Universidade de Brasília
Instituto de Ciências Exatas
Departamento de Estatística

Dissertação de Mestrado

# Regressão Binomial Negativa Inflacionada de Zeros Geograficamente Ponderada

**por**

**Marcos Douglas Rodrigues de Sousa**

Brasília, 21 de Setembro de 2022

# Regressão Binomial Negativa Inflacionada de Zeros Geograficamente Ponderada

**por**

**Marcos Douglas Rodrigues de Sousa**

Dissertação apresentada ao Departamento de Estatística da Universidade de Brasília, como requisito parcial para obtenção do título de Mestre em Estatística.

Orientador: Prof. Dr. Alan Ricardo da Silva

Brasília, 21 de Setembro de 2022

Dissertação submetida ao Programa de Pós-Graduação em Estatística do Departamento de Estatística da Universidade de Brasília como parte dos requisitos para a obtenção do título de Mestre em Estatística.

Texto aprovado por:

Prof. Dr. Alan Ricardo da Silva

Orientador, EST/UnB

Prof. Dr. André Luiz Fernandes Cançado

EST/UnB

Prof. Francisco José A. Cysneiros

EST/UFPE

*A persistência é o caminho do êxito.*

(Charles Chaplin)

# Agradecimentos

Agradeço a Deus por estar comigo em todo tempo.

Agradeço ao meu orientador, Prof. Alan Ricardo da Silva, pela paciência e disponibilidade, sem ele não seria possível a realização deste trabalho.

Agradeço aos meus pais e irmã por serem meu suporte e estarem ao meu lado.

Agradeço a todos os amigos e colegas de curso pelo companherismo durante esse período de mestrado.

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001.

# Resumo

O objetivo deste trabalho é trazer uma abordagem sobre a modelagem de dados de contagem, considerando a existência de zeros na distribuição. Pressupondo a utilização de dados espacias, em que o fenômeno em análise não apresente estacionariedade, a regressão geograficamente ponderada surge para solucionar este problema. Sendo assim, este trabalho traz uma extensão da regressão binomial negativa geograficamente ponderada (RBNGP) para incluir a distribuição binomial negativa inflacionada de zeros, sendo intitulada regressão binomial negativa inflacionada de zeros geograficamente ponderada (RBNIZGP).

Para verificar a performance de ajuste do modelo RBNIZGP, foram utilizados alguns dados simulados de distribuições Poisson, binomial negativa, Poisson inflacionado de zeros e binomial negativa inflacionada de zeros, sem variação espacial. E por último, para verificação da qualidade do ajuste no caso de variação espacial, foram utilizados dados reais sobre casos de COVID-19 na Coréia do Sul, sendo dados que foram analisados por (Weinstein et al., 2021).

Os resultados das simulações mostraram que o modelo RBNIZGP foi capaz de modelar os dados com distribuição Poisson, binomial negativa, Poisson inflacionada de zeros e binomial negativa inflacionada de zeros, sem variação espacial, por meio de uma grande parâmetro de suavização. Já no estudo de caso real, os resultados mostraram que localmente, os modelos ajustados poderiam ser Poisson ou binomial negativo, refinando dessa forma a análise, e mostrando a flexibilidade do modelo RBNIZGP.

**Palavras-Chave:** Dados espaciais; Dados de contagem; Não estacionariedade; Regressão geograficamente ponderada; Regressão binomial negativa inflacionada de zeros

# Abstract

The goal of this work is to bring an approach to the modeling of count data, considering the existence of zeros in the distribution. Assuming the use of spatial data, in which the phenomenon under analysis does not present stationarity, the geographically weighted regression appears to solve this problem. Therefore, this work brings an extension of the geographically weighted negative binomial regression (GWNBR) to include a zero-inflated negative binomial distribution, entitled geographically weighted zero-inflated negative binomial regression (GWZINBR).

To verify the performance of the fit of the RBNIZGP model, some simulated data from distributions, zero-inflated poisson and zero-inflated negative binomial, without spatial space, were used. Finally, adjustment was used in the case of selection of the real quality of data on COVID-19 cases in South Korea, with data from South Korea being analyzed by (Weinstein et al., 2021).

The results of the simulations showed that the RBNIZGP model was able to model the data with Poisson, negative binomial, zero inflated Poisson and zero inflated negative binomial distributions, without spatial variation, by means of a large bandwidth. In the real case study, the results showed that locally, the adjusted models could be Poisson or negative binomial, thus refining the analysis, and showing the flexibility of the GWZINBR model.

**Keywords:** Spatial data; Count data; Non-stationarity; Geographically weighted regression; Zero inflated negative binomial regression

# Sumário

# Lista de Tabelas

# Lista de Figuras

# Capítulo 1

# Introdução

Modelos globais de regressão espacial com coeficientes únicos para toda a área de estudo são utéis para descrever a relação observada de dependência espacial, considerando que o fenômeno em estudo apresente estacionariedade. Um processo espacial é definido como estacionário se sua distribuição de probabilidade possui invariância no espaço. Entretanto, quando o fenômemo em análise não apresenta estacionariedade, o modelo global de regressão espacial pode não ser o melhor modelo para explicar o fenômeno estudado. Sendo assim, surge a necessidade da estimação de modelos locais de regressão espacial (Brunsdon et al., 2000).

A regressão geograficamente ponderada - RGP (ou do inglês, *Geographically Weighted Regression - GWR*), proposta por Brunsdon et al. (1996) surge como uma alternativa para o problema da não estacionariedade em dados espaciais, sendo uma extensão do modelo de regressão linear tradicional, permitindo que ocorram variações locais nos parâmetros. Portanto, a ideia da RGP é realizar um ajuste de modelo de regressão para cada ponto no conjunto de dados, ponderando as observações por uma função de distância a esse ponto.

Na RGP, existe o pressuposto de que a variável resposta $y$ deve seguir uma distribuição normal, assim como a distribuição dos erros $\varepsilon_j$. Entretanto, na prática, existem muitas situações em que a variável respostal $y$ não segue uma distribuição normal, como por exemplo em casos de dados discretos com contagens de observações. Para estes casos, as distribuições mais utilizadas são a Poisson e a binomial negativa.

O modelo de regressão de Poisson é frequentemente utilizado para a análise de dados de contagem e se baseia nos pressupostos inerentes ao processo e à distribuição de Poisson. Assim como o modelo de Poisson, a regressão binomial negativa também é utilizada para modelar dados de contagem. Entretanto, quando os dados apresentam um número excessivo de zeros, as distribuições Poisson inflacionada de zeros (PIZ) (ou do inglês, *Zero Inflated Poisson* - ZIP) (Lambert, 1992) e binomial negativa inflacionada de zeros (BNIZ) (ou inglês, *Zero Inflated Negative Binomial* - ZINB) podem ser boas alternativas, já que foram propostas justamente para modelar dados de contagem com excesso de observações iguais a zero (Lawless, 1987).

Os modelos globais resumem os dados para toda a região de estudo e também geram estatísticas que não podem ser analisadas de acordo com o sistema geográfico de informação, diferentemente dos modelos locais que são analisados considerando o sistema geográfico de informação (Brunsdon et al., 2000). Ao modelar dados de contagem, com quantidade excessiva de zeros, a distribuição de probabilidade do modelo global pode-se considerar uma BNIZ. Entretanto, quando há um foco nas diferenças no espaço, ou seja, nos modelos locais, essa distribuição pode não vir a ser uma BNIZ.

Para ilustrar essa ideia, a Figura 1.1 mostra a distribuição dos casos de COVID-19 na Coréia do Sul antes da quarentena (Weinstein et al., 2021), onde pode-se observar uma quantidade excessiva de zeros, caracterizando uma possível distribuição binomial negativa inflacionada de zeros.

A Figura 1.2 mostra a distribuição de frequência do número de casos de Covid-19 antes da quarentena na Coréia do Sul, onde nota-se uma concentração muito grande de observações iguais a zero, mostrando uma distribuição altamente assimétrica à direita. No entanto, ao observar as distribuições de frequências locais de regiões específicas (Figura 1.4), considerando uma distância euclidiana de 50 Km, verifica-se que localmente as distribuições são distintas, sendo que em alguns locais, a distribuição pode ser uma Poisson inflacionada de zeros, Poisson ou ainda binomial negativa.

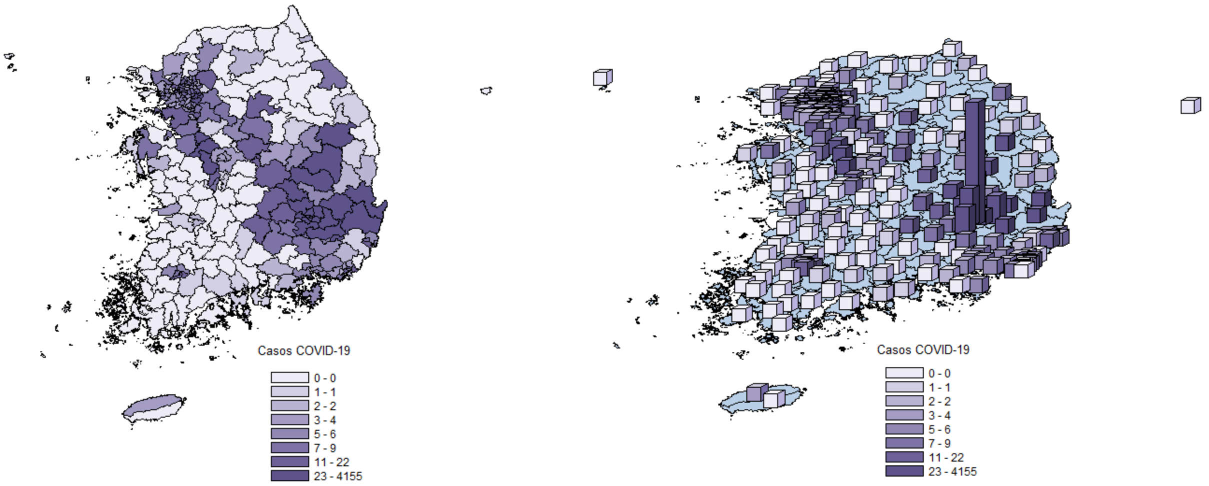



Figura 1.1: Distribuição espacial dos casos de COVID-19 antes da quarentena na Coréia do Sul
Fonte: Weinstein et al. (2021).

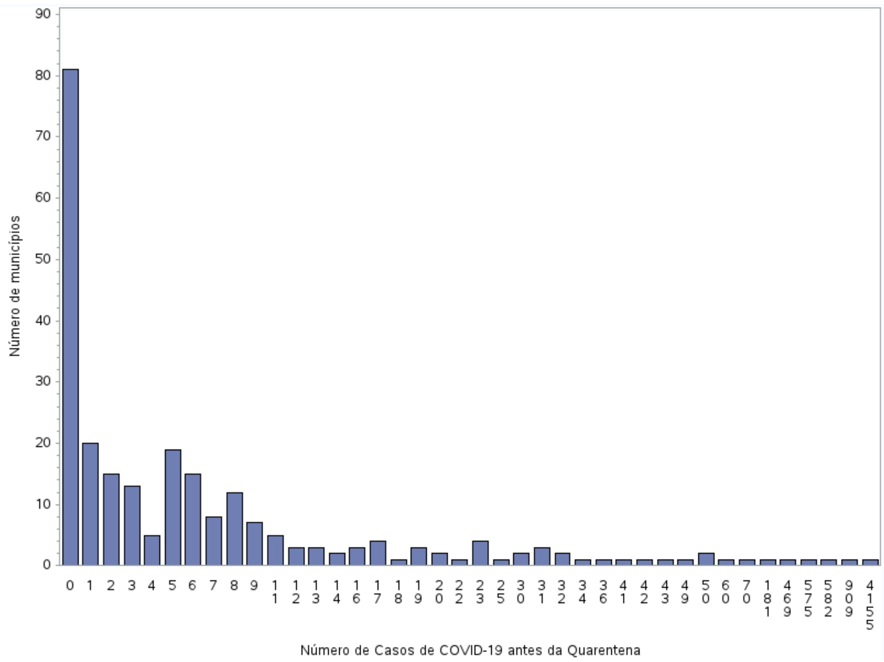



Figura 1.2: Distribuição de frequências do número de casos diários de Covid-19 antes da quarentena na Coréia do Sul

Na Figura 1.3 estão representados os mapas de algumas regiões que geraram a Figura 1.4, considerando um raio de 50km. A RGP calcula as estimativas locais considerando as observações ponderadas dentro desse raio. Para cada região, pode-se observar uma diferença no número

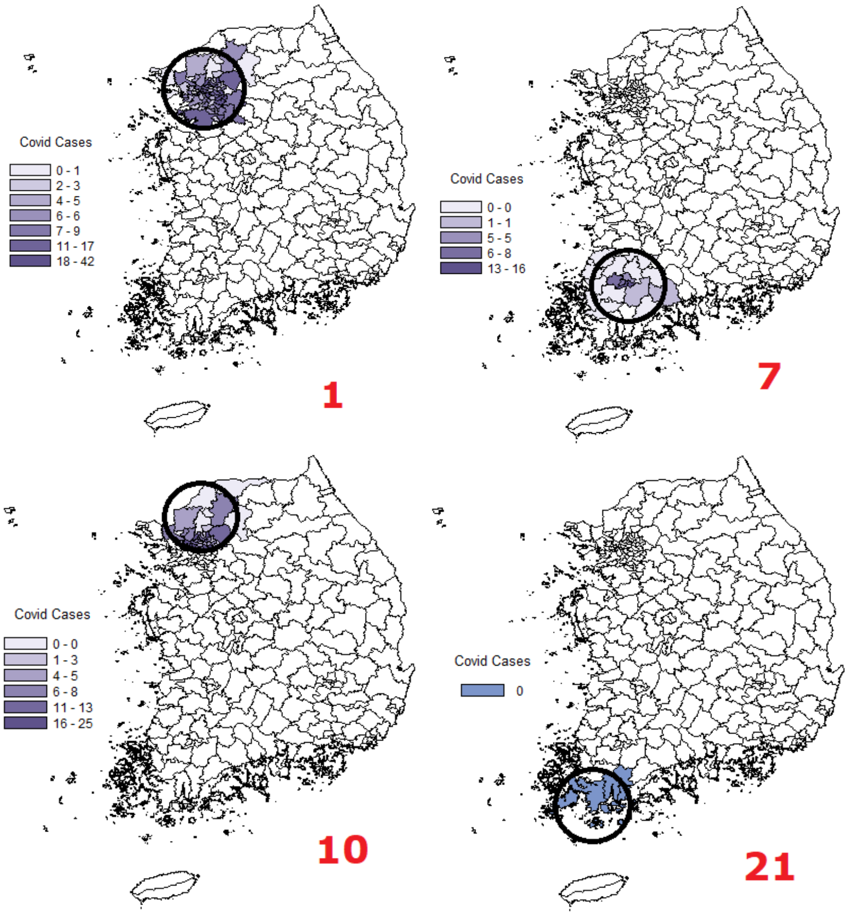



Figura 1.3: Mapa da Coréia do Sul - Algumas regiões (raio de 50km)

de casos de COVID-19, sendo que a frequência de casos está entre 0 e 42 na região 1, entre 0 e 16 casos na região 7, entre 0 e 25 casos na região 10 e por fim na região 21, observa-se que o número de casos de COVID-19 para esta região é totalmente nulo. Com referência às distribuições de frequências da Figura 1.4, nota-se que nas regiões 1, 7 e 10, os dados se assemelham à distribuições Poisson, Poisson inflacionada de zeros e binomial negativa, respectivamente.

Na Figura 1.5, que representa novamente o mapa da Coréia do Sul, são ilustrados os raios de 150km e 600km. Já nas Figuras 1.6 e 1.7, onde foram consideradas as distâncias euclidianas de 150 Km e 600 Km, respectivamente, pode-se notar uma quantidade significativa de casos de COVID-19 iguais a zero, ou seja, com o aumento da distância, a distribuição volta a ter uma tendência de distribuição binomial negativa inflacionada de zeros, ou ainda Poisson inflacionada

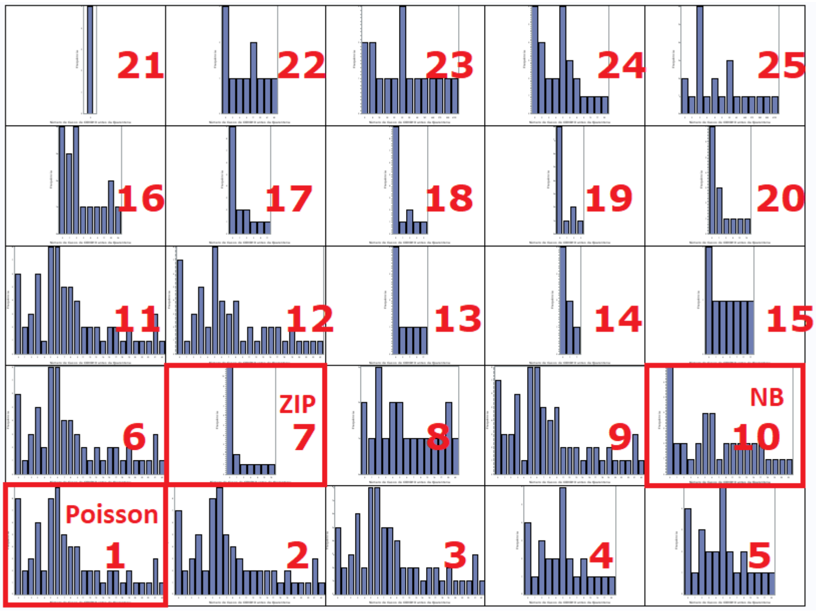



Figura 1.4: Distribuição de frequências do número de casos de Covid-19 antes da quarentena em regiões específicas da Coréia do Sul (raio de 50 KM)

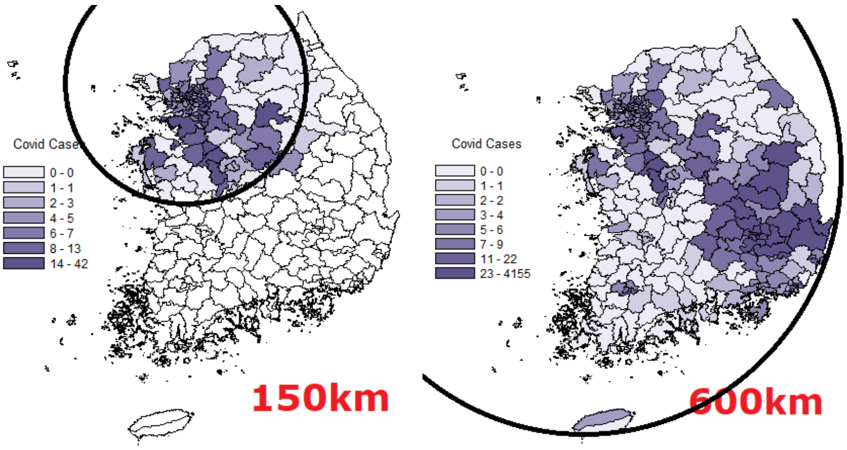



Figura 1.5: Mapa da Coréia do Sul (raios de 150km e 600km)

de zeros e binomial negativa em alguns casos.

