbf{D}^{-1}$, então

$$(\mathbf{y}-\boldsymbol{\mu})'\mathbf{D}^{-1}(\mathbf{y}-\boldsymbol{\mu}) = \sum_{i=1}^{n}\frac{(y_i-\mu_i)^2}{\sigma_i^2} \sim \chi^2_{(n)}. \tag{5.7}$$

Uma medida utilizada para testar o ajuste dos MLG's, refere-se a uma generalização da estatística χ^2 proposta originalmente por Pearson (1900) para o modelo multinomial. Considere uma resposta y_i, uma estimativa do valor esperado $\hat{\mu}_i$, uma função de variação $v(\hat{\mu}_i)$ e o parâmetro de dispersão $\phi = 1$, a estatística generalizada de Pearson é dada por (McCullagh e Nelder, 1989)

$$Q(\hat{\boldsymbol{\mu}}) = \sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i)^2}{v(\hat{\mu}_i)}. \tag{5.8}$$

Quando o modelo for adequado, para uma amostra suficientemente grande, a estatística $Q(\hat{\boldsymbol{\mu}})$ tenderá a seguir uma distribuição χ^2 com o número de graus de liberdade residual.

Contudo, no modelo de regressão de Touchard tem-se $\mu(\hat{\lambda}_i, \hat{\delta}_i)$ e $v(\hat{\lambda}_i, \hat{\delta}_i)$ e para essa situação, não há na literatura uma orientação sobre a distribuição amostral da estatística

$$Q(\hat{\boldsymbol{\lambda}}, \hat{\boldsymbol{\delta}}) = \sum_{i=1}^{n} \frac{[y_i - \mu(\hat{\lambda}_i, \hat{\delta}_i)]^2}{v(\hat{\lambda}_i, \hat{\delta}_i)}. \tag{5.9}$$

Em contrapartida, se houver uma proposta hipotética para os parâmetros do modelo, tem-se que

$$Q(\boldsymbol{\lambda}, \boldsymbol{\delta}) = \sum_{i=1}^{n} \frac{[y_i - \mu(\lambda_i, \delta_i)]^2}{v(\lambda_i, \delta_i)} \tag{5.10}$$

segue aproximadamente uma distribuição χ^2 com n graus de liberdade, por conta do teorema do limite central. Para exemplificar, considere a seguinte simulação computacional.

Exemplo 5.1. Neste exemplo para geração de dados, considera-se um modelo com oito parâmetros e uma amostra de tamanho $n = 1000$, com $\boldsymbol{\beta} = (0.5, 0.3, -0.2, 0.2, -0.1)$' e $\boldsymbol{\alpha} = (1.0, 0.1, -0.1)'$, com covariáveis $x_i = (1, x_{i1}, x_{i2}, x_{i3}, x_{i4})'$ e $z_i = (1, x_{i1}, x_{i3})'$, tais que $x_{i1}, x_{i2}, x_{i3}, x_{i4}$ sejam, respectivamente, realizações de Bernoulli(0,6), Bernoulli(0,3), Normal (1,1) e Normal (2,1), para $i = 1, ..., 1000$. Com base nesses dados, são obtidos $n = 1000$ pares de parâmetros (λ_i, δ_i), em que $\lambda_i = \exp\{\mathbf{x}_i'\boldsymbol{\beta}\}$ e $\delta_i = \exp\{\mathbf{z}_i'\boldsymbol{\alpha}\}$. Utilizando as covariáveis e os parâmetros hipotéticos até o final do experimento, realiza-se os seguintes procedimentos:

1. para cada $i = 1, ..., 1000$, obtém-se uma realização $y_i \sim Touchard(\lambda_i, \delta_i)$;

2. utiliza-se o vetor de resposta $(y_i, ..., y_{1000})$ e o conjunto de covariáveis para obter as estimativas de máxima verossimilhança $\hat{\lambda}$ e $\hat{\alpha}$;

3. registram-se as estatísticas qui-quadrado de Pearson;

4. repetem-se os passos 1 a 3 r vezes.

A tabela exibe os resultados da simulação, os valores críticos empíricos correspondentes a $r = 1000$ réplicas das estatísticas $Q(\hat{\boldsymbol{\lambda}}, \hat{\boldsymbol{\delta}})$ e $Q(\boldsymbol{\lambda}, \boldsymbol{\delta})$. Para a estatística $Q(\boldsymbol{\lambda}, \boldsymbol{\delta})$, tem-se que $\hat{\text{E}}[Q(\boldsymbol{\lambda}, \boldsymbol{\delta})] \approx 997$ e $\hat{\text{Var}}[Q(\boldsymbol{\lambda}, \boldsymbol{\delta})] \approx 2396$, notam-se proximidades entre os valores empíricos e teóricos. Já para $Q(\hat{\boldsymbol{\lambda}}, \hat{\boldsymbol{\delta}})$ obteve-se $\hat{\text{E}}[Q(\hat{\boldsymbol{\lambda}}, \hat{\boldsymbol{\delta}})] \approx 999$ e $\hat{\text{Var}}[Q(\hat{\boldsymbol{\lambda}}, \hat{\boldsymbol{\delta}})] \approx 334$, e produziu valores críticos distantes daqueles da distribuição χ^2. Esses resultados indicam que sua distribuição amostral não seja χ^2.