Dessa forma, a Figura 1.8 apresenta um fluxograma com o relacionamento entre as distribuições binomial negativa inflacionada de zeros, Poisson inflacionada de zeros, binomial negativa e Poisson. Ou seja, quando os parâmetros inflacionados do modelo binomial negativo inflacionado de zeros são não significativos (ou se de fato são todos iguais a zero), então a distribuição

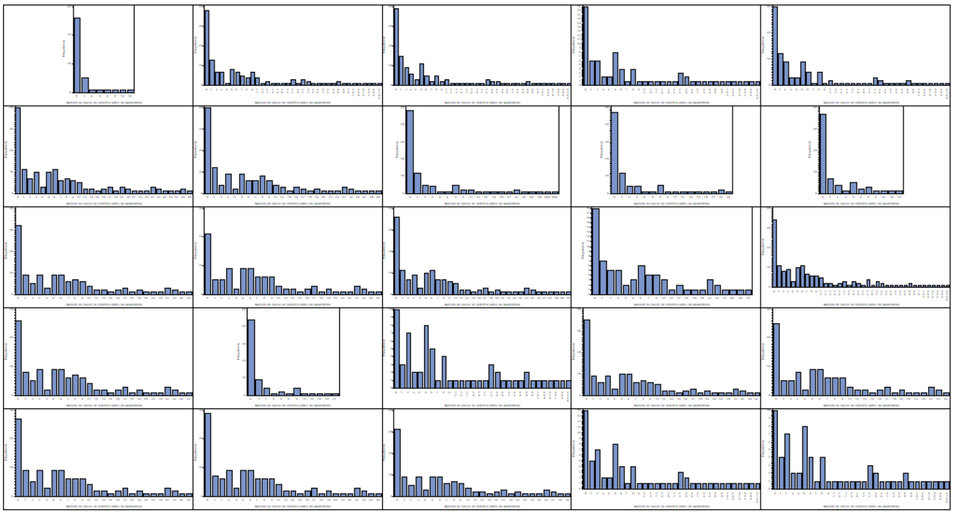


Figura 1.6: Distribuição de frequências do número de casos de Covid-19 antes da quarentena em regiões específicas da Coréia do Sul (raio de 150 KM)

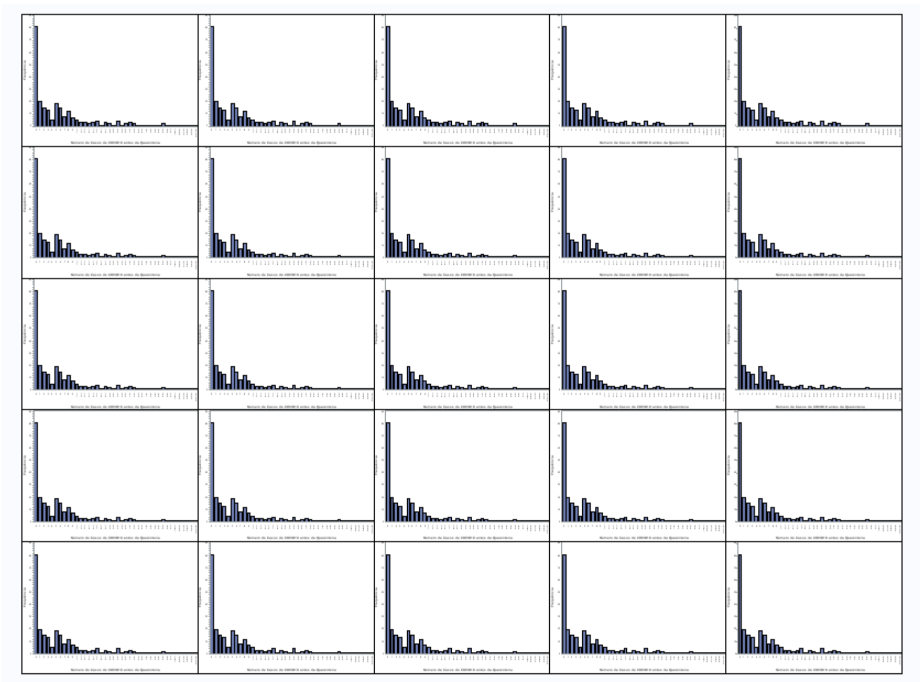


Figura 1.7: Distribuição de frequências do número de casos de Covid-19 antes da quarentena em regiões específicas da Coréia do Sul (raio de 600 KM)

poderá ser binomial negativa. Já em relação ao parâmetro de superdispersão da distribuição binomial negativa inflacionada de zeros, se este não for considerado significativo (ou se de fato ele for igual a zero), então a distribuição será a Poisson inflacionada de zeros e se ainda os parâmetros inflacionados da distribuição Poisson inflacionada de zeros forem não significativos (ou se de fato forem todos iguais a zero), então a distribuição será a Poisson. Por fim, considerando o caso da binomial negativa, se o parâmetro de superdispersão também não for considerado significativo (ou se de fato for igual a zero), a distribuição também será uma Poisson. Ou seja, tem-se como caso geral o modelo binomial negativo inflacionado de zeros e como caso mais simples o modelo de Poisson.

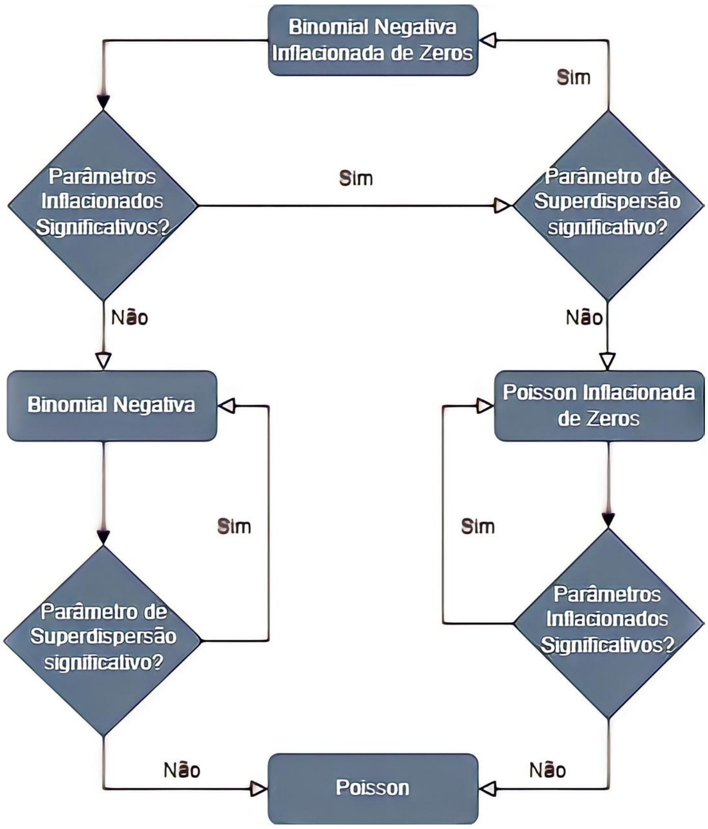



Figura 1.8: Relação entre os modelos binomial negativo inflacionado de zeros, Poisson inflacionado de zeros, binomial negativo e Poisson

A estrutura apresentada na Figura 1.8 será demonstrada teoricamente nos capítulos 2 e 3. No capítulo 3 será demonstrada a ideia de significância dos parâmetros de superdispersão significativos para os modelos binomial negativo inflacionado de zeros e Poisson inflacionado de zeros. Naya et al. (2008) propuseram uma estrutura parecida com a apresentada na Figura 1.8, no entanto, se basearam somente na distribuição Poisson inflacionada de zeros.

Com base na ideia dos trabalhos de Atkinson et al. (2003), Nakaya et al. (2005),Da Silva e Rodrigues (2014), que desenvolveram os modelos de RGP com as distribuições binomial, Poisson e binomial negativa, respectivamente, o objetivo deste trabalho é propor uma extensão da regressão binomial negativa geograficamente ponderada (RBNGP) para incluir a distribuição binomial negativa inflacionada de zeros, aqui denominada regressão binomial negativa inflacionada de zeros geograficamente ponderada (RBNIZGP), a qual incluirá a regressão Poisson inflacionada de zeros geograficamente ponderada (RPIZGP), a regressão binomial negativa geograficamente ponderada (RBNGP) e a regressão Poisson geograficamente ponderada (RPGP). Ou seja, ela pode ser considerada um modelo geral para dados de contagem espacialmente dependentes.

A ideia da regressão geograficamente ponderada para o modelo Poisson inflacionado de zeros foi desenvolvida anteriormente, nos trabalhos de Kalagirou (2016), Purhadi et al. (2015) e Purhadi et al. (2021). A partir das condicionais apresentadas na Figura 1.8, o modelo binomial negativo inflacionado de zeros poderá ser reduzido localmente para o modelo Poisson inflacionado de zeros, binomial negativo ou Poisson.

Este trabalho está organizado da seguinte forma: No capítulo 2 será apresentada uma revisão bibliográfica sobre os modelos lineares generalizados. No capítulo 3 serão apresentados os modelos inflacionados de zeros, com destaque para os modelos binomial negativo inflacionado de zeros e Poisson inflacionado de zeros. O capítulo 4 mostrará o modelo de regressão geograficamente ponderado e o capítulo 5 mostrará a modelagem da regressão geograficamente ponderada com a inclusão do modelo binomial negativo inflacionado de zeros. O capítulo 6 apresentará os materias e métodos a serem utilizados no trabalho para ilustrar a flexibilidade da RBNIZGP e o capítulo 7 apresenta a análise dos resultados. Por fim, no capítulo 8 serão apresentadas as conclusões, limitações do trabalho e recomendações para trabalhos futuros.

# Capítulo 2

# Modelos Lineares Generalizados

## 2.1 Introdução

O modelo linear generalizado - MLG é uma extensão do modelo linear clássico. Ele permite analisar a relação entre um conjunto de variáveis independentes $X_1, ..., X_n$ e a variável dependente $Y$. A variável resposta $Y$ deve seguir uma distribuição pertencente à família exponencial. Sendo assim, o objetivo deste Capítulo é fazer uma pequena introdução do MLG, detalhando os modelos binomial, Poisson e binomial negativo.

## 2.2 Família exponencial

Uma variável aleatória $Y$ pode ter sua distribuição pertencente à família exponencial uniparamétrica se sua função (densidade) de probabilidade for expressa da seguinte forma (Nelder e Wedderburn, 1972):

$$f(y;\theta) = h(y)\exp\left(\eta(\theta)t(y) - b(\theta)\right) \qquad (2.1)$$

sendo que $\theta$ é um parâmetro escalar e as funções $h(y)$, $t(y)$, $b(\theta)$ e $\eta(\theta)$ possuem valores reais conhecidos.

No caso da família exponencial multiparamétrica (que é uma generalização da uniparamétrica), a sua função(densidade) de probabilidade se caracteriza da seguinte forma:

$$f(y;\theta) = h(y)\exp\left(\sum_{i=1}^{k}\eta_i(\theta)t_i(y) - b(\theta)\right) \tag{2.2}$$

sendo que $\theta$ é um vetor de parâmetros e que $h(y)$ e $t_1(y), ..., t_k(y)$ são funções com valores reais, de observação de $y$ não dependendo de $\theta$ e $b(\theta)$ e $\eta_1(\theta), ..., \eta_k(\theta)$ são funções com valores reais do parâmetro que possivelmente tem seu valor definido pelo vetor de $\theta$.

Muitas famílias podem ser apresentadas como exponenciais conforme (2.1), como por exemplo as distribuições binomial, Poisson, exponencial e geométrica. Já outras famílias podem ser apresentadas como exponenciais conforme (2.2), como por exemplo, as distribuições beta, gama, Weibull e Gaussiana.

A partir de (2.1) pode-se definir a forma canônica da família exponencial, quando $t(y)$ e $\eta(\theta)$ são funções do tipo identidade, sendo assim, a expressão é denotada da seguinte forma:

$$f(y;\theta) = h(y)\exp\left(\theta y - b(\theta)\right) \tag{2.3}$$

O parâmetro $\phi > 0$, associado à dispersão da distribuição, foi introduzido por Nelder e Wedderburn (1972) na forma canônica da família exponencial uniparamétrica:

$$f(y;\theta;\phi) = \exp\left(\frac{\theta y - b(\theta)}{\phi} + c(y,\phi)\right) \tag{2.4}$$

onde $\theta$ e $\phi$ são parâmetros escalares e $b(.)$ e $c(.)$ são funções reais conhecidas.

Com essa introdução de $\phi$, algumas das distribuições biparamétricas pertencentes à família exponencial são contempladas. Portanto, a família (5.1) engloba distribuições contínuas e discretas e possui uma abrangência maior do que a família (2.1).

## 2.3 Modelo linear generalizado

Os Modelos Lineares Generalizados - MLG foram propostos por Nelder e Wedderburn (1972) e a sua classe é definida pelos seguintes componentes:

$a)$ Componente aleatório: Variáveis aleatórias independentes pertencentes à família exponencial (5.1).

A esperança da variável aleatória $Y$ é dada da forma:

$$E(Y_i) = \mu_i = b'(\theta_i) \tag{2.5}$$

E a variância da variável aleatória $Y$ é dada da forma:

$$Var(Y_i) = \phi b''(\theta_i) = \phi V_i \tag{2.6}$$

$b)$ Componente sistemático: Variáveis explicativas na forma de uma soma linear de seus efeitos:

$$\eta_i = \sum_{r=1}^{p} x_{ir}\beta_r = \boldsymbol{X\beta} \tag{2.7}$$

onde $\boldsymbol{X} = (x_1, ..., x_n)^T$ representa a matriz do modelo ; $\boldsymbol{\beta} = (\beta_1, ..., \beta_n)$ representa o vetor de parâmetros desconhecidos; $\boldsymbol{\eta} = (\eta_1, ..., \eta_n)^T$ representa o preditor linear.

$c)$ Função de ligação: Função que associa, de forma adequada, a média da variável resposta $(\mu_i)$ à um preditor linear.

$$\eta_i = g(\mu_i) = \boldsymbol{X\beta} \tag{2.8}$$

onde a função de ligação $g(.)$ deve ser inversível e duplamente diferenciável.

Caso a função de ligação seja escolhida de tal forma que $g(\mu_i) = \theta_i = \eta_i$, o preditor linear modela diretamente o parâmetro canônico $\theta_i$, sendo denominada função de ligação canônica (Nelder e Wedderburn, 1972).

## 2.4 Algoritmos de estimação

Nesta sessão serão detalhados os algoritmos utilizados na estimação de parâmetros, sendo o Newton-Raphson e o Método Escore de Fisher.

### 2.4.1 Newton-Raphson

O método de máxima verossimilhança (MV) é muito utilizado na estimação dos parâmetros $\beta_1, ..., \beta_p$ de um modelo, sendo que os estimadores a serem obtidos possuem consistência e eficiência assintótica. O vetor escore de dimensão $p$ é formado pelas derivadas parciais de primeira ordem do logaritmo da função de verossimilhança e sua expressão é dada da forma (Conte, 1965) :

$$\mathbf{U}(\boldsymbol{\beta}) = \frac{\partial l(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} \tag{2.9}$$

onde $l(\boldsymbol{\beta})$ representa o logaritmo da função de verossimilhança.

A estimativa de máxima verossimilhança do vetor de parâmetros $\boldsymbol{\beta}$ é obtida igualando-se $\mathbf{U}(\boldsymbol{\beta}) = 0$. Quando não é possível encontrar uma solução para a estimação, utiliza-se um método iterativo tal como o de Newton-Raphson que é baseado na aproximação de Taylor para a função $f(x)$ na vizinhança do ponto $x_0$ :

$$f(x) = f(x_0) + (x - x_0)f'(x_0) = 0 \tag{2.10}$$

De forma mais geral obtêm-se:

$$x^{(m+1)} = x^{(m)} - \frac{f(x^{(m)})}{f'(x^{(m)})} \tag{2.11}$$

onde $x^{(m+1)}$ representa o valor de $x$ no passo $(m+1)$; $x^{(m)}$ representa o valor de $x$ no passo $(m)$; $f(x^{(m)})$ representa a função $f(x)$ avaliada em $x^{(m)}$ e $f'(x^{(m)})$ é a derivada da função $f(x)$ avaliada em $x^{(m)}$.

A versão do método de Newton-Raphson multivariado é dada (Nelder e Wedderburn, 1972):

$$\boldsymbol{\beta}^{(m+1)} = \boldsymbol{\beta}^{(m)} + (\boldsymbol{J}^{(m)})^{-1}\boldsymbol{U}^{(m)} \tag{2.12}$$

onde $\boldsymbol{\beta}^{(m)}$ e $\boldsymbol{\beta}^{(m+1)}$ são os vetores de parâmetros estimados nos passos $m$ e $(m+1)$, respectivamente; $\boldsymbol{U}^{(m)}$ representa o escore avaliado no passo $m$ e $(\boldsymbol{J}^{(m)})^{-1}$ representa a inversa da negativa da matriz de derivadas parciais de segunda ordem do $l(\boldsymbol{\beta})$, avaliada no passo $m$.

De forma resumida, o algoritmo de Newton-Raphson é exibido no Algoritmo 1 . Pode-se perceber que os valores iniciais devem ser fornecidos para o vetor de parâmetros.