Tabela 5.1: Percentis empíricos $\hat{q}_\pi$, $\pi = 90\%, 95\%, 97,5\%$ e 99%, correspondentes às estatísticas $Q(\hat{\boldsymbol{\lambda}}, \hat{\boldsymbol{\delta}})$ e $Q(\boldsymbol{\lambda}, \boldsymbol{\delta})$, com tamanho amostral igual a $n = 1000$, obtidos com base em $r = 1000$ realizações. Os valores entre parênteses referem-se aos percentis teóricos da distribuição assintótica.

estatística	90%	95%	97,5%	99%
$Q(\hat{\boldsymbol{\lambda}}, \hat{\boldsymbol{\delta}})$	1022.652	1029.688	1035.090	1041.581
$\chi^2_{(992)}$	(1049.494)	(1066.385)	(1081.180)	(1098.552)
$Q(\boldsymbol{\lambda}, \boldsymbol{\delta})$	1059.062	1082.263	1094.652	1107.447
$\chi^2_{(1000)}$	(1057.724)	(1074.679)	(1089.531)	(1106.969)

Apesar que a distribuição amostral de $Q(\hat{\boldsymbol{\lambda}}, \hat{\boldsymbol{\delta}})$ possa ser obtida computacionalmente, neste trabalho, por analogia ao método da validação cruzada, propõe-se o seguinte procedimento para a modelagem e a avaliação da adequabilidade do modelo:

1. divide-se aleatoriamente o conjunto de dados de tamanho n em duas partes mutuamente excludentes, com tamanhos $n1$ e $n2$, em que $n_1 + n_2 = n$ e $n_1 > n_2$;

2. aplica-se um modelo de regressão de Touchard para os dados da parte 1, de tamanho n_1, estimam-se seus coeficientes $\boldsymbol{\beta}$ e $\boldsymbol{\alpha}$;

3. aplica-se as estimativas dos coeficientes $\boldsymbol{\beta}$ e $\boldsymbol{\alpha}$ na segunda parte dos dados e encontra-se n_2 estimativas de λ e δ, e, em seguida, de μ e v, sob a hipótese de

que o modelo ajustado descreva de forma adequada y da segunda parte dos dados;

4. com base nessas respostas (parte 2) calcula-se $Q(\hat{\boldsymbol{\lambda}}, \hat{\boldsymbol{\delta}})$, comparando-a com os valores críticos de uma distribuição $\chi^2_{(n_2)}$.

Capítulo 6

A distribuição do número de partos no estado de Goiás

6.1 Introdução

Segundo a Organização Mundial da saúde (OMS), a taxa ideal de cesárea se encontra entre 10% e 15%. Nessa faixa estão incluídos os casos de intervenção cesárea sob recomendação médica, os quais permitem salvar vidas de mães e bebês. Por outro lado, taxas superiores a 15% não se associam a redução da mortalidade materna e neonatal, além de aumentar o risco das pacientes às complicações pós-operatórias e elevar os custos dos serviços de saúde. Diante do aumento acima do nível considerado ideal do número de cesáreas em muitos países, a OMS estabeleceu diretrizes para o atendimento de mulheres grávidas saudáveis, recomendando a redução de intervenções médicas desnecessárias (WHO, 2018).

No Brasil as taxas de cesáreas são muito superiores ao ideal estabelecido pela OMS. No caso particular de Goiás, nas segundas feiras do mês de maio de 2017, 74,2% dos partos foram cesáreas, segundo o Sistema de Informação sobre Nascidos Vivos (SINASC).

Ficou evidenciado pela análise a seguir que a taxa de cesárea é determinada pela condição socioeconômica da região em análise, de modo que quanto mais rica a região, maior a taxa de cesáreas. A taxa de cesáreas das três microrregiões mais pobres de Goiás (Entorno de Brasília, Vão do Paranã e Chapada dos Veadeiros) é

51,1% menor do que as treze demais microrregiões do estado. Por outra perspectiva, considerando a divisão do estado nas cinco mesorregiões geográficas (IMB, 2013) e juntando-se as duas mais pobres (Leste e Norte), observou-se taxa de cesáreas igual a 58,5%. Outra variável que indica a relação entre a taxa de cesárea e o poder econômico é o diferencial das taxas de cesáreas entre os estabelecimentos empresariais e os demais estabelecimentos (públicos ou sem fins lucrativos). No período de análise, o percentual de cesáreas em relação ao total de partos foi de 93,4% nos estabelecimentos empresariais, enquanto que nos demais estabelecimentos foi de 55,7%.