**Algoritmo 1:** Newton-Raphson

**Entrada:** $\boldsymbol{\beta_0}, \boldsymbol{\beta_n}$

**enquanto** $(abs(\boldsymbol{\beta_n} - \boldsymbol{\beta_0}) > 10^{-6})$ **faça**

  $\boldsymbol{U} = \partial l(\boldsymbol{\beta})/\partial \boldsymbol{\beta}$

  $\boldsymbol{J} = \partial^2 l/\partial \boldsymbol{\beta}^2$

  $\boldsymbol{\beta_0} = \boldsymbol{\beta_n}$

  $\boldsymbol{\beta_n} = \boldsymbol{\beta_0} - \boldsymbol{J}^{-1}\boldsymbol{U}$

**fim**

### 2.4.2 Método escore de Fisher

Existe um outro método que é utilizado quando as derivadas parciais de segunda ordem não são avaliadas de forma simples e fácil. Este método é o escore de Fisher, que é mais eficiente para essas situações. Ele envolve a substituição da matriz de derivadas parciais de segunda ordem pela matriz de valores esperados das derivadas parciais, ou seja, ocorre a substituição da matriz $\boldsymbol{J}$, pela matriz de informação esperada de Fisher $\boldsymbol{I}$ (Nelder e Wedderburn, 1972). Portanto,

$$\boldsymbol{\beta}^{(m+1)} = \boldsymbol{\beta}^{(m)} + (\boldsymbol{I}^{(m)})^{-1}\boldsymbol{U}^{(m)} \tag{2.13}$$

onde $\boldsymbol{I}$ possui a expressão:

$$[\boldsymbol{I}(\boldsymbol{\theta})]^{-1} = -E\left[\frac{\partial^2 l(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'}\right]^{-1} \tag{2.14}$$

No caso de um modelo de regressão, $\boldsymbol{I}$ passa a ter a expressão:

$$[\boldsymbol{I}(\boldsymbol{\beta})]^{-1} = -E\left[\frac{\partial^2 l(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}'}\right]^{-1} = \phi^{-1}\boldsymbol{X^T A X} \tag{2.15}$$

onde $\boldsymbol{A}$, é uma matriz diagonal de pesos que captura a informação sobre a distribuição e a função de ligação. Portanto, sua expressão é dada por:

$$a_i = \frac{1}{V(\mu_i)}\left(\frac{\partial\mu_i}{\partial\eta_i}\right)^2 \tag{2.16}$$

Considerando as funções de ligação canônicas, tem-se $a_i = V_i$, pois $V_i = V(\mu_i) = \frac{\partial\mu_i}{\partial\eta_i}$. Nota-se também que a informação é inversamente proporcional ao parâmetro de dispersão. Com isso, o vetor escore $\boldsymbol{U}(\boldsymbol{\beta})$ pode ser reescrito da forma (Nelder e Wedderburn, 1972):

$$\boldsymbol{U}(\boldsymbol{\beta}) = \phi^{-1}\boldsymbol{X^T W \Delta}(\boldsymbol{y}-\boldsymbol{\mu}) \tag{2.17}$$

onde $\boldsymbol{\Delta} = diag(\partial\eta_1/\partial\mu_1, ..., \partial\eta_n/\partial\mu_n) = diag(g'(\mu_1), ..., g'(\mu_n))$. Portanto, a matriz diagonal $\boldsymbol{\Delta}$ é formada pelas derivadas de primeira ordem da função de ligação. Ao substituir $\boldsymbol{I}$ e $\boldsymbol{U}$ em (2.13) e eliminando $\phi$, tem-se (Nelder e Wedderburn, 1972):

$$\boldsymbol{X^T A}^{(m)}\boldsymbol{X\beta}^{(m+1)} = \boldsymbol{X^T A}^{(m)}\boldsymbol{X\beta}^{(m)} + \boldsymbol{X^T A}^{(m)}\boldsymbol{\Delta}^{(m)}(\boldsymbol{y}-\boldsymbol{\mu}^{(m)}) \tag{2.18}$$

ou, ainda,

$$\boldsymbol{X^T A}^{(m)}\boldsymbol{X\beta}^{(m+1)} = \boldsymbol{X^T A}^{(m)}[\boldsymbol{\eta}^{(m)} + \boldsymbol{\Delta}^{(m)}(\boldsymbol{y}-\boldsymbol{\mu}^{(m)})] \tag{2.19}$$

Por definição, a variável dependente é dada por (Nelder e Wedderburn, 1972):

$$\boldsymbol{z} = \boldsymbol{\eta} + \boldsymbol{\Delta}(\boldsymbol{y}-\boldsymbol{\mu}) \tag{2.20}$$

Portanto:

$$\boldsymbol{\beta}^{(m+1)} = [\boldsymbol{X^T A}^{(m)}\boldsymbol{X}]^{-1}\boldsymbol{X^T A}^{(m)}\boldsymbol{z}^{(m)} \tag{2.21}$$

A Equação (2.21) é válida para qualquer MLG, mostrando que a solução das equações por máxima verossimilhança é equivalente ao cálculo repetido de uma regressão linear ponderada de uma variável dependente ajustada $\boldsymbol{z}$ sobre a matriz $\boldsymbol{X}$ usando uma matriz de pesos $\boldsymbol{A}$ que se decodifica nas iterações (Nelder e Wedderburn, 1972).

A matriz de covariância de $\hat{\boldsymbol{\beta}}$, considerando $n \to \infty$, é dada pelo inverso da matriz de informação de Fisher $\boldsymbol{I}$ (2.15), ou seja:

$$\widehat{Cov}(\hat{\boldsymbol{\beta}}) = \phi[\boldsymbol{X^T}\hat{\boldsymbol{A}}\boldsymbol{X}]^{-1} \tag{2.22}$$

onde $\hat{\boldsymbol{A}}$ é a matriz de pesos $\boldsymbol{A}$ avaliada em $\hat{\boldsymbol{\beta}}$ .

A qualidade do ajuste de um MLG é avaliada através da função *Deviance*, e o critério de parada do algoritmo é baseado nessa função *Deviance*, proposta por Nelder e Wedderburn (1972).

$$D = 2\sum_{i=1}^{n}[L(y_i; y_i) - L(\mu_i; y_i)] \tag{2.23}$$

A Equação (2.23) é uma distância entre o logaritmo da função verossimilhança do modelo saturado (com $n$ parâmetros) e do modelo sobre investigação (com $p$ parâmetros) avaliado na estimativa de máxima verossimilhança $\hat{\boldsymbol{\beta}}$. Sendo que quanto menor o valor da função *Deviance*, melhor será o ajuste.

## 2.5 Casos específicos

Nesta seção serão apresentados alguns modelos específicos. Os modelos binomial, Poisson e binomial negativo serão detalhados a seguir.

### 2.5.1 Regressão binomial

O modelo binomial é usado no estudo de dados na forma de proporções ou dados binários. Suponha que $Y = m\pi$ tenha distribuição binomial $B(m, \pi)$, onde $\pi$ é definido como a proporção de sucessos em $m$ ensaios Bernoulli independentes, com probabilidade de sucesso $\pi$ (Dobson e Barnett, 2008).

O modelo é dado pela seguinte função massa de probabilidade:

$$P(Y \geq y) = \sum_{i=y}^{m} \binom{m}{i} \pi^i (1-\pi)^{m-i} \tag{2.24}$$

A probabilidade de se obter exatamente $y$ sucessos é dada pela função de probabilidade :

$$f(y; m, \pi) = \binom{m}{y} \pi^y (1-\pi)^{m-y} \tag{2.25}$$

Ao reescrever (2.25) em termos de família exponencial, conforme (5.1), a expressão resulta na seguinte forma (Dobson e Barnett, 2008):

$$f(y; m, \pi) = \exp\left( y \log\left(\frac{\pi}{1-\pi}\right) + m\log(1-\pi) + \log\binom{m}{y} \right) \tag{2.26}$$

Com base em Nelder e Wedderburn (1972), os parâmetros da distribuição binomial estão especificados na Tabela 2.1.

Os componentes do MLG da binomial são:

a) O componente aleatório: $Y_i \sim$ binomial $(m_i, \pi_i)$

b) O componente sistemático: $\boldsymbol{\eta} = \boldsymbol{X}\boldsymbol{\beta}$, onde $\boldsymbol{X}$ representa a matriz do modelo e $\boldsymbol{\eta}$ o vetor de parâmetros.

c) A função de ligação canônica: No caso da binomial, pode-se utilizar várias funções de ligação. Na Tabela 2.2, observa-se as funções de ligação Logit, Probit e Log-log.

Tabela 2.1: Parâmetros da distribuição binomial

| $\phi$ | $\theta$ | $b(\theta)$ | $\boldsymbol{\mu}$ | $V(\boldsymbol{\mu})$ | $c(y)$ |
|---|---|---|---|---|---|
| 1 | $\log\left(\frac{\mu}{1-\mu}\right)$ | $m\log(1+e^{\theta})$ | $\frac{me^{\theta}}{1+e^{\theta}}$ | $\frac{\mu}{m}(m-\mu)$ | $\log\binom{m}{y}$ |

Tabela 2.2: Funções de Ligação - Binomial

| **Logit** | **Probit** | **Log-log** |
|---|---|---|
| $\log\left(\frac{\pi}{1-\pi}\right) = \frac{e^{\boldsymbol{X_i\beta}}}{e^{\boldsymbol{X_i\beta}}+1}$ | $\phi^{-1}(\pi_i) = \phi(\boldsymbol{X_i\beta})$ | $\log(-\log(1-\pi_i)) = 1 - e^{-e^{\boldsymbol{X_i\beta}}}$ |

Neste trabalho, a função de ligação canônica a ser utilizada será a Logit, sendo:

$$g(\mu) = \theta = \log\left(\frac{\pi}{1-\pi}\right) = \log\left(\frac{\mu}{1-\mu}\right) \tag{2.27}$$

A log-verossimilhança do modelo de regressão binomial é dada por:

$$L(\pi, \boldsymbol{y}) = \sum_{i=1}^{n}\left[y_i\log(\pi_i) - (-(1-y_i)\log(1-\pi_i))\right] \tag{2.28}$$

Ao reescrever (2.28), considerando a parametrização pela função de ligação canônica, tem-se:

$$L(\boldsymbol{\beta}, \boldsymbol{y}) = \sum_{i=1}^{n} y_i(\boldsymbol{X_i\beta}) - \sum_{i=1}^{n}\log(1+e^{\boldsymbol{X_i\beta}}) \tag{2.29}$$

No caso binomial, a função *Deviance* é dada por:

$$D(\boldsymbol{y}, \hat{\boldsymbol{\mu}}) = 2\sum_{i=1}^{n}\left[y_i\log\left(\frac{y_i}{\hat{\mu_i}}\right) + (1-y_i)\log\left(\frac{1-y_i}{1-\hat{\mu_i}}\right)\right] \tag{2.30}$$

O Algoritmo 2 mostra o algoritmo de estimação dos parâmetros da regressão binomial.

**Algoritmo 2:** Regressão binomial

**Entrada:** $D_0 = 0, \quad difD = 1, \quad itr = 1$

$\boldsymbol{\mu} = (\boldsymbol{y} + \overline{y})/2$

$\boldsymbol{\eta} = \log[\boldsymbol{\mu}/(1 - \boldsymbol{\mu})]$

**enquanto** *(*$abs(difD) > 10^{-6} \quad \& \quad itr < 100$*)* **faça**

- $\boldsymbol{A} = \boldsymbol{\mu}(1 - \boldsymbol{\mu})$
- $\boldsymbol{z} = \boldsymbol{\eta} + (\boldsymbol{y} - \boldsymbol{\mu})/\boldsymbol{A}$
- $\boldsymbol{\beta} = [\boldsymbol{X^T A X}]^{-1}\boldsymbol{X^T A z}$
- $\boldsymbol{\eta} = \boldsymbol{X\beta}$
- $\boldsymbol{\mu} = \frac{\exp(\boldsymbol{\eta})}{1+\exp(\boldsymbol{\eta})}$
- $D = 2\sum_{i=1}^{n}[y_i \log(y_i/\mu_i) + (1 - y_i)\log(\frac{1-y_i}{1-\mu_i})]$
- $difD = D - D_0$
- $D_0 = D$
- $itr = itr + 1$

**fim**

### 2.5.2 Regressão Poisson

A distribuição de Poisson é a referência mais utilizada para modelar dados de contagem. A regressão Poisson baseia-se nos pressupostos inerentes ao processo e à distribuição de Poisson. A sua função de probabilidade é dada da forma (Casella e Berger, 2014):

$$f(y;\mu) = \frac{e^{-\mu}\mu^y}{y!} \tag{2.31}$$

Ao reescrever (2.31) em termos de família exponencial, conforme (5.1), a expressão resulta na seguinte forma (Nelder e Wedderburn, 1972):

$$f(y;\mu) = \exp\Bigg([y\log(\mu) - \mu] - \log(y!)\Bigg) \tag{2.32}$$

Com base na teoria de MLG, os parâmetros da distribuição Poisson estão especificados na Tabela 2.3.

Tabela 2.3: Parâmetros da distribuição Poisson

| $\phi$ | $\theta$ | $b(\theta)$ | $\mu$ | $V(\mu)$ | $c(y)$ |
|---|---|---|---|---|---|
| 1 | $\log(\mu)$ | $\mu = \exp(\theta)$ | $b'(\theta) = \exp(\theta)$ | $b''(\theta) = \exp(\theta) = \mu$ | $-\log(y!)$ |

Os componentes do MLG da distribuição Poisson são:

a) O componente aleatório: $Y_i \sim$ Poisson($\mu_i$),

b) O componente sistemático: $\boldsymbol{\eta} = \boldsymbol{X\beta}$, onde $\boldsymbol{X}$ representa a matriz do modelo e $\boldsymbol{\eta}$ o vetor de parâmetros.

c) E a função de ligação canônica: $g(\boldsymbol{\mu}) = \theta = \log(\boldsymbol{\mu})$, onde $\log(\boldsymbol{\mu}) = \boldsymbol{X\beta}$.

Para o ajuste da regressão Poisson, considerando uma taxa $\mu_i/t_i$, onde $t_i$ reflete uma variável *offset*, que pode ser o tempo de exposição ou a área de interesse do evento, o modelo fica da seguinte forma:

$$\log(\boldsymbol{\mu}) = \boldsymbol{X\beta} + \log(\boldsymbol{t}) \tag{2.33}$$

A log-verossimilhança do modelo de regressão de Poisson é dado por:

$$L(\boldsymbol{\mu}, \boldsymbol{y}) = \sum_{i=1}^{n} \left[ -\hat{\mu_i} + y_i \log(\hat{\mu_i}) - \log(y_i!) \right] \tag{2.34}$$

Ao reescrever (2.34), considerando a parametrização pela função de ligação canônica, tem-se:

$$\begin{aligned} L(\boldsymbol{\beta}, \boldsymbol{y}) &= \sum_{i=1}^{n} \left[ -e^{\boldsymbol{X_i\beta}} + y_i \log(e^{\boldsymbol{X_i\beta}}) - \log(y_i!) \right] \\ &= \sum_{i=1}^{n} \left[ -e^{\boldsymbol{X_i\beta}} + y_i \boldsymbol{X_i\beta} - \log(y_i!) \right] \end{aligned} \tag{2.35}$$

No caso da Poisson, a função *Deviance* é dada por:

$$D(\boldsymbol{y}, \hat{\boldsymbol{\mu}}) = 2 \sum_{i=1}^{n} \left[ y_i \log\left(\frac{y_i}{\hat{\mu}_i}\right) - (y_i - \hat{\mu}_i) \right] \tag{2.36}$$

O modelo de regressão Poisson desempenha, na análise de dados categorizados, o mesmo papel do modelo normal, na análise de dados contínuos. A diferença fundamental é que a estrutura multiplicativa para as médias da regressão Poisson é mais apropriada do que a estrutura aditiva das médias do modelo normal (Nelder e Wedderburn, 1972).

Com base no algoritmo do escore de Fisher apresentado na seção 2.4, o Algoritmo 3 mostra o algoritmo de estimação dos parâmetros da regressão Poisson.

**Algoritmo 3:** Regressão Poisson

**Entrada:** $D_0 = 0, \quad difD = 1, \quad offset = 0, \quad itr = 1$

$\boldsymbol{\mu} = (\boldsymbol{y} + \overline{y})/2$

$\boldsymbol{\eta} = \boldsymbol{g}(\boldsymbol{\mu}) = \log(\boldsymbol{\mu})$

**enquanto** $(abs(difD) > 10^{-6} \quad \& \quad itr < 100)$ **faça**

  $\boldsymbol{A} = \boldsymbol{\mu}$

  $\boldsymbol{z} = \boldsymbol{\eta} + (\boldsymbol{y} - \boldsymbol{\mu})/\boldsymbol{A} - offset$

  $\boldsymbol{\beta} = [\boldsymbol{X^T A X}]^{-1}\boldsymbol{X^T A z}$

  $\boldsymbol{\eta} = \boldsymbol{X\beta} + offset$

  $\boldsymbol{\mu} = \boldsymbol{g^{-1}}(\boldsymbol{\eta}) = \exp(\boldsymbol{\eta})$

  $D = 2\sum_{i=1}^{n}[y_i \log(y_i/\mu_i) - (y_i - \mu_i)]$

  $difD = D - D_0$

  $D_0 = D$

  $itr = itr + 1$

**fim**

No caso da função de ligação canônica da Poisson, que é dada por $g(\boldsymbol{\mu}) = \log(\boldsymbol{\mu})$, e se forem observados valores $y_i = 0$, utiliza-se a substituição de $y + c$ ao invés de $y$, tal que $E[g(y + c)]$ seja o mais próximo possível de $g(\boldsymbol{\mu})$. Portanto, na regressão Poisson com função logarítmica, utiliza-se $c = 1/2$.

### 2.5.3 Regressão binomial negativa

A distribuição binomial negativa é definida em termos da variável aleatória $Y$ como o número de fracassos até ocorrer o $k$-ésimo sucesso. A sua expressão é dada da forma (Casella e

Berger, 2014):

$$f(y,p,k) = \binom{y+k-1}{k-1} p^k (1-p)^y \tag{2.37}$$

onde $k > 0$ e $0 < p < 1$.

A binomial negativa é uma generalização da distribuição Geométrica (para $k = 1$), e a expressão pode ser dada conforme os termos do parâmetro de superdispersão $\alpha$, como (Casella e Berger, 2014):

$$f(y,p,\alpha) = \binom{y+\frac{1}{\alpha}-1}{\frac{1}{\alpha}-1} p^{\frac{1}{\alpha}} (1-p)^y \tag{2.38}$$

onde $\alpha = \frac{1}{k}$.

As formas especiais desta distribuição surgiram, em 1679, com Pascal e Fermat. Quando $k$ é inteiro, essa distribuição também é denominada como Pascal. Em 1907, Gosset ("Student") utilizou a distribuição binomial negativa como um modelo para contagens no lugar da distribuição Poisson (Hinde e Demétrio, 1998).

Ao reescrever (2.37) em termos de família exponencial, conforme (5.1), considerando que o parâmetro $k$ seja conhecido, tem-se (Nelder e Wedderburn, 1972):

$$f(y,p,k) = \exp\left( y\log(1-p) + k\log(p) + \log\binom{y+k-1}{k-1} \right) \tag{2.39}$$

Considerando a teoria de MLG para a distribuição da binomial negativa, os parâmetros importantes estão especificados na Tabela 2.4:

Tabela 2.4: Parâmetros da distribuição binomial negativa

| $\phi$ | $\theta$ | $b(\theta)$ | $\mu$ | $V(\mu)$ | $c(y)$ |
|---|---|---|---|---|---|
| 1 | $\log(1-p)$ | $-k\log(p) =$ $-k\log(1-\exp(\theta))$ | $b'(\theta)= \frac{k(1-p)}{p}$ | $b''(\theta)=\frac{k(1-p)}{p^2}$ | $\log\binom{y+k-1}{k-1}$ |

Os componentes do MLG da binomial negativa são:

a) O componente aleatório: $Y_i \sim$ binomial negativa$(p_i, k)$

b) O componente sistemático: $\boldsymbol{\eta} = \boldsymbol{X\beta}$, onde $\boldsymbol{X}$ representa a matriz do modelo e $\boldsymbol{\eta}$ o vetor de parâmetros.

c) E a função de ligação canônica: $g(\boldsymbol{\mu}) = \theta = \log(1-\boldsymbol{p}) = \log(\frac{\boldsymbol{\mu}}{\boldsymbol{\mu}+k})$, sendo que $\boldsymbol{p} = \frac{k}{\boldsymbol{\mu}+k}$. Portanto $\log(\frac{\boldsymbol{\mu}}{\boldsymbol{\mu}+k}) = \boldsymbol{X\beta}$.

Análogo a regressão Poisson, na regressão binomial negativa também pode-se utilizar uma variável $offset$:

$$\log\left(\frac{\boldsymbol{\mu}}{\boldsymbol{\mu}+k}\right) = \boldsymbol{X\beta} + \log(\boldsymbol{t}) \tag{2.40}$$

onde $\boldsymbol{t}$ reflete a variável $offset$.

A log-verossimilhança do modelo de regressão binomial negativo é dada por:

$$L(\boldsymbol{\mu}, k, \boldsymbol{y}) = \sum_{i=1}^{n}\left[\log\left\{\frac{\Gamma(k+y_i)}{\Gamma(k)\Gamma(1+y_i)}\right\} + k\log(k) + y_i\log(\mu_i) - (y_i+k)\log(k+\mu_i)\right] \tag{2.41}$$

Ao reescrever (2.41), considerando a parametrização pela função de ligação canônica, tem-se:

$$L(\boldsymbol{\beta}, k, \boldsymbol{y}) = \sum_{i=1}^{n}\left[\log\left\{\frac{\Gamma(k+y_i)}{\Gamma(k)\Gamma(1+y_i)}\right\} + k\log(k) + y_i\log(e^{\boldsymbol{X_i\beta}}) - (y_i+k)\log(k+e^{\boldsymbol{X_i\beta}})\right] \tag{2.42}$$

No caso da binomial negativa, a função *Deviance* é dada por:

$$D(\boldsymbol{y}, \hat{\boldsymbol{\mu}}, \boldsymbol{\alpha}) = 2\sum_{i=1}^{n}\left[y_i\log\left(\frac{y_i}{\hat{\mu_i}}\right) - \left(y_i + k\right)\log\left(\frac{1+\frac{1}{k}y_i}{1+\frac{1}{k}\hat{\mu_i}}\right)\right] \tag{2.43}$$

A variância da binomial negativa pode ser especificada em termos da média como $Var(\boldsymbol{\mu}) = \boldsymbol{\mu}(1+\alpha\boldsymbol{\mu})$. Este fato caracteriza o modelo binomial negativo como um dos modelos mais adequados para estudar superdispersão, ou seja, $Var(\boldsymbol{\mu}) > \boldsymbol{\mu}$ (Hilbe, 2011).

A principal motivação para a distribuição binomial negativa se baseia num processo de contagem heterogêneo, onde $Y \sim Poisson(\theta)$ e $\theta \sim$ Gama $(\alpha, \beta)$ (Paula, 2013):

$$g(\theta, \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} \theta^{\alpha-1} e^{-\beta\theta}, \alpha, \beta > 0 \tag{2.44}$$

com $E(\theta) = \theta = \frac{\alpha}{\beta}$ e $V(\theta) = \frac{\alpha}{\beta^2}$.

Com isso, tem-se uma mistura de distribuições: Poisson-Gama, resultando, marginalmente, na distribuição binomial negativa.

A família binomial negativa de distribuições inclui a distribuição de Poisson como um caso limite. Ao considerar a função de probabilidade da binomial negativa com $p$ sendo a probabilidade de fracasso e se $k \to \infty$ e $\mu = \frac{kp}{1-p}$, então:

$$f(y, p, k) = \frac{\Gamma(k+y)}{\Gamma(k)y!} p^y (1-p)^k \tag{2.45}$$

Como $\mu = \frac{kp}{1-p}$, $p = \frac{\mu}{k+\mu}$, ao reparametrizar (2.45) em termos de $\mu$, a expressão é dada da forma:

$$f(y, p, k) = \frac{\Gamma(k+y)}{\Gamma(k)y!} \left( \frac{\mu}{k+\mu} \right)^y \left( 1 - \frac{\mu}{k+\mu} \right)^k = \frac{\mu^y}{y!} \frac{\Gamma(k+y)}{\Gamma(k)(k+\mu)^y} \frac{1}{(1+\frac{\mu}{k})^k} \tag{2.46}$$

Sendo assim, ao considerar $k \to \infty$, a binomial negativa se caracteriza como uma opção à distribuição Poisson, como resultado obtêm-se a seguinte expressão:

$$\lim_{k \to \infty} f(y, p, k) = \frac{e^{-\mu} \mu^y}{y!} \tag{2.47}$$

Diferentemente da regressão Poisson, a regressão binomial negativa não utiliza a função de ligação canônica. O modelo tradicional da regressão binomial negativa, que é denominado como NB-2, se baseia na função de ligação logatítmica $g(\boldsymbol{\mu}) = \theta = \log(\boldsymbol{\mu})$ (Hilbe, 2011). Portanto, neste caso da regressão NB-2, resultados distintos podem surgir nos erros padrões das estimativas dos métodos de estimação de parâmetros apresentados na seção anterior (Newton-

Raphson e Escore de Fisher). Sendo assim, uma alteração é feita no Método Escore de Fisher para corrigir os erros padrões, permitindo que sejam calculados com base na matriz de informação (Hilbe, 2011).

Uma nova matriz diagonal $A_0$ é definida:

$$a_{i0} = \frac{1}{V(\boldsymbol{y})}\left(\frac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{\eta}}\right)^2 + (\boldsymbol{y} - \boldsymbol{\mu})\frac{V(\boldsymbol{\mu})g''(\boldsymbol{\mu}) + V'(\boldsymbol{\mu})g'(\boldsymbol{\mu})}{V(\boldsymbol{\mu})^2 g'(\boldsymbol{\mu})^3} \tag{2.48}$$

Portanto, a expressão para $\boldsymbol{A}$ é dada da forma:

$$\begin{aligned} A &= \sum_{i=1}^{n} \frac{\mu_i}{1+\alpha\mu_i} + (y_i - \mu_i)\frac{\alpha\mu_i}{(1+2\alpha\mu_i + \mu_i^2\alpha^2)} \\ &= \sum_{i=1}^{n} \frac{e^{\boldsymbol{X_i\beta}}}{1+\frac{e^{\boldsymbol{X_i\beta}}}{k}} + \frac{(y_i - e^{\boldsymbol{X_i\beta}})e^{\boldsymbol{X_i\beta}}}{k\left(1+\frac{e^{\boldsymbol{X_i\beta}}}{k}\right)^2} \\ &= \sum_{i=1}^{n} \frac{ke^{\boldsymbol{X_i\beta}}}{k+e^{\boldsymbol{X_i\beta}}} + \frac{(y_i - e^{\boldsymbol{X_i\beta}})e^{\boldsymbol{X_i\beta}}k}{(k+e^{\boldsymbol{X_i\beta}})^2} \\ &= \sum_{i=1}^{n} \frac{ke^{\boldsymbol{X_i\beta}}}{k+e^{\boldsymbol{X_i\beta}}}\left(\frac{y_i - e^{\boldsymbol{X_i\beta}}}{k+e^{\boldsymbol{X_i\beta}}} + 1\right) \end{aligned} \tag{2.49}$$

Para o cálculo da variância do parâmetro $\alpha$, pode-se utilizar também o Método Escore de Fisher, ou ainda utilizar a segunda derivada da log-verossimilhança do modelo. Neste caso, a segunda derivada da log-verossimilhança do modelo binomial negativo é dado da forma:

$$H = \sum_{i=1}^{n} \psi'(k+\boldsymbol{y}) - \psi'(k) + \frac{1}{k} - \frac{2}{(k+\boldsymbol{\mu})} + \frac{(\boldsymbol{y}+k)}{[(k+\boldsymbol{\mu})(k+\boldsymbol{\mu})]} \tag{2.50}$$

onde $\psi(.)$ e $\psi'(.)$, são as funções digama e trigama, respectivamente, sendo expressas: $\psi(z) = \frac{\partial \log\Gamma(z)}{\partial z}$ e $\psi'(z) = \frac{\partial\psi(z)}{\partial z} = \frac{\partial^2 \log\Gamma(z)}{\partial z^2}$.

O Algoritmo 4 mostra o algoritmo de estimação dos parâmetros da regressão binomial negativa.

**Algoritmo 4:** Regressão Binomial Negativa

**Entrada:** $k = 1, \quad ddpar = 1, \quad itr = 0$

$\boldsymbol{\mu} = (\boldsymbol{y} + \overline{y})/2$

$\boldsymbol{\eta} = \log(\boldsymbol{\mu})$

$\boldsymbol{\beta = [X^T X]^{-1} X^T \eta}$

**enquanto** *$(abs(ddpar) > 10^{-6})$* **faça**

  $dpar = 1, parold = k$

  **enquanto** *$(abs(dpar) > 10^{-6})$* **faça**

    $\boldsymbol{g} = \sum_{i=1}^{n}[\partial\Gamma(k + y_i) - \partial\Gamma(k) + \log(k) + 1 - \log(k + \mu_i) - (k + y_i)/(k + \mu_i)]$

    $\boldsymbol{H} = \sum_{i=1}^{n}[\partial^2\Gamma(k+y_i) - \partial^2\Gamma(k) + 1/k - 2/(k+\mu_i) + (y_i+k)/((k+\mu_i)^T(k+\mu_i))]$

    $k_0 = k$

    $k = k_0 - \boldsymbol{H^{-1} g}$

    $dpar = k - k_0$

  **fim**

  **Entrada:** $D_0 = 0, \quad difD = 1$

  $\boldsymbol{\mu} = (\boldsymbol{y} + \overline{y})/2$

  $\boldsymbol{\eta} = \log(\boldsymbol{\mu})$

  $\alpha = 1/k$

  **enquanto** *$(abs(difD) > 10^{-6})$* **faça**

    $\boldsymbol{A} = (\boldsymbol{\mu}/(1 + \alpha\boldsymbol{\mu})) + (\boldsymbol{y} - \boldsymbol{\mu})(\alpha\boldsymbol{\mu}/(1 + 2\alpha\boldsymbol{\mu} + \alpha^2\boldsymbol{\mu}^2))$

    $\boldsymbol{z} = \boldsymbol{\eta} + (\boldsymbol{y - \mu})/(\boldsymbol{A}(1 + \alpha\boldsymbol{\mu})) - offset$

    $\boldsymbol{\beta = [X^T A X]^{-1} X^T A z}$

    $\boldsymbol{\eta = X\beta} + offset$

    $\boldsymbol{\mu} = \exp(\boldsymbol{\eta})$

    $D = 2\sum_{i=1}^{n}[y_i \log(y_i/\mu_i) - (y_i + 1/\alpha)\log((1 + \alpha y_i)/(1 + \alpha\mu_i))]$

    $difD = D - D_0$

    $D_0 = D$

    $itr = itr + 1$

  **fim**

**fim**

Considerando uma função $g(.)$ diferenciável em $\theta$, se ocorre a convergência em distribuição $\hat{\theta}_n \to N(\theta, \sigma_n^2)$, então $g(\hat{\theta}_n) \to N(g(\theta_n), [g'(\theta)]^2 \times \sigma_n^2)$ também converge em distribuição (Casella e Berger, 2014). As estimativas de máxima verossimilhança local são assintoticamente Normais, não viesadas e consistentes (Staniswalis, 1989). Portanto, tem-se

$$\hat{\alpha} = \frac{1}{\hat{k}} \tag{2.51}$$

$$Var(\hat{\alpha}) = -\frac{1}{H\hat{k}^4} \tag{2.52}$$

sendo que $Var(\hat{k}) = -\frac{1}{H}$, onde $H$ é dado por (2.50).

Os modelos inflacionados de zeros (Poisson inflacionado de zeros e binomial negativo inflacionado de zeros) são misturas de distribuições da família exponencial, entretanto, esses modelos não fazem parte da família exponencial. Uma estrutura similar para estes modelos será apresentada a seguir.

# Capítulo 3

# Modelos Inflacionados de Zeros (Poisson e Binomial Negativo)

## 3.1 Introdução

Lambert (1992) introduziu o uso de modelos de regressão para contagens com distribuição inflacionada de zeros, com o modelo Poisson inflacionado de zeros, sendo um dos mais utilizados para dados com excesso de zeros. O modelo binomial negativo inflacionado de zeros também se torna uma boa alternativa para este tipo de modelagem.

A estrutura apresentada na Figura 1.8 será demonstrada neste Capítulo. Por meio dos parâmetros de superdispersão, pode-se verificar que a distribuição binomial negativa inflacionada de zeros se reduz para a distribuição binomial negativa, se tais parâmetros não forem considerados significativos. Considerando que os parâmetros inflacionados não são significativos na distribuição Poisson inflacionada de zeros, então ela se reduz para uma distribuição de Poisson e se os parâmetros não forem considerados significativos na distribuição binomial negativa inflacionada de zeros, então ela se reduz para uma distribuição binomial negativa.

### 3.2 Regressão Poisson inflacionada de zeros

A distribuição Poisson inflacionada de zeros foi descrita por Cohen (1963), SINGH (1963) e Johnson e Kotz (1969). Essa distribuição surge para a aplicação de dados de contagem em que a frequência apresenta excesso de valores iguais a zero.

A distribuição Poisson inflacionada de zeros segue uma estrutura que pode ser dividida em dois estados. A primeira parte se refere ao estado zero, onde apenas valores nulos são observados e a segunda parte se refere ao estado Poisson, onde os valores diferentes de zero são observados (Hall, 2000). Portanto, a sua expressão é dada da forma (Lambert, 1992):

$$f(y,p,\mu) = \begin{cases} p + (1-p)e^{-\mu}, & y = 0 \\ (1-p)\frac{e^{-\mu}\mu^y}{y!}, & y > 0 \end{cases} \tag{3.1}$$

onde $\mu$ representa a média da distribuição de Poisson e $p$ representa a probabilidade da ocorrência de zeros.

A média da distribuição Poisson inflacionada de zeros pode ser calculada da forma (Lambert, 1992):

$$\begin{aligned} E(Y) &= \sum_{y=1}^{\infty} y f(y,p,\mu) = \sum_{y=0}^{\infty} y(1-p)\frac{e^{-\mu}\mu^y}{y!} \\ &= (1-p)e^{-\mu}\sum_{y=1}^{\infty} y\frac{\mu^y}{y!} = (1-p)e^{-\mu}\mu e^{\mu} \\ &= (1-p)\mu \end{aligned} \tag{3.2}$$

e para obter a variância, é necessário calcular primeiro $E(Y^2)$:

$$\begin{aligned}
E(Y^2) &= \sum_{y=1}^{\infty} y^2 f(y,p,\mu) = \sum_{y=0}^{\infty} y^2(1-p)\frac{e^{-\mu}\mu^y}{y!} \\
&= (1-p)e^{-\mu}\sum_{y=1}^{\infty} y^2\frac{\mu^y}{y!} = (1-p)e^{-\mu}\mu e^{\mu}(\mu+1) \\
&= (1-p)(\mu^2+\mu) \qquad (3.3)
\end{aligned}$$

Com isso, a variância, $V(Y)$ pode ser obtida:

$$\begin{aligned}
V(Y) &= E(Y^2) - [E(Y)]^2 = (1-p)(\mu^2+\mu) - [(1-p)\mu]^2 \\
&= (\mu^2+\mu-p\mu^2-p\mu) - (\mu-p\mu)^2 \\
&= (\mu^2+\mu-p\mu^2-p\mu) - (\mu^2-2p\mu^2-p^2\mu^2) \\
&= \mu(1-p)(1+p\mu) \qquad (3.4)
\end{aligned}$$

indicando que a distribuição marginal de $Y$ exibe superdispersão, se $p > 0$. Isso se reduz ao modelo de Poisson quando $p = 0$.

O modelo Poisson inflacionado de zeros foi proposto por Lambert (1992) e os parâmetros $\mu$ e $p$ são estimados com base na teoria de MLG. Portanto, para o ajuste do modelo Poisson inflacionado de zeros, a função de ligação se baseia em duas partes (Ridout et al., 1998):

$i)$ A função de ligação da parte não inflacionada de zeros, sendo a mesma do modelo Poisson, e expressa como:

$$\log(\boldsymbol{\mu}) = \boldsymbol{X\beta} \qquad (3.5)$$

$ii)$ A função de ligação da parte inflacionada de zeros, sendo o preditor linear dado pela função de ligação logit, e expressa como:

$$\log\left(\frac{\boldsymbol{p}}{1-\boldsymbol{p}}\right) = \boldsymbol{G\gamma} \qquad (3.6)$$

onde $\boldsymbol{X}$ e $\boldsymbol{G}$ representam as matrizes de covariáveis para o modelo Poisson e para o modelo logístico, respectivamente e $\boldsymbol{\beta}$ e $\boldsymbol{\gamma}$ representam os vetores dos parâmetros dos modelos.

O número de parâmetros que pode ser estimado em um modelo de regressão Poisson inflacionado de zeros depende da quantidade de variáveis disponíveis. Com base no modelo Poisson inflacionado de zeros proposto por Lambert (1992), as matrizes $\boldsymbol{G}$ e $\boldsymbol{X}$ podem possuir conjuntos diferentes de fatores experimentais e efeitos de covariáveis que pertencem à probabilidade $p$ (no estado zero) e à média da Poisson $\mu$ (no estado Poisson). Se as mesmas covariáveis afetam $\mu$ e $p$, então pode-se pensar em $p$ como função de $\mu$ para a redução do número de parâmetros.

Para a obtenção dos parâmetros do modelo Poisson inflacionado de zeros com covariáveis, Lambert (1992) propôs o ajuste do modelo Poisson inflacionado de zeros utilizando o estimador de máxima verossimilhança por meio do algoritmo EM proposto por Dempster et al. (1977). Se $\mu$ e $p$ não forem funcionalmente relacionados, a log-verossimilhança para a regressão Poisson inflacionada de zeros com a parametrização padrão (3.5) e (3.6) é dada da forma:

$$
\begin{aligned}
L(\boldsymbol{\gamma},\boldsymbol{\beta},\boldsymbol{y}) &= \log\left(\prod_{y_i=0}^{n}[p+(1-p)e^{-\mu}]\prod_{y_i>0}^{n}(1-p)\frac{e^{-\mu}\mu^{y_i}}{y!}\right) \\
&= \sum_{y_i=0}\log\left[\left(\frac{p}{1-p}+e^{-\mu}\right)(1-p)\right]+\sum_{y_i>0}[\log(1-p)-\mu+y_i\log(\mu)-\log(y_i!)]
\end{aligned}
\tag{3.7}
$$

$$
\begin{aligned}
&= \sum_{y_i=0}\log(e^{\boldsymbol{G_i\gamma}}+\exp(-e^{\boldsymbol{X_i\beta}}))+\sum_{y_i=0}\log(1-p) \\
&\quad +\sum_{y_i>0}\log(1-p)+\sum_{y_i>0}(y_i\boldsymbol{X_i\beta}-e^{\boldsymbol{X_i\beta}})-\sum_{y_i>0}\log(y_i!) \\
&= \sum_{y_i=0}\log(e^{\boldsymbol{G_i\gamma}}+\exp(-e^{\boldsymbol{X_i\beta}}))+\sum_{y_i>0}(y_i\boldsymbol{X_i\beta}-e^{\boldsymbol{X_i\beta}})- \\
&\quad \sum_{i=1}^{n}\log(1+e^{\boldsymbol{G_i\gamma}})-\sum_{y_i>0}\log(y_i!)
\end{aligned}
$$

onde $\boldsymbol{G_i}$ e $\boldsymbol{X_i}$ representam as i-ésimas linhas de $\boldsymbol{G}$ e $\boldsymbol{X}$.