Esforços nacionais e internacionais têm sido empenhados para se mudar esse quadro, como por exemplo, a *Iniciativa Hospital Amigo da Criança* (IHAC), idealizada pelo Fundo das Nações Unidas para a Infância (Unicef) e a OMS (UNICEF, 2008). O Ministério da Saúde confere o selo de qualidade IHAC aos hospitais que cumprem alguns quesitos, como o de assegurar cuidados que promovam a redução de procedimentos invasivos, o que inclui partos cesáreas. Atualmente, há 324 estabelecimentos de saúde certificados com esse selo em todo o país, sendo que 20 deles se encontram no estado de Goiás (GO). Nesses hospitais, considerando o mesmo período analisado, a taxa de partos cesárea foi de 46,1%, enquanto que nos demais hospitais foi de 85,2%.

O propósito deste Capítulo é exemplificar a aplicação do modelo de regressão de Touchard para estudar essa distribuição de partos. Essa modelagem permitiria avaliar melhor a contribuição de cada um desses aspectos (como as características socioeconômicas e o perfil do estabelecimento) nas contagens desses partos. Inicialmente, uma breve descrição das variáveis que compõem a base de dados é feita na Seção 6.2. De maneira exploratória, a distribuição do número diário de partos segundo o tipo de parto (normal ou cesárea) é relacionada com possíveis variáveis regressoras, como a existência de selo IHAC no estabelecimento de saúde, a natureza jurídica do estabelecimento (pública, entidade empresarial ou entidade sem fins lucrativos) e regiões geográficas.

6.2 Descrição das variáveis

A massa de dados contempla 908 registros diários feitos por 162 estabelecimentos de saúde distribuídos em 89 municípios goianos. Esses dados, referentes às segundas-feiras do mês de maio de 2017, foram coletados do Sistema de Informação sobre Nascidos Vivos (SINASC) em março de 2018. Informações acerca das regiões de saúde foram obtidas junto à Secretaria da Saúde do Estado de Goiás (SES-GO), e os dados socioeconômicos foram levantados junto ao Instituto Brasileiro de Geografia e Estatística (IBGE). Os dados organizados se encontram disponíveis em `https://1drv.ms/f/s!Apx60k7TMXzegaAZlSPMbQ9HoVwcsg`.

A Tabela 6.1 apresenta a descrição das variáveis. Cada estabelecimento é identificado pelo seu número no Cadastro Nacional de Estabelecimentos de Saúde (`cnes`), ou pelo seu nome (`hosp`). A variável `y` representa o número de partos registrados por um estabelecimento em determinada segunda-feira, e `normal` é uma variável indicadora que assume valor 1, se esse número refere-se a partos normais, ou 0, se cesáreas. É importante ressaltar que consideraram-se apenas os dias em que pelo menos um parto, seja normal ou cesárea, foi registrado no estabelecimento de saúde. Por isso, a massa de dados não apresenta concentração de zeros. A Tabela 6.2 mostra as distribuições dos números diários de partos (`y`) de acordo com o tipo, normal (1) ou cesárea (0). Enquanto a moda do número de partos normais é zero, a distribuição do número de cesáreas apresenta concentração de valores 1. Essas distribuições parecem depender de determinadas variáveis, como pode ser visto a seguir.

A localização do hospital é dada pelas variáveis `munic` (município) — ou `cod.mun` (código do município) —, `micro.r` (microrregião), `meso.r` (mesorregião) e `hlth.r` (região de saúde). A Tabela 6.3 apresenta as taxas de partos normais e cesáreas observadas nas microrregiões de Goiás em maio de 2017. As taxas de intervenções cesáreas são elevadas, acima de 55%, com exceção da Chapada dos Veadeiros, Entorno de Brasília e Vão do Paranã. Todavia, essas regiões são consideradas extremamente pobres (IMB, 2013). Assim, para o nosso estudo, para indicar o hospital que se encontra nessas microrregiões, define-se a variável `poor.mi`. Com respeito à divisão geográfica por mesorregiões, a Tabela 6.4 evidencia menores taxas de ce-

sáreas no Leste e no Norte goiano. Essas mesoregiões contemplam municípios com pobreza extrema (IMB, 2013). Para indicar partos registrados nesses locais, define-se a variável `poor.me`. A Secretaria de Estado da Saúde de GO (SES-GO) define uma divisão geográfica denominada regiões de saúde. A Tabela 6.5, mostra as distribuições das taxas de partos normais e cesáreas segundo essas 18 regiões, as quais novamente apontam para uma menor incidência de cesáreas em localidades mais pobres (indicadas por `poor.rs`).

Sugere-se também um efeito na distribuição dos partos proporcionado pelo tipo do estabelecimento (Tabela 6.6). Nas entidades empresariais, a taxa de cesáreas é 93% do total de partos, valor considerado alarmante, pois está absurdamente acima do limite recomendado pela OMS (WHO, 2018). A Tabela 6.7 mostra que em estabelecimentos que não atendem pelo SUS (Sistema Único de Saúde) incidem taxas igualmente elevadas de cesáreas.

Em contraste, nos 20 hospitais participantes do programa *Iniciativa Hospital Amigo da Criança* (IHAC) do Unicef e da OMS (UNICEF, 2008) no estado de GO, a taxa de cesáreas é de 46,1%, enquanto nos demais hospitais essa taxa é de 85,2%.

Finalmente, o porte do estabelecimento de saúde é um fator naturalmente importante. Como seu indicador, neste trabalho definimos

$$\texttt{LPT.4} = \ln(\texttt{PT.4} + 1), \tag{6.1}$$

em que `PT.4` denota a quantidade média diária de partos (normais + cesáreas) registrados no mês anterior. A correlação linear entre o indicador do porte `LPT.4` e a contagem `y` é aproximadamente igual a 0,57, e a dispersão entre ambas se encontra ilustrada na Figura 6.1.

Portanto, por uma perspectiva descritiva é possível observar que a distribuição de partos por tipo (normal ou cesárea) depende de variáveis regressoras, tais como a existência de selo IHAC no estabelecimento de saúde, a natureza jurídica do estabelecimento (pública, entidade empresarial ou entidade sem fins lucrativos), condições socioeconômicas (representadas por regiões geográficas, incluindo-se o PIB municipal e o índice de desenvolvimento humano), o porte do hospital e o tamanho

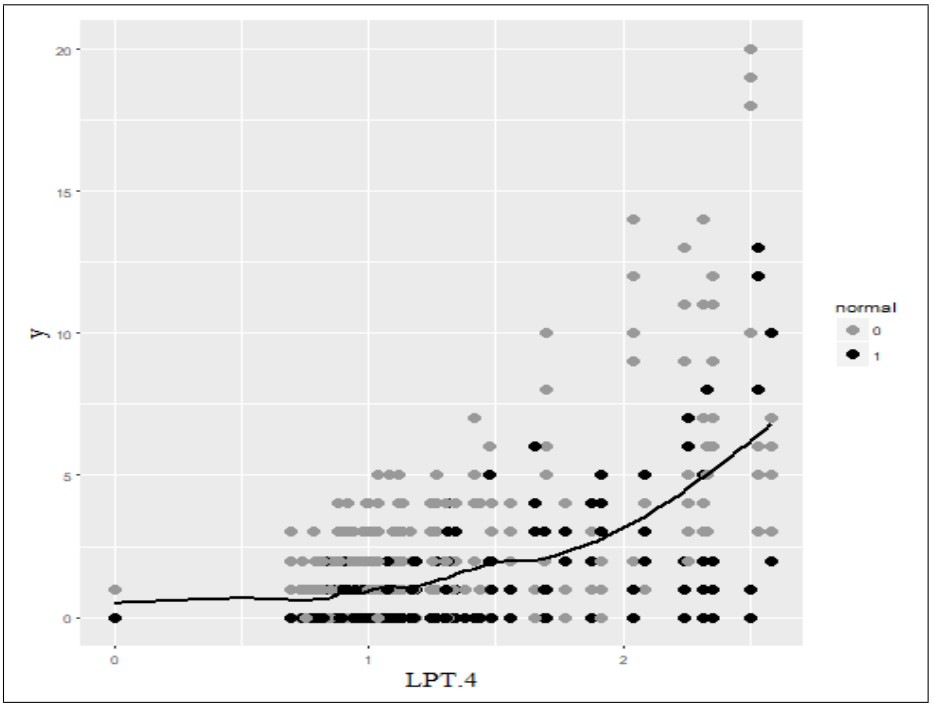

Figura 6.1: `LPT.4` *versus* `y`. As linhas sólidas representam as médias condicionaisy|`LPT.4` ajustadas não parametricamente pelo método LOESS.

populacional.

A seguir, será realizado um estudo com base no modelo de regressão de Touchard, na tentativa de encontrar as variáveis regressoras mais relevantes, separando os efeitos de cada uma delas na distribuição de `y`.

Tabela 6.1: Variáveis encontradas no arquivo `partos.csv`.