A soma das exponenciais no primeiro termo de (3.7) dificulta a maximização de $L(\boldsymbol{\gamma}, \boldsymbol{\beta}, \boldsymbol{y})$. Lambert (1992) traz uma modificação em (3.7), acrescentando uma variável aleatória $z$. Portanto, supondo que seja possível identificar quais zeros pertencem ao estado zero e quais zeros pertencem ao estado Poisson, considera-se que $z_i = 1$ quando $y_i$ for do estado zero e $z_i = 0$ quando $y_i$ for do estado Poisson. A primeira etapa do Algoritmo consiste em substituir $\boldsymbol{z}$ pela esperança condicional dado por $\boldsymbol{y}$, $\boldsymbol{\gamma^{(m)}}$ e $\boldsymbol{\beta^{(m)}}$. Esta esperança condicional pode ser calculada da forma: (Lambert, 1992):

$$z_i^{(m)} = \begin{cases} [1 + e^{-\boldsymbol{G_i}\boldsymbol{\gamma}^{(k)} - e^{\boldsymbol{X_i}\boldsymbol{\beta}^{(k)}}}]^{-1}; \quad y_i = 0 \\ 0; \quad y_i > 0 \end{cases} \tag{3.8}$$

então a log-verossimilhança com os dados completos ($\boldsymbol{y}$,$\boldsymbol{z}$) será:

$$\begin{aligned} L(\boldsymbol{\gamma}, \boldsymbol{\beta}, \boldsymbol{y}, \boldsymbol{z}) &= \log\left(\prod_{i=1}^{n} f(\boldsymbol{y}, \boldsymbol{z}, \boldsymbol{\gamma}, \boldsymbol{\beta})\right) = \sum_{i=1}^{n} \log[f(z_i|\boldsymbol{\gamma}) f(y_i|z_i, \boldsymbol{\beta})] \\ &= \sum_{i=1}^{n} \log(f(z_i|\boldsymbol{\gamma})) + \sum_{i=1}^{n} \log(f(y_i|z_i, \boldsymbol{\beta})) \\ &= \sum_{i=1}^{n} (z_i \boldsymbol{G_i}\boldsymbol{\gamma} - \log(1 + e^{\boldsymbol{G_i}\boldsymbol{\gamma}})) + \sum_{i=1}^{n} (1 - z_i)(y_i \boldsymbol{X_i}\boldsymbol{\beta} - e^{\boldsymbol{X_i}\boldsymbol{\beta}}) - \\ &\quad \sum_{i=1}^{n} (1 - z_i) \log(y_i!) \end{aligned}$$

$$L(\boldsymbol{\gamma}, \boldsymbol{\beta}, \boldsymbol{y}, \boldsymbol{z}) = L_c(\boldsymbol{\gamma}, \boldsymbol{y}, \boldsymbol{z}) + L_c(\boldsymbol{\beta}, \boldsymbol{y}, \boldsymbol{z}) - \sum_{i=1}^{n} (1 - z_i) \log(y_i!) \tag{3.9}$$

Para demonstrar a ideia apresentada na Figura 1.8, considerando que $\boldsymbol{\gamma} = 0$, ou seja, $\boldsymbol{z} = 0$

tem-se:

$$\begin{aligned} L(\boldsymbol{\gamma}, \boldsymbol{\beta}, \boldsymbol{y}, \boldsymbol{z}) &= \sum_{i=1}^{n} \log(f(z_i|\boldsymbol{\gamma})) + \sum_{i=1}^{n} \log(f(y_i|z_i, \boldsymbol{\beta})) \\ &= -n \log(1 + e^0) + \sum_{i=1}^{n} (y_i \boldsymbol{X_i} \boldsymbol{\beta} - e^{\boldsymbol{X_i} \boldsymbol{\beta}}) - \sum_{i=1}^{n} \log(y_i!) \end{aligned} \quad (3.10)$$

Nota-se que $-n \log(1 + e^0)$ é uma constante que não influencia na estimação do parâmetro $\boldsymbol{\beta}$, sendo assim, a expressão se resume aproximadamente:

$$\begin{aligned} L(\boldsymbol{\gamma}, \boldsymbol{\beta}, \boldsymbol{y}, \boldsymbol{z}) &= \sum_{i=1}^{n} \log(f(z_i|\boldsymbol{\gamma})) + \sum_{i=1}^{n} \log(f(y_i|z_i, \boldsymbol{\beta})) \\ &\approx \sum_{i=1}^{n} (y_i \boldsymbol{X_i} \boldsymbol{\beta} - e^{\boldsymbol{X_i} \boldsymbol{\beta}}) - \sum_{i=1}^{n} \log(y_i!) \end{aligned} \quad (3.11)$$

Portanto, pode-se observar que esta log-verossimilhança (3.11) se torna a log-verossimilhança do modelo de regressão de Poisson, expresso em (2.35). Na log-verossimilhança (3.9) pode-se maximizar sepadamente $L_c(\boldsymbol{\gamma}, \boldsymbol{y}, \boldsymbol{z})$ e $L_c(\boldsymbol{\beta}, \boldsymbol{y}, \boldsymbol{z})$. Pode-se observar que a Equação (3.9) é linear em $Z$.

Com a substituição (3.8), pode-se maximizar a log-verossimilhança (3.9) e então perceber que a expressão se resume à soma da log-verossimilhança da regressão logística não ponderada de $\boldsymbol{z}^{(m)}$ em $\boldsymbol{G}$ (sendo um termo que não envolve $\boldsymbol{\beta}$) e a log-verossimilhança da regressão de Poisson ponderada de $\boldsymbol{y}$ em $\boldsymbol{X}$ (um termo que não envolve $\boldsymbol{\gamma}$) (Hall, 2000).

A segunda etapa da iteração $(m + 1)$ consiste em obter $\boldsymbol{\gamma}^{(m+1)}$, maximizando $L_c(\boldsymbol{\gamma}, \boldsymbol{y}, \boldsymbol{z})$ por meio da regressão logística não ponderada de $\boldsymbol{z}^{(m)}$ em $\boldsymbol{G}$. Em seguida obter $\boldsymbol{\beta}^{(m+1)}$, maximizando $L_c(\boldsymbol{\beta}, \boldsymbol{y}, \boldsymbol{z}) = \sum_{i=1}^{n} (1 - z_i)(y_i \boldsymbol{X}_i \boldsymbol{\beta} - e^{\boldsymbol{X}_i \boldsymbol{\beta}}) - \sum_{i=1}^{n} (1 - z_i) \log(y_i!)$, por meio da regressão de Poisson ponderada com pesos $(1 - z_i)$ (Hall, 2000).

De forma resumida, essa estimação do Algoritmo Expectation Maximization (EM) pode ser vista no Algoritmo 5.

Baseando-se em (2.15), para obtenção da matriz de informação de fisher, é necessário reali-

**Algoritmo 5:** Algoritmo EM - Poisson inflacionada de zeros

**Entrada:** $\boldsymbol{\beta}, \boldsymbol{\gamma}$

Estimar $\boldsymbol{\beta}$ para $y > 0$, por meio regressão Poisson.

$\boldsymbol{\gamma_0} = (\sum_{i=1}^{n} I_{(y_i=0)} - \sum_{i=1}^{n} \exp(-\exp(\boldsymbol{X_i}\boldsymbol{\beta})))/n$

Estimar $\boldsymbol{\gamma} = \log(\boldsymbol{\gamma_0}/(1-\boldsymbol{\gamma_0}))$

$DiffD = 1, \quad OldD = 0$

**enquanto** *($abs(DiffD) > 10^{-6}$)* **faça**

  **Passo E:** Estimar a esperança condicional $z_i$

  **se** $y_i = 0$ **então**

    $z_i = 1/(1+\exp(-\boldsymbol{G_i}\boldsymbol{\gamma} - \exp(\boldsymbol{X_i}\boldsymbol{\beta})))$

  **senão**

    $0$

  **fim**

  **Passo M para $\boldsymbol{\beta}$:** Estimação do modelo Poisson ponderado por $(1-\boldsymbol{z})$, minimizando a $Deviance$ $(D_1)$:

  { $\boldsymbol{\eta} = \boldsymbol{X}\boldsymbol{\beta} + offset$

  $\boldsymbol{\mu} = \exp(\boldsymbol{\eta})$

  $D_1 = \sum_{i=1}^{n}[(1-z_i)(y_i\eta_i - \mu_i)]$

  }

  **Passo M para $\boldsymbol{\gamma}$:** Estimação do modelo logístico não ponderado, utilizando $z$ como variável resposta, minimizando a $Deviance$ $(D_2)$

  { $\boldsymbol{\eta} = \boldsymbol{G}\boldsymbol{\gamma}$

  $D_2 = \sum_{i=1}^{n}(z_i\eta_i - \sum_{i=1}^{n}(\log(1+\exp(\eta_i)))$

  }

  **Maximização:** $OldD = D$

  $D = D_1 + D_2$

  $DiffD = OldD - D$

**fim**

zar o cálculo das segundas derivadas referentes aos parâmetros $\boldsymbol{\beta}$ e $\boldsymbol{\gamma}$ em $L(\boldsymbol{\gamma}, \boldsymbol{\beta}, \boldsymbol{y}, \boldsymbol{z})$. Sendo assim, a matriz de informação correspondente ao modelo Poisson inflacionado de zeros, considerando os parâmetros $\boldsymbol{\beta}$ e $\boldsymbol{\gamma}$ é dada:

$$\boldsymbol{I}(\boldsymbol{\beta}, \boldsymbol{\gamma}) = \begin{pmatrix} I_{11} & I_{12} \\ I_{21} & I_{22} \end{pmatrix} \tag{3.12}$$

E a matriz inversa é dada:

$$[I(\boldsymbol{\beta}, \boldsymbol{\gamma})]^{-1} = \begin{pmatrix} Var(\hat{\boldsymbol{\beta}}) & Cov(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}}) \\ Cov(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}}) & Var(\hat{\boldsymbol{\gamma}}) \end{pmatrix} \tag{3.13}$$

Sendo assim, as equações com as segundas derivadas são apresentadas a seguir:

$$I_{11} = \frac{\partial^2 L(\boldsymbol{\gamma}, \boldsymbol{\beta}, \boldsymbol{y}, \boldsymbol{z})}{\partial \beta_j \partial \beta_r} = \sum_{i,j,r;y=0} \frac{\exp(-e^{\boldsymbol{X_i\beta}})e^{\boldsymbol{X_i\beta}} X_{ij} X_{ir} \left( e^{\boldsymbol{X_i\beta}} - \frac{e^{\boldsymbol{X_i\beta}} \exp(-e^{\boldsymbol{X_i\beta}}))}{e^{\boldsymbol{G_i\gamma}} + \exp(-e^{\boldsymbol{X_i\beta}})} - 1 \right)}{e^{\boldsymbol{G_i\gamma}} + \exp(-e^{\boldsymbol{X_i\beta}})} - \sum_{i;y>0} X_{ij} X_{ir} e^{\boldsymbol{X_i\beta}} \tag{3.14}$$

$$I_{22} = \frac{\partial^2 L(\boldsymbol{\gamma}, \boldsymbol{\beta}, \boldsymbol{y}, \boldsymbol{z})}{\partial \gamma_j \partial \gamma_r} = \sum_{i;y=0} \frac{e^{\boldsymbol{G_i\gamma}} G_{ij} G_{ir} \exp(-e^{\boldsymbol{X_i\beta}})}{(e^{\boldsymbol{G_i\gamma}} + \exp(-e^{\boldsymbol{X_i\beta}}))^2} - \sum_{i=1}^{n} \frac{e^{\boldsymbol{G_i\gamma}} G_{ij} G_{ir}}{(1 + e^{\boldsymbol{G_i\gamma}})^2} \tag{3.15}$$

$$I_{12} = I_{21} = \frac{\partial^2 L(\boldsymbol{\gamma}, \boldsymbol{\beta}, \boldsymbol{y}, \boldsymbol{z})}{\partial \gamma_j \partial \beta_r} = \sum_{i;y=0} \frac{e^{\boldsymbol{G_i\gamma}} G_{ij} \exp(-e^{\boldsymbol{X_i\beta}}) e^{\boldsymbol{X_i\beta}} X_{ir}}{(e^{\boldsymbol{G_i\gamma}} + \exp(-e^{\boldsymbol{X_i\beta}}))^2} \tag{3.16}$$

As demonstrações dessas derivadas (3.14), (3.15) e (3.16), podem ser vistas no Apêndice B.

Seguindo a ideia feita em (3.11), para mostrar que a Poisson inflacionada de zeros irá se reduzir em uma Poisson, pode-se considerar $\boldsymbol{\gamma} = 0$, ou equivalentemente, todas as equações referentes ao estado $y_i = 0$ sejam iguais a 0 e $y_i > 0$ passa a ser $y_i \geq 0$, isto é, considerando todos os dados, e substituir esta expressão nas segundas derivadas. Sendo assim, a segunda derivada, em função do parâmetro $\boldsymbol{\beta}$ pode ser expressa da forma:

$$\begin{aligned} I_{11} = \frac{\partial^2 L(\boldsymbol{\gamma},\boldsymbol{\beta},\boldsymbol{y},\boldsymbol{z})}{\partial\beta_j\partial\beta_r} &= \sum_{i;y=0} \frac{\exp(-e^{\boldsymbol{X_i\beta}})e^{\boldsymbol{X_i\beta}}X_{ij}X_{ir}\left(e^{\boldsymbol{X_i\beta}} - \frac{e^{\boldsymbol{X_i\beta}}\exp(-e^{\boldsymbol{X_i\beta}}))}{e^{\boldsymbol{G_i\gamma}}+\exp(-e^{\boldsymbol{X_i\beta}})} - 1\right)}{e^{\boldsymbol{G_i\gamma}} + \exp(-e^{\boldsymbol{X_i\beta}})} - \sum_{i;y>0} X_{ij}X_{ir}e^{\boldsymbol{X_i\beta}} \\ &= 0 - \sum_{i;y\geq 0} X_{ij}X_{ir}e^{\boldsymbol{X_i\beta}} = -\sum_{i=1}^{n} X_{ij}X_{ir}e^{\boldsymbol{X_i\beta}} \\ &= -\boldsymbol{X^T\mu X} \end{aligned} \tag{3.17}$$

Pode-se observar que (3.17) resulta na segunda derivada da log-veromilhança do modelo Poisson (2.35) com relação a $\boldsymbol{\beta}$. Portanto, se o parâmetro $\boldsymbol{\gamma} = 0$, a expressão retorna à uma Poisson. Desconsiderar a parte $\sum_{y_i=0}\{\}$ ou $0 \times \sum_{y_i=0}\{\}$ e fazer com que a parte $\sum_{y_i>0}\{\}$ se torne $\sum_{y_i\geq 0}\{\}$ é uma forma computacional para fazer com que (3.13) possa gerar a variância de $\boldsymbol{\beta}$ na regressão Poisson.

E as outras derivadas, usando o mesmo argumento, resulta nas seguintes expressões:

$$\begin{aligned} \frac{\partial^2 L(\boldsymbol{\gamma},\boldsymbol{\beta},\boldsymbol{y},\boldsymbol{z})}{\partial\gamma_j\partial\gamma_r} &= \sum_{i;y=0} \frac{e^{\boldsymbol{G_i\gamma}}G_{ij}G_{ir}\exp(-e^{\boldsymbol{X_i\beta}})}{(e^{\boldsymbol{G_i\gamma}} + \exp(-e^{\boldsymbol{X_i\beta}}))^2} - \sum_{i=1}^{n} \frac{e^{\boldsymbol{G_i\gamma}}G_{ij}G_{ir}}{(1+e^{\boldsymbol{G_i\gamma}})^2} \\ &= 0 \end{aligned} \tag{3.18}$$

$$\begin{aligned} \frac{\partial^2 L(\boldsymbol{\gamma},\boldsymbol{\beta},\boldsymbol{y},\boldsymbol{z})}{\partial\gamma_j\partial\beta_r} &= \sum_{i;y=0} \frac{e^{\boldsymbol{G_i\gamma}}G_{ij}\exp(-e^{\boldsymbol{X_i\beta}})e^{\boldsymbol{X_i\beta}}X_{ir}}{(e^{\boldsymbol{G_i\gamma}} + \exp(-e^{\boldsymbol{X_i\beta}}))^2} \\ &= 0 \end{aligned} \tag{3.19}$$

## 3.3 Regressão binomial negativa inflacionada de zeros

A distribuição binomial negativa inflacionada de zeros, com parâmetros $\mu$, $k$ e $p$, pode ser expressa da forma (Yau et al., 2003):

$$f(y, p, \mu, k) = \begin{cases} p + (1-p)\left(\frac{k}{k+\mu}\right)^k, & y = 0 \\ (1-p)\frac{\Gamma(y+k)}{\Gamma(k)y!}\left(\frac{k}{k+\mu}\right)^k\left(\frac{\mu}{k+\mu}\right)^y, & y > 0 \end{cases} \tag{3.20}$$

onde $p$ representa a probabilidade de zeros e $\mu$ representa a média da distribuição binomial negativa.

De forma análoga ao modelo Poisson inflacionado de zeros, a distribuição binomial negativa inflacionada de zeros apresenta uma modelagem para as contagens nulas, sendo chamada de estado zero e para as contagens não nulas, sendo denominada como estado binomial negativo (Garay et al., 2011). A binomial negativa inflacionada de zeros (3.20) se reduz para a distribuição Poisson inflacionada de zeros (3.1) quando $k \to \infty$ (Yau et al., 2003). Portanto, quando $y = 0$, tem-se:

$$\begin{aligned} f(y, p, \mu, k) &= \lim_{k\to\infty} p + (1-p)\left(\frac{k}{k+\mu}\right)^k \\ &= \lim_{k\to\infty} p + (1-p)\frac{1}{(1+\frac{\mu}{k})^k} \\ &= p + (1-p)e^{-\mu} \end{aligned} \tag{3.21}$$

E quando $y > 0$, tem-se:

$$\begin{aligned} f(y, p, \mu, k) &= \lim_{k\to\infty} (1-p)\frac{\Gamma(y+k)}{\Gamma(k)y!}\left(\frac{k}{k+\mu}\right)^k\left(\frac{\mu}{k+\mu}\right)^y \\ &= (1-p)\lim_{k\to\infty}\frac{\mu^y}{y!}\frac{\Gamma(k+y)}{\Gamma(k)(k+\mu)^y}\frac{1}{(1+\frac{\mu}{k})^k} \\ &= (1-p)\frac{\mu^y e^{-\mu}}{y!} \end{aligned} \tag{3.22}$$

Portanto, nota-se que (3.21) e (3.22) se reduz a Poisson inflacionada de zeros (3.1).

Com base na densidade do modelo binomial negativo inflacionado de zeros, pode-se calcular a média da distribuição:

$$\begin{aligned}
E(Y) &= \sum_{y=1}^{\infty} y f(y,p,\mu,k) = \sum_{y=1}^{\infty} y(1-p)\frac{\Gamma(y+k)}{\Gamma(k)y!}\left(\frac{k}{k+\mu}\right)^k \left(\frac{\mu}{k+\mu}\right)^y \\
&= (1-p)\frac{1}{\Gamma(k)}\left(\frac{k}{k+\mu}\right)^k \sum_{y=1}^{\infty} y\frac{\Gamma(y+k)}{y!}\frac{\mu^y}{(k+\mu)^y} \\
&= (1-p)\frac{1}{\Gamma(k)}\left(\frac{k}{k+\mu}\right)^k \Gamma(1+k)\mu\left(\frac{k}{k+\mu}\right)^{-k} k^{-1} \\
&= (1-p)\mu \qquad (3.23)
\end{aligned}$$

E para obter a variância, é necessário calcular a $E(Y^2)$:

$$\begin{aligned}
E(Y^2) &= \sum_{y=1}^{\infty} y^2 f(y,p,\mu,k) = (1-p)\frac{1}{\Gamma(k)}\left(\frac{k}{k+\mu}\right)^k \sum_{y=1}^{\infty} y^2\frac{\Gamma(y+k)}{y!}\frac{\mu^y}{(k+\mu)^y} \\
&= (1-p)\frac{(k+(1+k)\mu)\mu}{k} \\
&= (1-p)\mu\frac{(k+\mu+\mu k)}{k} \\
&= (1-p)\mu(1+\frac{\mu}{k}+\mu) \qquad (3.24)
\end{aligned}$$

Com isso, a variância, $V(Y)$ pode ser obtida:

$$\begin{aligned}
V(Y) &= E(Y^2) - [E(Y)]^2 = (1-p)\mu(1+\frac{\mu}{k}+\mu) - [(1-p)\mu]^2 \\
&= (1-p)\mu(1+\frac{\mu}{k}+\mu) - (\mu^2 - 2p\mu^2 + p^2\mu^2) \\
&= (\mu + \frac{\mu^2}{k} + \mu^2 - p\mu - \frac{p\mu^2}{k} - p\mu^2) - (\mu^2 - 2p\mu^2 + p^2\mu^2) \\
&= \mu(1 + \frac{\mu}{k} + p\mu - p - \frac{p\mu}{k} - p^2\mu) \\
&= (1-p)(1+\frac{\mu}{k}+p\mu)\mu \qquad (3.25)
\end{aligned}$$

Considerando $k \to \infty$ em (3.25), tem-se:

$$
\begin{aligned}
V(Y) &= \lim_{k\to\infty}(1-p)(1+\frac{\mu}{k}+p\mu)\mu \\
&= (1-p)(1+0+p\mu)\mu \\
&= (1-p)(1+p\mu)\mu
\end{aligned} \tag{3.26}
$$

sendo (3.26) a variância da distribuição Poisson inflacionada de zeros, obtida em (3.4).