variável	descrição
`cnes`	identificação no Cadastro Nacional de Estabelecimentos de Saúde
`hosp`	nome do estabelecimento de saúde
`munic`	nome do município
`cod.mun`	código do município
`date`	data do registro no formato ano-mês-dia
`PT.4`	quantidade média diária de partos registrados no mês anterior
`sus`	1 (atendimento pelo SUS) ou 0 (caso contrário)
`h.type`	tipo do estabelecimento (público, entidade empresarial ou sem fins lucrativos)
`pub`	1 (público) ou 0 (caso contrário)
`emp`	1 (entidade empresarial) ou 0 (caso contrário)
`ihac`	1 (com selo IHAC) ou 0 (caso contrário)
`micro.r`	microrregião
`meso.r`	mesorregião
`hlth.r`	região da saúde
`poor.rs`	1 (região da saúde com alta taxa de pobreza) ou 0 (caso contrário)
`poor.me`	1 (mesoregião com alta taxa de pobreza) ou 0 (caso contrário)
`poor.mi`	1 (microrregião com alta taxa de pobreza) ou 0 (caso contrário)
`rm`	1 (região metropolitana de Goiânia) ou 0 (caso contrário)
`ride`	1 (região integrada de desenvolvimento do DF e entorno) ou 0 (caso contrário)
`y`	total diário de partos
`normal`	1 (normal) ou 0 (cesárea)
`cXsus`	(1-`normal`)× `sus`
`cXemp`	(1-`normal`)× `emp`
`LPT.4`	$\ln(\texttt{PT.4} + 1)$
`pop`	população do município
`pib`	produto interno bruto do município
`idhm`	índice de desenvolvimento humano
`idhm.E`	índice de desenvolvimento humano (educação)
`idhm.L`	índice de desenvolvimento humano (longevidade)
`idhm.R`	índice de desenvolvimento humano (renda)

Fontes: DATASUS/SINASC, SES-GO e IBGE.

Tabela 6.2: Número diário de partos (`y`) registrados nas segundas-feiras de maio de 2017 no estado de GO, segundo o tipo, normal (1) ou cesárea (0).

`normal`	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	18	19	20	Sum
0	60	197	85	39	30	12	8	4	1	3	3	3	2	1	2	1	2	1	454
1	298	85	36	11	5	8	4	1	3	0	1	0	1	1	0	0	0	0	454
Total	358	282	121	50	35	20	12	5	4	3	4	3	3	2	2	1	2	1	908

Tabela 6.3: Taxas de partos normais (1) e cesáreas (0), por microrregiões do estado de Goiás em maio de 2017.

microrregião	0	1
Anápolis	0.76	0.24
Anicuns	1.00	0.00
Aragarças	0.91	0.09
Catalão	0.88	0.12
Ceres	0.73	0.27
Chapada dos Veadeiros	0.17	0.83
Entorno de Brasília	0.55	0.45
Goiânia	0.72	0.28
Iporá	1.00	0.00
Meia Ponte	0.78	0.22
Pires do Rio	0.57	0.43
Porangatu	0.77	0.23
Quirinópolis	1.00	0.00
Rio Vermelho	0.69	0.31
São Miguel do Araguaia	0.72	0.28
Sudoeste de Goiás	0.75	0.25
Vale do Rio dos Bois	0.96	0.04
Vão do Paranã	0.50	0.50

Tabela 6.4: Taxas de partos normais (1) e cesáreas (0), por mesorregiões do estado de Goiás em maio de 2017.

mesorregião	0	1
Centro Goiano	0.74	0.26
Leste Goiano	0.54	0.46
Noroeste Goiano	0.75	0.25
Norte Goiano	0.68	0.32
Sul Goiano	0.80	0.20

Tabela 6.5: Taxas de partos normais (1) e cesáreas (0), por regiões de saúde do estado de Goiás, em maio de 2017.

	0	1
Central	0.75	0.25
Centro Sul	0.63	0.37
Entorno Norte	0.47	0.53
Entorno Sul	0.66	0.34
Estrada de Ferro	0.84	0.16
Nordeste I	0.20	0.80
Nordeste II	0.47	0.53
Norte	0.69	0.31
Oeste I	0.97	0.03
Oeste II	1.00	0.00
Pirineus	0.73	0.27
Rio Vermelho	0.80	0.20
São Patrício I	0.77	0.23
São Patrício II	0.61	0.39
Serra da Mesa	0.78	0.22
Sudoeste I	0.77	0.23
Sudoeste II	0.81	0.19
Sul	0.84	0.16

Tabela 6.6: Taxas de partos normais (1) e cesáreas (0), por tipo de estabelecimento, em maio de 2017.

	0	1
Administração Pública	0.55	0.45
Entidades Empresariais	0.93	0.07
Entidades sem Fins Lucrativos	0.57	0.43

Tabela 6.7: Taxas de partos normais (1) e cesáreas (0), por estabelecimento que atende pelo SUS ou não, em maio de 2017.

sus	0	1
não	0.95	0.05
sim	0.67	0.33

6.3 Modelagem e Análise

Como foi introduzido no Capítulo 4, suponha que a contagem y_i seja descrita por uma distribuição de Touchard com parâmetros λ_i and δ_i, tais que

$$\lambda_i = \exp\{\mathbf{x}_i'\boldsymbol{\beta}\} \tag{6.2}$$

e

$$\delta_i = \mathbf{z}_i'\boldsymbol{\alpha}, \tag{6.3}$$

em que $\mathbf{x}_i' = (1, x_{i1}, \ldots, x_{ip})$ e $\mathbf{z}_i' = (1, z_{i1}, \ldots, z_{iq})$ são vetores constituídos por valores das covariáveis referentes ao i-ésimo registro da massa de dados; os termos unitários remetem aos interceptos correspondentes em λ e δ; e $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_p)' \in \mathbb{R}^{p+1}$ e $\boldsymbol{\alpha} = (\alpha_0, \alpha_1, \ldots, \alpha_q)' \in \mathbb{R}^{q+1}$ são vetores de coeficientes desconhecidos. A estimação desses coeficientes é feita por máxima verossimilhança, tendo sido estabelecidos como critérios gerais de seleção do modelo o AIC (*Akaike Information Criterion*), o BIC (*Bayesian Information Criterion*), o coeficiente χ^2 e o $pseudo - R^2$ (Capítulo 2).

Em particular, tem-se $p, q \leq 16$, sendo que as variáveis regressoras disponíveis para o nosso estudo são `normal`, `sus`, `pub`, `emp`, `ihac`, `poor.rs`, `poor.me`, `poor.mi`, `rm`, `ride`, `cXsus`, `cXemp`, `LPT.4`, `lpop`, `lpib` e `idhm` (Tabela 6.1), em que `lpib` = ln(`pib`)/100 e `lpop` = ln(`pop`)/100.

Iniciamos com o modelo somente com os interceptos β_0 e α_0 (Modelo 0), pois sua log verossimilhança ($l = -1264.499$) é base de referência para a obtenção dos $pseudo - R^2$ de alguns modelos subsequentes exemplificados na Tabela 6.8. Por exemplo, o Modelo 1 considera apenas a variável `normal`, que indica o tipo de parto normal (1) ou cesárea (0). Apesar dessa variável proporcionar algum ganho (maior verossimilhança e valores de AIC e BIC inferiores ao modelo 0, o p-valor da estatística χ^2 é baixo, o que sugere falta de ajuste (Tabela 6.9). Em outro extremo, incluindo-se todas as 16 variáveis regressoras, tanto para os coeficientes $\boldsymbol{\beta}$ como para os $\boldsymbol{\alpha}$, embora haja verossimilhança elevada, os valores de AIC e BIC não são minimizados. Além disso, o ganho no valor do pseudo-R^2 não é substancial em comparação com

os modelos com menos coeficientes.

Dos modelos mostrados na Tabela 6.9, o Modelo 6 apresentou o menor valor AIC, tendo apresentado um valor estatististicamente não significativo para a estatística χ^2 (p-valor = 0.519). A Tabela 6.10 mostra as estimativas dos coeficientes, os quais mostraram-se estatisticamente significativos.

Cabe salientar que para o modelo de Touchard, dependendo da situação, a contribuição de cada coeficiente na contagem esperada não pode ser examinada com base no seu sinal. Por exemplo, o valor +1.75 relativo à variável `cXemp` sugere que o número esperado de cesáreas em estabelecimentos empresariais tende a ser mais elevado, em contraste com os casos tais que `cXemp`=0. Mas com respeito às variáveis `ihac` e `LPT.4`, ambas possuem coeficientes β e α, com sinais invertidos. Para esses casos é preciso calcular a média ajustada para avaliar os efeitos proporcionados por essas variáveis, de acordo com o perfil do estabelecimento de saúde.

Tabela 6.8: Exemplos de modelos.

modelo	p	q	variáveis para λ	variáveis para δ
0	0	0	—	—
1	1	1	`normal`	`normal`
2	3	2	`ihac,LPT.4,lpib`	`ihac,LPT.4`
3	7	3	`normal,pub,ihac,rm,cXemp,LPT.4,lpib`	`normal,ihac,LPT.4`
4	6	4	`normal,sus,cXsus,cXemp,LPT.4,lpib`	`sus,emp,cXsus,lpib`
5	8	3	`sus, pub, emp, ihac, cXsus, cXemp, LPT.4, lpib`	`normal,ihac,LPT.4`
6	8	4	`sus, pub, emp, ihac, cXsus, cXemp, LPT.4, lpib`	`normal,ihac,rm,LPT.4`
7	8	5	`normal,sus,pub,rm,cXsus,cXemp,LPT.4,lpib`	`sus,pub,emp,cXsus,LPT.4`
8	16	16	`normal,sus,pub,emp,ihac,poor.rs,poor.me,poor.mi, rm,ride,cXsus,cXemp,LPT.4,lpop,lpib,idhm`	`normal,sus,pub,emp,ihac,poor.rs,poor.me,poor.mi, rm,ride,cXsus,cXemp,LPT.4,lpop,lpib,idhm`

Tabela 6.9: Resultados gerais. Estimativas obtidas com base na massa de dados para estimação ($n_1 = 726$), e estatísticas $\chi^2_{(182)}$ calculadas sobre a massa de teste ($n_2 = 182$).