A estimação dos parâmetros do modelo binomial negativo inflacionado de zeros na presença de covariáveis segue a mesma ideia proposta por Lambert (1992) na estimação dos parâmetros do modelo Poisson inflacionado de zeros (Garay et al., 2011). Portanto, as funções de ligação também se baseiam nas partes não inflacionadas e inflacionadas de zero, sendo expressas da seguinte forma, respectivamente:

$i)$ A função de ligação da parte não inflacionada de zeros, sendo a mesma do modelo binomial negativo, é expressa como:

$$\log(\boldsymbol{\mu}) = \boldsymbol{X\beta} \tag{3.27}$$

$ii)$ A função de ligação da parte inflacionada de zeros, sendo o preditor linear dado pela função de ligação logit, é expressa como:

$$\log\left(\frac{\boldsymbol{p}}{1-\boldsymbol{p}}\right) = \boldsymbol{G\gamma} \tag{3.28}$$

onde $\boldsymbol{X}$ e $\boldsymbol{G}$ representam as matrizes de covariáveis para o modelo binomial negativo e para o modelo binomial, respectivamente e $\boldsymbol{\beta}$ e $\boldsymbol{\gamma}$ representam os vetores dos parâmetros dos modelos.

Assim como o modelo Poisson inflacionado de zeros, o modelo binomial negativo inflacionado de zeros deve ser ajustado de forma simultânea, considerando as duas partes distintas, sendo que as funções de ligação também são análogas ao modelo Poisson inflacionado de zeros, ou seja, para as contagens nulas utiliza-se a função logística e para as não nulas, a ligação logarítmica (Fumes, 2009).

As estimativas dos parâmetros do modelo binomial negativo inflacionado de zeros na presença de covariáveis, segue a mesma proposta de Lambert (1992) para o modelo Poisson inflacionado de zeros (Garay et al., 2011). Portanto, a função de log-verossimilhança é dada da forma:

$$
\begin{aligned}
L(\boldsymbol{\gamma}, \boldsymbol{\beta}, \boldsymbol{y}, \boldsymbol{k}) &= \log\left(\prod_{y_i=0}^{n} p + (1-p)\left(\frac{k}{k+\mu}\right)^k \prod_{y_i>0}^{n} (1-p)\frac{\Gamma(y_i+k)}{\Gamma(k)y_i!}\left(\frac{k}{k+\mu}\right)^k \left(\frac{\mu}{k+\mu}\right)^{y_i}\right) \\
&= \sum_{y_i=0} \log\left[\left(\frac{p}{1-p} + \left(\frac{k}{k+\mu}\right)^k\right)(1-p)\right] + \\
&\quad \sum_{y_i>0} \log\left[(1-p)\frac{\Gamma(y_i+k)}{\Gamma(k)y_i!}\left(\frac{k}{k+\mu}\right)^k \left(\frac{\mu}{k+\mu}\right)^{y_i}\right]\Bigg)
\end{aligned}
$$

$$
\begin{aligned}
&= -\sum_{i=1}^{n} \log(1+e^{\boldsymbol{G_i\gamma}}) + \sum_{y_i>0}\left[k \log\left(\frac{k}{k+e^{\boldsymbol{X_i\beta}}}\right) + y_i \log\left(\frac{e^{\boldsymbol{X_i\beta}}}{k+e^{\boldsymbol{X_i\beta}}}\right)\right] + \\
&\quad \sum_{y_i=0} \log\left[e^{\boldsymbol{G_i\gamma}} + \left(\frac{k}{k+e^{\boldsymbol{X_i\beta}}}\right)^k\right] + \\
&\quad \sum_{y_i>0} [-\log\Gamma(y_i+1) - \log\Gamma(k) + \log\Gamma(y_i+k)] \qquad (3.29)
\end{aligned}
$$

onde $\boldsymbol{G_i}$ e $\boldsymbol{X_i}$ representam as i-ésimas linhas de $\boldsymbol{G}$ e $\boldsymbol{X}$.

De forma análoga ao modelo Poisson inflacionado de zeros, o algoritmo EM passa a ser utilizado para tal estimação. Para o algoritmo, há a introdução da variável aleatória $\boldsymbol{z}$, sendo uma variável indicadora da resposta, na qual assume 1 se a observação pertencer a parte inflacionada de zeros e 0 se a observação não pertencer a parte inflacionada (Fumes, 2009; Garay et al., 2011).

A etapa $E$ do algoritmo consiste em substituir $\boldsymbol{z}$ pela esperança condicional dada por

$\boldsymbol{y}, \boldsymbol{\gamma}^{(m)}$ e $\boldsymbol{\beta}^{(m)}$ (Wang et al., 2015).

$$z_i^{(m)} = \begin{cases} \left(1 + e^{-\boldsymbol{G_i}\boldsymbol{\gamma}^{(m)}} \left[\frac{\hat{k}^m}{e^{\boldsymbol{X_i}\hat{\boldsymbol{\beta}}^{(m)}} + \hat{k}^{(m)}}\right]^{\hat{k}^{(m)}}\right)^{-1}; & y_i = 0 \\ 0; \quad y_i > 0, \end{cases} \tag{3.30}$$

então a log-verossimilhança com os dados completos ($\boldsymbol{y}$,$\boldsymbol{z}$) será:

$$\begin{aligned} L(\gamma, \boldsymbol{\beta}, \boldsymbol{y}, \boldsymbol{k}) &= \sum_{i=1}^{n} \Bigg\{ z_i \boldsymbol{G_i}\boldsymbol{\gamma} - \log(1 + e^{\boldsymbol{G_i}\boldsymbol{\gamma}}) + \\ & (1 - z_i) \log \left( \frac{\Gamma(k + y_i)}{\Gamma(k)\Gamma(1 + y_i)} \left[\frac{e^{\boldsymbol{X_i}\boldsymbol{\beta}}}{k + e^{\boldsymbol{X_i}\boldsymbol{\beta}}}\right]^{y_i} \left[\frac{k}{k + e^{\boldsymbol{X_i}\boldsymbol{\beta}}}\right]^{k} \right) \Bigg\} \end{aligned} \tag{3.31}$$

Para demonstrar a ideia apresentada na Figura 1.8, considerando que $\boldsymbol{\gamma} = 0$, ou seja, $\boldsymbol{z} = 0$ tem-se:

$$L(\boldsymbol{\gamma}, \boldsymbol{\beta}, \boldsymbol{y}, \boldsymbol{k}) = \left\{ -n\log(1 + e^0) + \sum_{i=1}^{n} \log \left( \frac{\Gamma(k + y_i)}{\Gamma(k)\Gamma(1 + y_i)} \left[\frac{e^{\boldsymbol{X_i}\boldsymbol{\beta}}}{k + e^{\boldsymbol{X_i}\boldsymbol{\beta}}}\right]^{y_i} \left[\frac{k}{k + e^{\boldsymbol{X_i}\boldsymbol{\beta}}}\right]^{k} \right) \right\} \tag{3.32}$$

Análogo a Poisson inflacionada de zeros, nota-se que $-n\log(1 + e^0)$ é uma constante que não influencia na estimação do parâmetro $\boldsymbol{\beta}$, sendo assim, a expressão se resume aproximadamente:

$$\begin{aligned} L(\boldsymbol{\gamma}, \boldsymbol{\beta}, \boldsymbol{y}, \boldsymbol{k}) &= -n\log(1 + e^0) + \sum_{i=1}^{n} \log \left( \frac{\Gamma(k + y_i)}{\Gamma(k)\Gamma(1 + y_i)} \left[\frac{e^{\boldsymbol{X_i}\boldsymbol{\beta}}}{k + e^{\boldsymbol{X_i}\boldsymbol{\beta}}}\right]^{y_i} \left[\frac{k}{k + e^{\boldsymbol{X_i}\boldsymbol{\beta}}}\right]^{k} \right) \\ &\approx \sum_{i=1}^{n} \log \left( \frac{\Gamma(k + y_i)}{\Gamma(k)\Gamma(1 + y_i)} \left[\frac{e^{\boldsymbol{X_i}\boldsymbol{\beta}}}{k + e^{\boldsymbol{X_i}\boldsymbol{\beta}}}\right]^{y_i} \left[\frac{k}{k + e^{\boldsymbol{X_i}\boldsymbol{\beta}}}\right]^{k} \right) \\ &\approx \sum_{i=1}^{n} \left[ \log \left\{ \frac{\Gamma(k + y_i)}{\Gamma(k)\Gamma(1 + y_i)} \right\} + k\log(k) + y_i \log(e^{\boldsymbol{X_i}\boldsymbol{\beta}}) - (y_i + k)\log(k + e^{\boldsymbol{X_i}\boldsymbol{\beta}}) \right] \end{aligned} \tag{3.33}$$

Portanto, pode-se observar que esta log-verossimilhança (3.33) se torna a log-verossimilhança do modelo de regressão binomial negativo, expresso em (2.42)

Com base em $\boldsymbol{z}$, pode-se maximizar a log-verossimilhança e perceber que a expressão se resume à soma da log-verossimilhança da regressão logística não ponderada de $\boldsymbol{z}^{(m)}$ em $\boldsymbol{G}$ (sendo um termo que não envolve $\boldsymbol{\beta}$) e a log-verossimilhança da regressão de binomial negativa ponderada de $\boldsymbol{y}$ em $\boldsymbol{X}$ (um termo que não envolve $\boldsymbol{\gamma}$) (Garay et al., 2011).

A segunda etapa da iteração $(m+1)$ consiste em obter $\boldsymbol{\gamma}^{(m+1)}$, maximizando $L_c(\boldsymbol{\gamma}, \boldsymbol{y}, \boldsymbol{z})$ por meio da regressão logística não ponderada de $\boldsymbol{z}^{(m)}$ em $\boldsymbol{G}$. Em seguida obter $\boldsymbol{\beta}^{(m+1)}$, maximizando $L_c(\boldsymbol{\beta}, \boldsymbol{y}, \boldsymbol{z})$, por meio da regressão binomial negativa ponderada com pesos $(1 - z_i)$ (Garay et al., 2011).

Como foi feito na no Algoritmo da regressão Poisson inflacionada de zeros, pode-se ver no Algoritmo 6 a estimação para a regressão binomial negativa inflacionada de zeros. Neste caso, observa-se a mesma estrutura feita para o Algoritmo da Poisson inflacionada de zeros, a diferença consiste na utilização dos modelos logístico e binomial negativo, compondo assim, o modelo binomial negativo inflacionado de zeros.

A matriz de informação observada correspondente ao modelo binomial negativo inflacionado de zeros, considerando os parâmetros $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$ e $\boldsymbol{k}$ é dada por:

$$\boldsymbol{I}(\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{k}) = \begin{pmatrix} I_{11} & I_{12} & I_{13} \\ I_{21} & I_{22} & I_{23} \\ I_{31} & I_{32} & I_{33} \end{pmatrix} \tag{3.34}$$

E a matriz inversa é dada:

$$\boldsymbol{I}(\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{k})^{-1} = \begin{pmatrix} Var(\hat{\boldsymbol{\beta}}) & Cov(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}}) & Cov(\hat{\boldsymbol{\beta}}, \hat{k}) \\ Cov(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}}) & Var(\hat{\boldsymbol{\gamma}}) & Cov(\hat{k}, \hat{\boldsymbol{\gamma}}) \\ Cov(\hat{\boldsymbol{\beta}}, \hat{k}) & Cov(\hat{k}, \hat{\boldsymbol{\gamma}}) & Var(\hat{k}) \end{pmatrix} \tag{3.35}$$

**Algoritmo 6:** Algoritmo EM - binomial negativa inflacionada de zeros

**Entrada:** $\boldsymbol{\beta}, \gamma$

Estimar $\boldsymbol{\beta}$ para $y > 0$, por meio regressão binomial negativa.

$\boldsymbol{\gamma_0} = (\sum_{i=1}^{n} I_{(y_i=0)} - \sum_{i=1}^{n}(k/(\mu_i + k))^k)/n$

Estimar $\boldsymbol{\gamma} = \log(\boldsymbol{\gamma_0}/(1 - \boldsymbol{\gamma_0}))$

$DiffD = 1, \quad OldD = 0$

**enquanto** *($abs(DiffD) > 10^{-6}$)* **faça**

- **Passo E:** Estimar a esperança condicional $z_i$
- **se** $y_i = 0$ **então**
  - $$z_i = \left(1 + e^{-\boldsymbol{G_i}\boldsymbol{\gamma}}\left[\frac{\hat{k}}{e^{\boldsymbol{X_i}\hat{\boldsymbol{\beta}}} + \hat{k}}\right]^{\hat{k}}\right)^{-1}$$
- **senão**
  - 0
- **fim**
- **Passo M para $\boldsymbol{\beta}$:** Estimação do modelo binomial negativo ponderado por $(1 - \boldsymbol{z})$, minimizando a $Deviance$ $(D_1)$ :
- { $\boldsymbol{\eta} = \boldsymbol{X\beta} + offset$
- $\boldsymbol{\mu} = \exp(\boldsymbol{\eta})$
- $\boldsymbol{M} = \frac{\Gamma(k+\boldsymbol{y})}{\Gamma(k)\Gamma(\boldsymbol{y}+1)}\left(\frac{\boldsymbol{\mu}}{\boldsymbol{\mu}+k}\right)^{\boldsymbol{y}}\left(\frac{k}{\boldsymbol{\mu}+k}\right)^{k}$
- $D_1 = \sum_{i=1}^{n}[(1 - z_i)(\log(M_i))$
- }
- **Passo M para $\boldsymbol{\gamma}$:** Estimação do modelo logístico não ponderado, utilizando $z_i$ como variável resposta, minimizando a $Deviance$ $(D_2)$ :
- { $\boldsymbol{\eta} = \boldsymbol{G\gamma}$
- $D_2 = \sum_{i=1}^{n}(z_i\eta_i - \sum_{i=1}^{n}(\log(1 + \exp(\eta_i)))$
- }
- **Maximização:** $OldD = D$
- $D = D_1 + D_2$
- $DiffD = OldD - D$

**fim**

Para obter os erros padrão para a matriz de informação observada, deve-se obter as segundas derivadas, referentes aos parâmetros $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$ e $\boldsymbol{k}$. Sendo assim, as equações com as segundas derivadas são dadas por:

$$
\begin{aligned}
I_{11} = \frac{\partial^2 L(\boldsymbol{\gamma}, \boldsymbol{\beta}, \boldsymbol{y}, \boldsymbol{k})}{\partial \beta_j \partial \beta_r} &= \sum_{i;y=0} -\frac{k^2([f_1(\boldsymbol{x_i})]^k f_2(\boldsymbol{x_i}))^2 X_{ij} X_{ir}}{h(\boldsymbol{G_i}, \boldsymbol{X_i})^2} + \frac{k^2 [f_1(\boldsymbol{x_i})]^k [f_2(\boldsymbol{x_i})]^2 X_{ij} X_{ir} \left(1 - \frac{1}{e^{\boldsymbol{X_i}\boldsymbol{\beta}}}\right)}{h(\boldsymbol{G_i}, \boldsymbol{X_i})} \\
&- \sum_{i;y>0} \left( \frac{k X_{ij} X_{ir} e^{\boldsymbol{X_i}\boldsymbol{\beta}}}{e^{\boldsymbol{X_i}\boldsymbol{\beta}} + k} \left[ \frac{y_i - e^{\boldsymbol{X_i}\boldsymbol{\beta}}}{e^{\boldsymbol{X_i}\boldsymbol{\beta}} + k} + 1 \right] \right)
\end{aligned} \tag{3.36}
$$

$$
I_{22} = \frac{\partial^2 L(\boldsymbol{\gamma}, \boldsymbol{\beta}, \boldsymbol{y}, \boldsymbol{k})}{\partial \gamma_j \partial \gamma_r} = \sum_{i;y=0} \frac{G_{ij} G_{ir} e^{\boldsymbol{G_i}\boldsymbol{\gamma}} [f_1(\boldsymbol{x_i})]^k}{h(\boldsymbol{G_i}, \boldsymbol{X_i})^2} - \sum_{i=1}^{n} \frac{e^{\boldsymbol{G_i}\boldsymbol{\gamma}} G_{ij} G_{ir}}{[1 + e^{\boldsymbol{G_i}\boldsymbol{\gamma}}]^2} \tag{3.37}
$$

$$
I_{12} = I_{21} = \frac{\partial^2 L(\boldsymbol{\gamma}, \boldsymbol{\beta}, \boldsymbol{y}, \boldsymbol{k})}{\partial \gamma_j \partial \beta_r} = \sum_{i;y=0} \frac{k G_{ij} X_{ir} e^{\boldsymbol{G_i}\boldsymbol{\gamma}} [f_1(\boldsymbol{x_i})]^k f_2(\boldsymbol{x_i})}{h(\boldsymbol{G_i}, \boldsymbol{X_i})^2} \tag{3.38}
$$

$$
\begin{aligned}
I_{33} &= \frac{\partial^2 L(\boldsymbol{\gamma}, \boldsymbol{\beta}, \boldsymbol{y}, \boldsymbol{k})}{\partial k^2} = \sum_{i;y>0} \left( \psi'(k + y_i) - \psi'(k) + k^{-1} + \frac{(y_i + k)}{(e^{\boldsymbol{X_i}\boldsymbol{\beta}} + k)^2} - \frac{2}{(e^{\boldsymbol{X_i}\boldsymbol{\beta}} + k)} \right) \\
&+ \sum_{i;y=0} \left( \frac{([f_1(\boldsymbol{x_i})]^k [\log(f_1(\boldsymbol{x_i})) + f_2(\boldsymbol{x_i})])^2}{h(\boldsymbol{G_i}, \boldsymbol{X_i})} \left(1 - \frac{1}{h(\boldsymbol{G_i}, \boldsymbol{X_i})}\right) + \frac{[f_1(\boldsymbol{x_i}]^k [k^{-1} [f_2(\boldsymbol{x_i})]^2]}{h(\boldsymbol{G_i}, \boldsymbol{X_i})} \right)
\end{aligned} \tag{3.39}
$$

$$
\begin{aligned}
I_{13} = I_{31} = \frac{\partial^2 L(\boldsymbol{\gamma}, \boldsymbol{\beta}, \boldsymbol{y}, \boldsymbol{k})}{\partial k \partial \beta_j} &= \sum_{i;y=0} \frac{[f_1(\boldsymbol{x_i})]^{2k+1} e^{\boldsymbol{X_i}\boldsymbol{\beta}} X_{ij} [\log(f_1(\boldsymbol{x_i})) + f_2(\boldsymbol{x_i})]}{h(\boldsymbol{G_i}, \boldsymbol{X_i})^2} \\
&- \sum_{i;y=0} \frac{[f_1(\boldsymbol{x_i})]^k X_{ij} (k f_2(\boldsymbol{x_i}) [\log(f_1(\boldsymbol{x_i})) + f_2(\boldsymbol{x_i})] - [f_2(\boldsymbol{x_i})]^2)}{h(\boldsymbol{G_i}, \boldsymbol{X_i})} \\
&+ \sum_{i;y>0} \left( f_2(\boldsymbol{x_i}) X_{ij} \left( \frac{y_i - e^{\boldsymbol{X_i}\boldsymbol{\beta}}}{e^{\boldsymbol{X_i}\boldsymbol{\beta}} + k} \right) \right)
\end{aligned} \tag{3.40}
$$

$$I_{23} = I_{32} = \frac{\partial^2 L(\boldsymbol{\gamma}, \boldsymbol{\beta}, \boldsymbol{y}, \boldsymbol{k})}{\partial k \partial \gamma_j} = \sum_{i;y=0} -\frac{[f_1(\boldsymbol{x_i})]^k [\log(f_1(\boldsymbol{x_i})) + f_2(\boldsymbol{x_i})]}{h(\boldsymbol{G_i}, \boldsymbol{X_i})^2} e^{\boldsymbol{G_i \gamma}} G_{ij} \quad (3.41)$$

onde $f_1(\boldsymbol{x_i}) = \frac{k}{k+e^{\boldsymbol{X_i\beta}}}$, $f_2(\boldsymbol{x_i}) = \frac{e^{\boldsymbol{X_i\beta}}}{k+e^{\boldsymbol{X_i\beta}}}$, $h(\boldsymbol{G_i}, \boldsymbol{X_i}) = e^{\boldsymbol{G_i\gamma}} + \left(\frac{k}{k+e^{\boldsymbol{X_i\beta}}}\right)^k$, $\psi(k) = \frac{\partial \log \Gamma(k)}{\partial k}$ e $\psi'(k) = \frac{\partial \psi(k)}{\partial k} = \frac{\partial^2 \log \Gamma(k)}{\partial k^2}$.