modelo	p+q+2	AIC	BIC	l (log veros.)	pseudo-R^2	χ^2 (p-valor)
0	2	2532.997	2542.172	-1264.499	—	239.615 (0.003)
1	4	2385.926	2404.276	-1188.963	0.06	245.099 (0.001)
2	7	2155.129	2187.242	-1070.564	0.15	166.879 (0.782)
3	12	1780.555	1835.606	-878.278	0.31	180.790 (0.511)
4	12	1751.003	1806.054	-863.502	0.32	176.964 (0.591)
5	13	1746.612	1806.251	-860.306	0.32	182.693 (0.472)
6	14	1745.165	1809.391	-858.583	0.32	180.424 (0.519)
7	15	1746.227	1815.040	-858.113	0.32	173.568 (0.661)
8	34	1768.583	1924.560	-850.292	0.33	177.486 (0.581)

Por exemplo, em função do porte `0 < LPT.4 < 2.6`, suponha que se deseja comparar a distribuição das cesáreas registradas em um hospital público da cidade

Tabela 6.10: Estimativas dos coeficientes do Modelo 6.

	estimate	s.e.	ratio
beta_0	-4.32	0.67	-6.49
sus	0.72	0.17	4.17
pub	0.32	0.11	2.75
emp	-0.97	0.22	-4.31
ihac	-0.80	0.21	-3.82
cXsus	-0.72	0.16	-4.54
cXemp	1.75	0.23	7.49
LPT.4	1.84	0.14	13.02
lpib	9.69	3.78	2.56
alpha_0	3.47	0.66	5.29
normal	-2.85	0.41	-6.94
ihac	1.50	0.56	2.68
rm	-0.63	0.34	-1.86
LPT.4	-1.11	0.41	-2.70

de Goiânia com selo IHAC que atenda pelo SUS (`sus=1`, `pub=1`,`emp=0`, `ihac=1`, `cXsus=1`, `cXemp=0` e `lpib=0.136`), com a das cesáreas registradas em um hospital empresarial da mesma cidade, sem o selo IHAC e que não atenda pelo SUS (`sus=0`, `pub=0`,`emp=1`, `ihac=0`, `cXsus=0`, `cXemp=1` e `lpib=0.136`), e a distribuição dos partos cesáreas registrados em um hospital sem fins lucrativos da mesma cidade, com o selo IHAC e que atenda pelo SUS (`sus=1`, `pub=0`,`emp=0`, `ihac=1`, `cXsus=1`, `cXemp=0` e `lpib=0.136`). A comparação entre as contagens esperadas, de acordo com esses perfís, se encontra ilustrada na Figura 6.2.

De modo semelhante à Figura 6.1, as contagens esperadas tendem a crescer naturalmente em função do porte `LPT.4`. Contudo, o modelo ajustado destaca diferenças proeminentes entre a curva dos estabelecimentos empresariais e as demais.

E, finalmente, a Figura 6.3 mostra os resíduos de Pearson proporcionados pelo modelo 6. A forma assimétrica da distribuição de Touchard, naturalmente, reflete-se nessa figura. A soma dos quadrados desses valores constitui a estatística χ^2, cujo valor observado foi igual a 180.424 e p-valor igual a 0.519 obtido sob a distribuição qui-quadrado com 182 graus de liberdade. Assim, não rejeita-se a hipótese de que o modelo Touchard se ajusta bem aos dados.

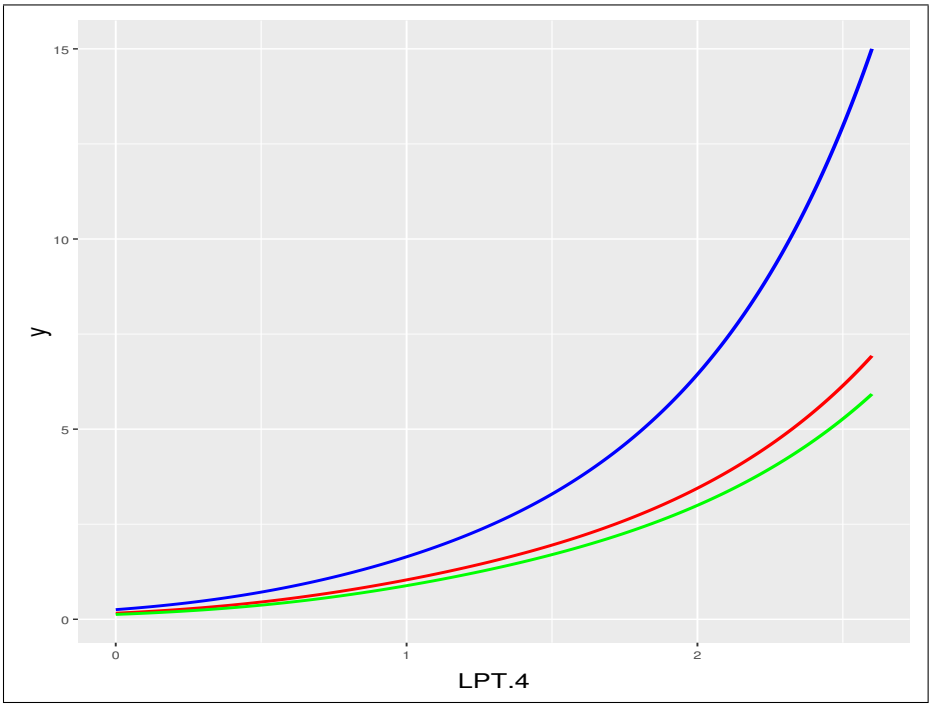

Figura 6.2: `LPT.4` *versus* `y`. As linhas sólidas representam as médias ajustadas de acordo com perfil do estabelecimento, em cor azul (hospital empresarial de Goiânia, sem o selo IHAC e que não atenda pelo SUS), cor vermelha (hospital público de Goiânia, com selo IHAC que atenda pelo SUS) e cor verde (entidades sem fins lucrativos de Goiânia, com selo IHAC e que atenda pelo SUS).

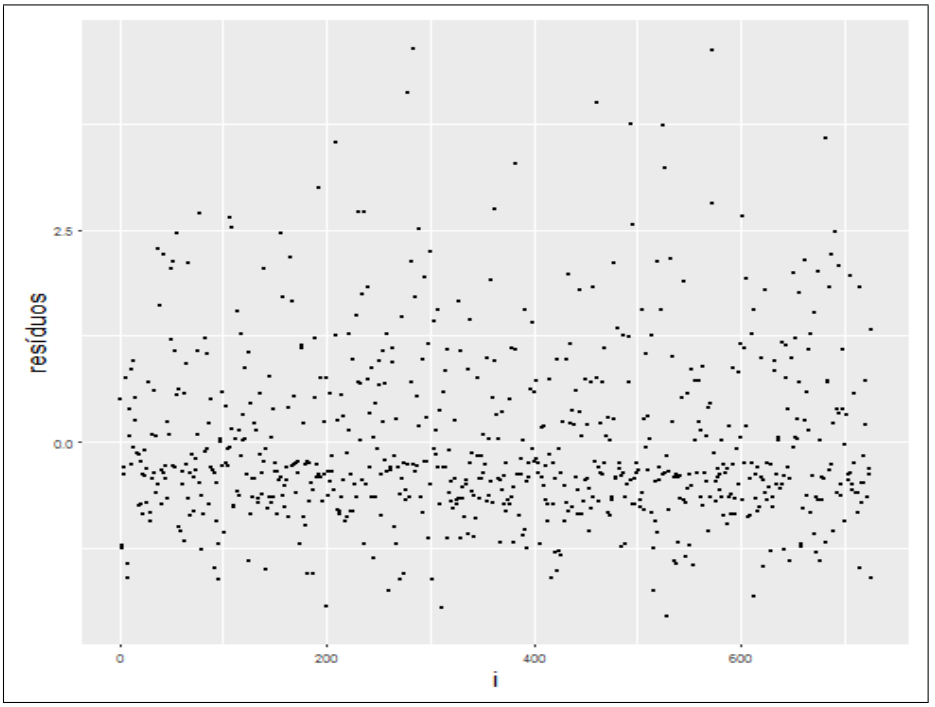

Figura 6.3: Resíduos de Pearson.

6.4 Comparação com outros modelos

As estimativas de máxima verossimilhança dos parâmetros dos modelos de regressão de Poisson, Binomial Negativa e COM-Poisson são apresentadas na Tabela 6.11. Os modelos apresentados obtiveram o melhor desempenho entre os modelos ajustados de cada categoria. Por exemplo, o modelo Poisson exibido na Tabela 6.11 corresponde ao modelo de Poisson que melhor se ajustou ao conjunto de dados e possui todos os coeficientes significativos. Segundo o critério de informação AIC, o modelo COM-Poisson apresenta um desempenho melhor que os modelos clássicos Poisson e Binomial Negativa.

O modelo Touchard e o COM-Poisson têm em comum o fato de serem extensões do modelo Poisson com dois parâmetros e ambos utilizaram variáveis para descrever o parâmetro de dispersão. No entanto, o modelo Touchard 6 com AIC=1745.165 (Tabela 6.9) proporciona um ajuste um pouco melhor que a regressão COM-Poisson.

Tabela 6.11: Estimativas de máxima verossimilhança dos coeficientes de regressão para o modelo Poisson, Binomial Negativa e COM-Poisson, e AIC.

coeficiente	Poisson	Bin. Negativa	COM-Poisson
beta_0	-1.54	-1.54	-2.08
sus	-0.21	-0.21	-0.72
pub	0.25	0.24	0.32
emp	-1.68	-1.68	-1.83
cXsus	0.17	0.20	0.88
cXemp	2.54	2.51	2.67
LPT.4	1.32	1.32	0.73
lpib			10.43
alpha_0			0.99
normal			-0.74
ihac			0.33
rm			0.19
LPT.4			-0.53
AIC	1802.03	1802.60	1760.11
BIC	1834.14	1834.72	1819.75

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.