As demonstrações dessas derivadas (3.36), (3.37), (3.38), (3.39), (3.40) e (3.41) podem ser vistas no Apêndice C.

Ao aplicar o $\lim_{k\to\infty}$, pode-se notar que a binomial negativa inflacionada de zeros irá se reduzir em uma Poisson inflacionada de zeros, como a seguir:

$$\begin{aligned}
\frac{\partial^2 L(\boldsymbol{\gamma}, \boldsymbol{\beta}, \boldsymbol{y}, \boldsymbol{k})}{\partial \beta_j \partial \beta_r} &= \lim_{k\to\infty} \sum_{i;y=0} -\frac{k^2 \left( [\frac{k}{k+e^{\boldsymbol{X_i\beta}}}]^k \frac{e^{\boldsymbol{X_i\beta}}}{k+e^{\boldsymbol{X_i\beta}}} \right)^2 X_{ij} X_{ir}}{\left( e^{\boldsymbol{G_i\gamma}} + \left[\frac{k}{k+e^{\boldsymbol{X_i\beta}}}\right]^k \right)^2} \\
&+ \sum_{i;y=0} \frac{k^2 [\frac{k}{k+e^{\boldsymbol{X_i\beta}}}]^k [\frac{e^{\boldsymbol{X_i\beta}}}{k+e^{\boldsymbol{X_i\beta}}}]^2 X_{ij} X_{ir} \left(1 - \frac{1}{e^{\boldsymbol{X_i\beta}}}\right)}{e^{\boldsymbol{G_i\gamma}} + [\frac{k}{k+e^{\boldsymbol{X_i\beta}}}]^k} \\
&- \sum_{i;y>0} \left( \frac{k X_{ij} X_{ir} e^{\boldsymbol{X_i\beta}}}{e^{\boldsymbol{X_i\beta}} + k} \left[ \frac{y_i - e^{\boldsymbol{X_i\beta}}}{e^{\boldsymbol{X_i\beta}} + k} + 1 \right] \right) \\
&= \sum_{i;y=0} \frac{\exp(-e^{\boldsymbol{X_i\beta}}) e^{\boldsymbol{X_i\beta}} X_{ij} X_{ir} \left( e^{\boldsymbol{X_i\beta}} - \frac{e^{\boldsymbol{X_i\beta}} \exp(-e^{\boldsymbol{X_i\beta}})}{\exp(-e^{\boldsymbol{X_i\beta}}) + e^{\boldsymbol{G_i\gamma}}} - 1 \right)}{e^{\boldsymbol{G_i\gamma}} + \exp(-e^{\boldsymbol{X_i\beta}})} \\
&- \sum_{i;y>0} X_{ij} X_{ir} e^{\boldsymbol{X_i\beta}} \qquad (3.42)
\end{aligned}$$

Pode-se perceber que (3.42) se torna igual a (3.14).

$$
\begin{aligned}
\frac{\partial^2 L(\boldsymbol{\gamma},\boldsymbol{\beta},\boldsymbol{y},\boldsymbol{k})}{\partial\gamma_j\partial\gamma_r} &= \lim_{k\to\infty}\sum_{i;y=0}\frac{G_{ij}G_{ir}e^{\boldsymbol{G_i\gamma}}[\frac{k}{k+e^{\boldsymbol{X_i\beta}}}]^k}{\left(e^{\boldsymbol{G_i\gamma}}+[\frac{k}{k+e^{\boldsymbol{X_i\beta}}}]^k\right)^2}-\sum_{i=1}^{n}\frac{e^{\boldsymbol{G_i\gamma}}G_{ij}G_{ir}}{[1+e^{\boldsymbol{G_i\gamma}}]^2}\\
&= \sum_{i;y=0}\frac{G_{ij}G_{ir}e^{\boldsymbol{G_i\gamma}}\exp(-e^{\boldsymbol{X_i\beta}})}{[e^{\boldsymbol{G_i\gamma}}+\exp(-e^{\boldsymbol{X_i\beta}})]^2}-\sum_{i=1}^{n}\frac{e^{\boldsymbol{G_i\gamma}}G_{ij}G_{ir}}{[1+e^{\boldsymbol{G_i\gamma}}]^2} \qquad (3.43)
\end{aligned}
$$

Também percebe-se que (3.43) se torna igual a (3.15).

$$
\begin{aligned}
\frac{\partial^2 L(\boldsymbol{\gamma},\boldsymbol{\beta},\boldsymbol{y},\boldsymbol{k})}{\partial\gamma_j\partial\beta_r} &= \lim_{k\to\infty}\sum_{i;y=0}\frac{kG_{ij}X_{ir}e^{\boldsymbol{G_i\gamma}}[\frac{k}{k+e^{\boldsymbol{X_i\beta}}}]^k\frac{e^{\boldsymbol{X_i\beta}}}{k+e^{\boldsymbol{X_i\beta}}}}{\left(e^{\boldsymbol{G_i\gamma}}+[\frac{e^{\boldsymbol{X_i\beta}}}{k+e^{\boldsymbol{X_i\beta}}}]^k\right)^2}\\
&= \sum_{i;y=0}\frac{G_{ij}X_{ir}e^{\boldsymbol{X_i\beta}}\exp(-e^{\boldsymbol{X_i\beta}})e^{\boldsymbol{G_i\gamma}}}{[e^{\boldsymbol{G_i\gamma}}+\exp(-e^{\boldsymbol{X_i\beta}})]^2} \qquad (3.44)
\end{aligned}
$$

E por último, percebe-se que (3.44) se torna igual a (3.16), cofirmando que a binomial negativa inflacionada de zeros se torna a Poisson inflacionada de zeros.

Seguindo a ideia feita em (3.17), para mostrar que a binomial negativa inflacionada de zeros irá se reduzir a uma binomial negativa, pode-se considerar $\boldsymbol{\gamma}=0$, ou equivalentemente, todas as equações referentes ao estado $y_i=0$ sejam iguais a 0 e $y_i>0$ passa a ser $y_i\geq 0$, isto é, considerando todos os dados, e substituir esta expressão nas segundas derivadas. Sendo assim, a segunda derivada, em função do parâmetro $\boldsymbol{\beta}$ pode ser expressa da forma:

$$
\begin{aligned}
\frac{\partial^2 L(\boldsymbol{\gamma},\boldsymbol{\beta},\boldsymbol{y},\boldsymbol{k})}{\partial k\partial\gamma_j} &= \sum_{i;y=0}-\frac{\left(\frac{k}{e^{\boldsymbol{X_i\beta}}+k}\right)^k\left[\log\left(\left(\frac{k}{e^{\boldsymbol{X_i\beta}}+k}\right)+\left(\frac{e^{\boldsymbol{X_i\beta}}}{e^{\boldsymbol{X_i\beta}}+k}\right)\right)\right]}{\left(\frac{k}{e^{\boldsymbol{G_i\gamma}}+e^{\boldsymbol{X_i\beta}}+k}\right)^k}e^{\boldsymbol{G_i\gamma}}\\
&= 0 \qquad (3.45)
\end{aligned}
$$

$$\frac{\partial^2 L(\boldsymbol{\gamma},\boldsymbol{\beta},\boldsymbol{y},\boldsymbol{k})}{\partial\gamma_j\partial\gamma_r} = \sum_{i;y=0} \frac{G_{ij}G_{ir}e^{\boldsymbol{G_i\gamma}}\left[\frac{k}{k+e^{\boldsymbol{X_i\beta}}}\right]^k}{\left(\left[\frac{k}{k+e^{\boldsymbol{X_i\beta}}}\right]^k + e^{\boldsymbol{X_i\beta}}\right)^2} - \sum_{i=1}^{n}\frac{e^{\boldsymbol{G_i\gamma}}G_{ij}G_{ir}}{[1+e^{\boldsymbol{G_i\gamma}}]^2} = 0 \qquad (3.46)$$

$$\frac{\partial^2 L(\boldsymbol{\gamma},\boldsymbol{\beta},\boldsymbol{y},\boldsymbol{k})}{\partial\gamma_j\partial\beta_r} = \sum_{i;y=0} \frac{kG_{ij}X_{ir}e^{\boldsymbol{G_i\gamma}}[\frac{k}{k+e^{\boldsymbol{X_i\beta}}}]^k\frac{e^{\boldsymbol{X_i\beta}}}{k+e^{\boldsymbol{X_i\beta}}}}{\left(e^{\boldsymbol{G_i\gamma}}+[\frac{e^{\boldsymbol{X_i\beta}}}{k+e^{\boldsymbol{X_i\beta}}}]^k\right)^2} = 0 \qquad (3.47)$$

$$\begin{aligned}\frac{\partial^2 L(\boldsymbol{\gamma},\boldsymbol{\beta},\boldsymbol{y},\boldsymbol{k})}{\partial\beta_j\partial\beta_r} &= \sum_{i;y=0} -\frac{k^2\left(\left[\frac{k}{k+e^{\boldsymbol{X_i\beta}}}\right]^k\frac{e^{\boldsymbol{X_i\beta}}}{k+e^{\boldsymbol{X_i\beta}}}\right)^2 X_{ij}X_{ir}}{\left(e^{\boldsymbol{G_i\gamma}}+\left[\frac{k}{k+e^{\boldsymbol{X_i\beta}}}\right]^k\right)^2} \\ &+ \sum_{i;y=0}\frac{k^2\left[\frac{k}{k+e^{\boldsymbol{X_i\beta}}}\right]^k\left[\frac{e^{\boldsymbol{X_i\beta}}}{k+e^{\boldsymbol{X_i\beta}}}\right]^2 X_{ij}X_{ir}\left(1-\frac{1}{e^{\boldsymbol{X_i\beta}}}\right)}{e^{\boldsymbol{G_i\gamma}}+\left[\frac{k}{k+e^{\boldsymbol{X_i\beta}}}\right]^k} \\ &- \sum_{i;y>0}\left(\frac{kX_{ij}X_{ir}e^{\boldsymbol{X_i\beta}}}{e^{\boldsymbol{X_i\beta}}+k}\left[\frac{y_i-e^{\boldsymbol{X_i\beta}}}{e^{\boldsymbol{X_i\beta}}+k}+1\right]\right) \\ &= -\sum_{i=1}^{n}\left(\frac{kX_{ij}X_{ir}e^{\boldsymbol{X_i\beta}}}{e^{\boldsymbol{X_i\beta}}+k}\left(\frac{y_i-e^{\boldsymbol{X_i\beta}}}{e^{\boldsymbol{X_i\beta}}+k}+1\right)\right) \end{aligned} \qquad (3.48)$$

Pode-se observar que (3.48) resulta na segunda derivada da log-veromilhança do modelo binomial negativo (2.42) com relação a $\boldsymbol{\beta}$. Portanto, se o parâmetro $\boldsymbol{\gamma} = 0$, a expressão retorna à uma binomial negativa. Portanto, tal expressão (3.48) resulta em (2.49).

$$
\begin{aligned}
\frac{\partial^2 L(\boldsymbol{\gamma},\boldsymbol{\beta},\boldsymbol{y},\boldsymbol{k})}{\partial k \partial \beta_j} &= \sum_{i;y=0} \frac{\left(\frac{k}{e^{\boldsymbol{X_i\beta}}+k}\right)^{2k+1} e^{\boldsymbol{X_i\beta}} X_{ij}\left[\log\left(\frac{k}{e^{\boldsymbol{X_i\beta}}+k}\right)+\left(\frac{e^{\boldsymbol{X_i\beta}}}{e^{\boldsymbol{X_i\beta}}+k}\right)\right]}{\left(e^{\boldsymbol{G_i\gamma}}+\left(\frac{k}{e^{\boldsymbol{X_i\beta}}+k}\right)^{k}\right)^2} \\
&- \sum_{i;y=0} \frac{\left(\frac{k}{e^{\boldsymbol{X_i\beta}}+k}\right)^{k} X_{ij}\left(\frac{e^{\boldsymbol{X_i\beta}}k}{e^{\boldsymbol{X_i\beta}}+k}\left[\log\left(\frac{k}{e^{\boldsymbol{X_i\beta}}+k}\right)+\left(\frac{e^{\boldsymbol{X_i\beta}}}{e^{\boldsymbol{X_i\beta}}+k}\right)\right]-\left(\frac{e^{\boldsymbol{X_i\beta}}}{e^{\boldsymbol{X_i\beta}}+k}\right)^2\right)}{e^{\boldsymbol{G_i\gamma}}+\left(\frac{k}{e^{\boldsymbol{X_i\beta}}+k}\right)^{k}} \\
&+ \sum_{i;y>0}\left(\frac{e^{\boldsymbol{X_i\beta}}X_{ij}}{e^{\boldsymbol{X_i\beta}}+k}\left(\frac{y_i-e^{\boldsymbol{X_i\beta}}}{e^{\boldsymbol{X_i\beta}}+k}\right)\right) \\
&= \sum_{i=1}^{n}\left(\frac{e^{\boldsymbol{X_i\beta}}X_{ij}}{e^{\boldsymbol{X_i\beta}}+k}\left(\frac{y_i-e^{\boldsymbol{X_i\beta}}}{e^{\boldsymbol{X_i\beta}}+k}\right)\right) \qquad (3.49)
\end{aligned}
$$

$$
\begin{aligned}
\frac{\partial^2 L(\boldsymbol{\gamma},\boldsymbol{\beta},\boldsymbol{y},\boldsymbol{k})}{\partial k^2} &= \sum_{i;y=0} \frac{\left(\left(\frac{k}{e^{\boldsymbol{X_i\beta}}+k}\right)^{k}\left[\log\left(\frac{k}{e^{\boldsymbol{X_i\beta}}+k}\right)+\left(\frac{e^{\boldsymbol{X_i\beta}}}{e^{\boldsymbol{G_i\gamma}}+k}\right)\right]\right)^2}{e^{\boldsymbol{G_i\gamma}}+\left(\frac{k}{e^{\boldsymbol{X_i\beta}}+k}\right)^{k}} \times \\
&\left(1-\frac{1}{e^{\boldsymbol{G_i\gamma}}+\left(\frac{k}{e^{\boldsymbol{X_i\beta}}+k}\right)^{k}}\right)+\sum_{i;y=0}\frac{\left(\frac{k}{e^{\boldsymbol{X_i\beta}}+k}\right)^{k}+\left[k^{-1}\left(\frac{e^{\boldsymbol{X_i\beta}}}{e^{\boldsymbol{X_i\beta}}+k}\right)^2\right]}{e^{\boldsymbol{G_i\gamma}}+\left(\frac{k}{e^{\boldsymbol{X_i\beta}}+k}\right)^{k}} \\
&+ \sum_{i;y>0}\left(\psi'(k+y_i)-\psi'(k)+k^{-1}+\frac{(y_i+k)}{(e^{\boldsymbol{X_i\beta}}+k)^2}-\frac{2}{(e^{\boldsymbol{X_i\beta}}+k)}\right) \\
&= \sum_{i=1}^{n}\left(\psi'(k+y_i)-\psi'(k)+k^{-1}+\frac{(y_i+k)}{(e^{\boldsymbol{X_i\beta}}+k)^2}-\frac{2}{(e^{\boldsymbol{X_i\beta}}+k)}\right) \qquad (3.50)
\end{aligned}
$$

Por último, nota-se que (3.50) torna-se a expressão em (2.50), que representa a segunda derivada da log-veromilhança do modelo binomial negativo (2.42), considerando o parâmetro $k$.

Portanto, como na Poisson inflacionada de zeros, desconsiderar a parte $\sum_{y_i=0}\{\}$ ou $0\times\sum_{y_i=0}\{\}$ e fazer com que a parte $\sum_{y_i>0}\{\}$ se torne $\sum_{y_i\geq 0}\{\}$ é uma forma computacional para fazer com que (3.35) possa gerar a variância de $\boldsymbol{\beta}$ e $k$ na regressão binomial negativa.

# Capítulo 4

# Regressão Geograficamente Ponderada

## 4.1 Introdução

Os modelos de regressão espacial buscam determinar a relação entre variáveis, considerando a localização do espaço do fenômeno em estudo. Sendo assim, a análise espacial de dados consiste em estudos quantitativos de fenômenos localizados em deteminado espaço, onde a localização dos dados se torna fundamental para a interpretação dos resultados (Druck et al., 2004).

Na condição de que os dados espacias não sejam estacionários, ou seja, dados cuja distribuição de probabilidade varia no espaço, o modelo global pode gerar estimativas incertas e não atender aos resultados com precisão, sendo assim, a Regressão Geograficamente Ponderada (RGP) propõe contornar essa limitação dos modelos globais, sendo uma extensão do modelo de regressão linear tradicional, permitindo que ocorra variações nos locais dos parâmetros (Fotheringham et al., 2002).

O objetivo deste Capítulo é apresentar a Regressão Geograficamente Ponderada (RGP), proposta por Brunsdon et al. (1996), sendo uma técnica não paramétrica utilizada para modelagem de dados espaciais não estacionários. Uma breve apresentação da Regressão Geograficamente Ponderada para dados com distribuição Normal será feita, e em seguida, a Regressão Binomial (ou Logística) Geograficamente Ponderada será detalhada, assim como a Regressão Poisson Geograficamente Ponderada e por fim a a Regressão Binomial Negativa Geograficamente Ponderada.

## 4.2 Regressão geograficamente ponderada

Na análise de dados espaciais, a Regressão Geograficamente Ponderada surge com o intuito de descrever a relação entre as variáveis quando há variação no espaço. Portanto, a RGP realiza um ajuste local para cada ponto da região de estudo com base nas observações mais próximas e permite que os parâmetros da regressão possam variar de acordo com a localização de forma contínua (Fotheringham et al., 2002).

Ao definir o modelo, considera-se que existam $n$ observações e $k$ variáveis explicativas, com isso o modelo RGP é definido pela expressão (Fotheringham et al., 2002):

$$y_j = \sum_k \beta_k(u_i, v_i)x_{jk} + \varepsilon_j, \quad j = 1, ..., n. \tag{4.1}$$

onde $(u_i, v_i)$ compõe as coordenadas do $i-$ésimo ponto no espaço; $\beta_k(u_i, v_i)$ é o parâmetro para a $k-$ésima variável explicativa, em função do local da $i-$ésima observação, $x_{jk}$ é o valor da $k-$ésima variável explicativa para a $j$-ésima observação e $\varepsilon_j$ é o erro associado a $j-$ésima observação, supondo de que sejam independentes e indenticamente distribuídos, ou seja, $\varepsilon_j \sim N(0, \sigma^2)$.

Para se obter os parâmetros $\beta_k(u_i, v_i)$ utiliza-se as observações próximas ao ponto $i$, ou seja, onde se calculam as estimativas dos parâmetros (pontos de regressão). Em (4.1), a variância dos erros é considerada fixa, ainda que os parâmetros da regressão obtenham variação espacial.

O método de mínimos quadrados pode ser utilizado para a estimação dos parâmetros $\boldsymbol{\beta}(u_i, v_i)$, ou de forma simplificada $\boldsymbol{\beta}(i)$, atribuindo pesos $\boldsymbol{W}(u_i, v_i)$, ou de forma simplificada $\boldsymbol{W}(i)$ nas diferentes observações, sendo assim, tem - se a expressão:

$$\hat{\boldsymbol{\beta}}(i) = (\boldsymbol{X^T W}(i)\boldsymbol{X})^{-1}\boldsymbol{X^T W}(i)\boldsymbol{Y} \tag{4.2}$$

onde $\hat{\boldsymbol{\beta}}(i)$ representa o vetor de tamanho $k$ com as estimativas para os parâmetros no local $i$; $\boldsymbol{W}(i)$ representa a matriz de ponderação para o local $i$, com dimensão $n \times n$; $\boldsymbol{X}$ representa a matriz $n \times k$, com o valor das $k$ variáveis explicativas para as $n$ observações e por fim $\boldsymbol{y}$ representa o vetor de tamanho $n$, com o valor das variáveis respostas no ponto.

Para a matriz de ponderação $\boldsymbol{W}(i)$ em (4.2), deve-se calcular os valores para cada ponto de localidade $i$. Sendo assim, a função de ponderação espacial é a que determina os pesos $w_{ij}$ da matriz $\boldsymbol{W}(i)$ que serão calculados. A Figura 4.1 mostra um exemplo da função de ponderação espacial.

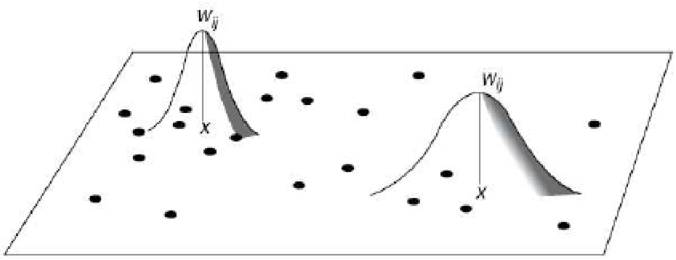



Figura 4.1: Função de ponderação espacial
Fonte: Fotheringham et al. (2002)

Fotheringham et al. (2002) propuseram algumas possibilidades de escolha para a função de ponderação, sendo:

**1.** Função de ponderação espacial, considerando a distância $d_{ij}$ entre o ponto amostral $j$ e o ponto de estimação $i$:

$$w_{ij} = \begin{cases} 1, & d_{ij} < d \\ 0, & \text{Caso contrário} \end{cases} \tag{4.3}$$

onde $d$ é fixo de forma arbitrária. Ao utilizar esta função, não há um decaimento da função para maiores distâncias, pois há uma descontinuidade considerável quando há proximidade com a distância $d$.

**2.** Função de ponderação espacial para casos com descontinuidade:

$$w_{ij} = e^{-\frac{d_{ij}^2}{2b^2}} \tag{4.4}$$

a Equação (4.4) é uma alternativa para não ocorrer descontinuidade, sendo uma função contínua exponencial quadrática e assume valores descrescentes quanto maior for a distância $d_{ij}$, na forma da distribuição Normal.

**3.** Função de ponderação espacial, mistura das estruturas (4.3) e (4.4):

$$w_{ij} = \begin{cases} \left(1 - \left(\frac{d_{ij}}{b}\right)^2\right)^2, & d_{ij} < b \\ 0, & \text{Caso contrário} \end{cases} \tag{4.5}$$

onde $b$ é conhecido como parâmetro de suavização (ou do inglês *bandwidth*) e quanto menor o seu valor, maior será o decaimento da função. Uma das funcionalidades da função bi-quadrática (4.5) é a facilidade no esforço computacional.

Existem casos em que os dados não estão espaçados de forma igualitária na região ou ainda, se concentram em áreas com tamanhos diferentes. Nestes casos, o recomendável é que o parâmetro de suavização da função de ponderação espacial possa variar segundo a disposição dos dados observados.

Na Figura 4.1, Fotheringham et al. (2002) mostram que áreas que possuem uma larga escala de pontos, utilizam a função $kernel$ (maior variância) com parâmetro de suavização menor e áreas com baixa escala de pontos utilizam um parâmetro de suavização maior. Alguns exemplos para funções de ponderação espacial, levando em conta a questão da dispersão serão descritos a seguir:

$$w_{ij} = \begin{cases} 1, & \text{se } j \text{ é um dos } R \text{ vizinhos mais próximos de } i \\ 0, & \text{Caso contrário} \end{cases} \tag{4.6}$$

onde $R$ é o parâmetro referente ao número de pontos incluídos na calibração do modelo. Outra opção é:

$$w_{ij} = \begin{cases} \left(1 - \left(\frac{d_{ij}}{b}\right)^2\right)^2, & \text{se } j \text{ é um dos } R \text{ vizinhos mais próximos de } i \\ 0, & \text{Caso contrário} \end{cases} \tag{4.7}$$

onde $R$ é o parâmetro referente ao número de pontos incluídos na calibração do modelo e $b$ representa a distância de $i$ até o $R-$ ésimo vizinho. Por fim pode-se utilizar:

$$w_{ij} = e^{-\frac{R_{ij}^2}{b}} \tag{4.8}$$

Com base na Equação (4.2), pode-se obter o erro padrão das estimativas locais do modelo RGP, ao reescrever (4.2) como:

$$\hat{\boldsymbol{\beta}}(i) = \boldsymbol{C}\boldsymbol{y} \tag{4.9}$$

onde $\boldsymbol{C} = (\boldsymbol{X^T}\boldsymbol{W}(i)\boldsymbol{X})^{-1}\boldsymbol{X^T}\boldsymbol{W}(i)$.

Com isso, a variância das estimativas dos parâmetros é dada da forma:

$$\widehat{Var}(\hat{\boldsymbol{\beta}}_{\boldsymbol{i}}) = \boldsymbol{C}\boldsymbol{C^T}\hat{\boldsymbol{\sigma}}^{\boldsymbol{2}} \tag{4.10}$$

sendo (4.10) análoga à regressão clássica quando $w_{ij} = 1, \forall i, j$ e $\hat{\sigma}^2$ é a soma de quadrados dos resíduos normalizados da regressão local, dada da forma:

$$\hat{\sigma}^2 = \sum_{j=1}^{n} \frac{(y_j - \hat{y_j})^2}{n - 2\nu_1 + \nu_2} \tag{4.11}$$

com $\nu_1 = tr(\boldsymbol{S})$ e $\nu_2 = tr(\boldsymbol{S^T S})$, onde o traço da matriz $\boldsymbol{S}$ é igual ao traço da matriz de projeção (ou em inglês, *hat matrix*) da RGP, sendo relacionada com os vetores $\hat{\boldsymbol{\mu}}$ e $\boldsymbol{y}$. As linhas $\boldsymbol{s_j}$ são denotadas por:

$$\boldsymbol{s_j} = \boldsymbol{X}_j(\boldsymbol{X^T W}(i)\boldsymbol{X})^{-1}\boldsymbol{X^T W}(i) \tag{4.12}$$

onde $\boldsymbol{X}_j$ representa a $j-$ésima linha da matriz do modelo $\boldsymbol{X}$ e $\boldsymbol{W}(i)$ a matriz de ponderação espacial.

O número efetivo de graus de liberdade é dado pela expressão $n - 2\nu_1 + \nu_2$ e o número efetivo de parâmetros estimados pelo modelo é dado pela expressão $2\nu_1 - \nu_2$. Os valores de $tr(\boldsymbol{S})$ e $tr(\boldsymbol{S^T S})$ são bem próximos, portanto, é possível utilizar $\nu_1$ como aproximação para o número efetivo de parâmetros no modelo (Fotheringham et al., 2002).

Na RGP, para cada estimativa de parâmetro global, têm-se $n$ estimativas de parâmetros locais, sendo que $n$ é o número de locais em que o modelo RGP é calibrado. A significância local para a estimativa do $k-$ésimo parâmetro no ponto $i$ pode ser avaliada por meio do $pseudo$ teste $t$, denotado por:

$$t_k(u_i, v_i) = \frac{\hat{\beta}_k(u_i, v_i)}{EP[\hat{\beta}_k(u_i, v_i)]} \tag{4.13}$$

cuja distribuição é aproximadamente Normal.

Para solucionar esse problema da utilização do $pseudo$ teste $t$, Da Silva e Fotheringham (2016) propuseram um ajuste neste teste, ajustando o nível de significância $\alpha$. Portanto, esse ajuste pode ser expresso da forma:

$$\alpha = \frac{p}{p_e}\xi_m = \frac{\xi_m}{\frac{p_e}{p}} \tag{4.14}$$

onde $p_e$ representa o número efetivo de parâmetros independentes estimados na RGP e é definido da forma: $p_e = 2tr(\boldsymbol{S}) - tr(\boldsymbol{S^T S})$, $p$ representa o número de parâmetros do modelo e $\xi_m$ representa o nível $\alpha$ desejado sem considerar a dependência espacial.

Enquanto isso, o AICc (critério AIC corrigido), sendo uma estatística para adequação relativa de comparação entre dois ou mais modelos, desenvolvido por Hurvich e Tsai (1989) é denotado da forma:

$$AIC_c = 2n\log(\hat{\sigma}) + n\log(2\pi) + \frac{n(n + tr(\boldsymbol{S})}{n - 2 - tr(\boldsymbol{S})} \tag{4.15}$$

onde $tr(\boldsymbol{S})$ é o número efetivo de parâmetros da RGP; $\hat{\sigma}$ representa a estimativa de máxima verossimilhança de $\sigma$ e $\boldsymbol{S}$ representa a matriz que relaciona $\hat{\boldsymbol{\mu}}$ e $\boldsymbol{y}$.

Na RGP, algumas medidas de ajuste e diagnóstico para a modelagem local são obtidas, como o coeficiente de determinação local $R_i^2$ que traz a informação do grau de ajuste dos modelos locais. Sendo assim, a sua expressão é dada da forma (Fotheringham et al., 2002):

$$R_i^2 = \frac{TSS_i^w - RSS_i^w}{TSS_i^w} \tag{4.16}$$

onde $TSS_i^w = \sum_j w_{ij}(y_j - \bar{y})^2$ e $RSS_i^w = \sum_j w_{ij}(y_j - \hat{y})^2, \quad j = 1, ..., n$, são a soma de quadrados total e residual geograficamente ponderadas, respectivamente.

### 4.2.1 Estimação do parâmetro de suavização ($bandwidth$)

Brunsdon et al. (1998) afirmam que devido à suavidade das funções e ao aspecto da $distance-decay$, os resultados da RGP são pouco influenciados pela escolha da função de ponderação, ainda que existem várias opções de função. Segundo Leung et al. (2000), a função mais utilizada é a (4.4). No entanto, Dempster et al. (2009) afirmam que a escolha do parâmetro de suavização é mais crítica, sendo que a escolha ótima para este valor na calibragem da RGP seja uma etapa importante.

Como pode-se ver na Figura 4.2, o parâmetro de suavização funciona como um fator de variabilidade da curva dos pesos.

A Validação Cruzada (do inglês, $Cross-Validation-CV$), sugerida por Cleveland (1979) pode ser utilizada para determinar o parâmetro de suavização, para a regressão local da forma:

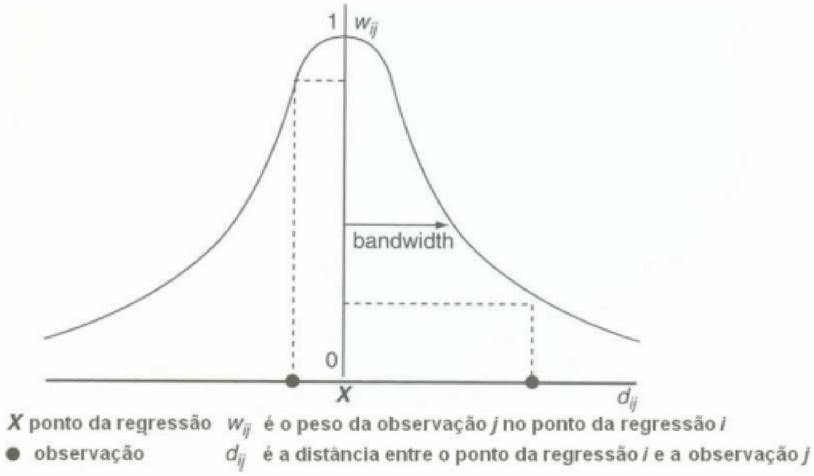



Figura 4.2: Parâmetro de suavização
Fonte: Fotheringham et al. (2002).

$$CV = \sum_{i=1}^{n} [y_i - \hat{y} \neq_i (b)]^2 \tag{4.17}$$

em que $\hat{y} \neq_i (b)$ passa a ser o valor ajustado para o ponto $y_i$, onde se omite a própria $i$-ésima observação do ajuste.

Esta abordagem tem a propriedade desejável de contabilizar o "efeito ao redor", já que quando $b$ se torna o menor possível, o modelo é calibrado apenas em amostras perto de $i$ e não do $i$ em si. Portanto o valor que minimiza a Equação (4.17) é o parâmetro de suavização ótimo do método de Validação Cruzada (Da Silva e Mendes, 2018). Esse parâmetro de suavização é usualmente estimado via algoritmo de Otimização por seção áurea ($Golden\ Section\ Search$).

## 4.3 Regressão binomial (logística) geograficamente ponderada

Para dados cuja variável resposta é binária, a metodologia da RGP pode ser feita por meio da Regressão Binomial Geograficamente Ponderada ou Regressão Logística Geograficamente Ponderada - RLGP, como é mais conhecida (do inglês, *Geographically Weighted Logistic Regression - GWLR*), introduzida por Atkinson et al. (2003). Portanto, a probabilidade de ocorrência do evento é dada por:

$$\log\left(\frac{\mu_j}{1-\mu_j}\right) = \sum_k \beta_k(u_i, v_i)x_{jk}, \quad j = 1, ..., n. \tag{4.18}$$

onde $(u_i, v_i)$ compõe as coordenadas do $i-$ésimo ponto no espaço; $\beta_k(u_i, v_i)$ é o parâmetro para a $k-$ésima variável explicativa, em função do local da $i-$ésima observação, $x_{jk}$ é o valor da $k-$ésima variável explicativa para a observação $i$ e $\mu_j$ é dado pela expressão:

$$\mu_j = \frac{\exp(\sum_k \beta_k(u_i, v_i)x_{jk})}{1 + \exp(\sum_k \beta_k(u_i, v_i)x_{jk})} \tag{4.19}$$

Sendo assim, o modelo de Regressão Logística Geograficamente Ponderado pode ser expresso da forma:

$$y_j \sim Binomial\left[\frac{\exp(\sum_k \beta_k(u_i, v_i)x_{jk})}{1 + \exp(\sum_k \beta_k(u_i, v_i)x_{jk})}\right] \tag{4.20}$$

A estimação dos parâmetros da RLGP é feita por meio da função log-verossimilhança, sendo denotada por:

$$L(\beta(u_i, v_i)|\boldsymbol{B}) = \sum_{j=1}^{n}\{y_j \log[\mu_j(\boldsymbol{\beta}(i))] + (1 - y_j)\log[1 - \mu_j(\boldsymbol{\beta}(i))]\}w(d_{ij}) \tag{4.21}$$

onde $\mu_j$ é expresso em (4.19) e depende de $\boldsymbol{\beta}(i)$ e $\boldsymbol{B}$ representa os dados $\{x_{jk}\}$, $\{y_j\}$ e $\{(u_i, v_i)\}$.

O Método Escore de Fisher passa a ser utilizado para a maximização de (4.21), entretanto, é feita uma modificação nesse método, onde se inclue a ponderação geográfica dada pela matriz de proximidade espacial $\boldsymbol{W}(i)$. Sendo assim, é feita uma multiplicação da matriz dos pesos do Método Escore de Fisher pela matriz de pesos da RGP (Fotheringham et al., 2002). Tal solução é dada por:

$$\boldsymbol{\beta}(u_i, v_i)^{(m+1)} = [\boldsymbol{X^T W}(u_i, v_i)\boldsymbol{A}(u_i, v_i)^{(m)}\boldsymbol{X}]^{-1}\boldsymbol{X^T W}(u_i, v_i)\boldsymbol{A}(u_i, v_i)^{(m)}\boldsymbol{z}(u_i, v_i)^{(m)} \tag{4.22}$$

onde $\boldsymbol{X}$ representa a matriz das covariáveis; $\boldsymbol{W}(u_i, v_i)$ representa a matriz diagonal de pesos da RGP e $\boldsymbol{A}(u_i, v_i)^{(m)}$ representa a matriz diagonal do MLG na interação $m$ para a localidade $i$. Com base na diagonal de $\boldsymbol{A}(u_i, v_i)^{(m)}$, os elementos $a_{ij}^{(m)}$, onde $j = 1, ..., n$, são dados por:

$$a_{ij}^{(m)} = \frac{1}{V(\mu_j)}\left(\frac{\partial \mu_j}{\partial \eta_j}\right)^2 = \mu_j(\boldsymbol{\beta}(u_i, v_i)^{(m)})(1 - \mu_j(\boldsymbol{\beta}(u_i, v_i)^{(m)})) \tag{4.23}$$

E $\boldsymbol{z}$ representa o vetor da variável dependente ajustada no algoritmo do Método Escore de Fisher, então para a RLGP, a expressão é dada por:

$$z_j(\boldsymbol{\beta}(i))^{(m)} = \boldsymbol{X}\boldsymbol{\beta}(i)^{(m)} + \frac{y_j - \mu_j(\boldsymbol{\beta}(\boldsymbol{i})^{(m)})}{\mu_j(\boldsymbol{\beta}(\boldsymbol{i})^{(m)})(1 - \mu_j(\boldsymbol{\beta}(\boldsymbol{i})^{(m)}))} \tag{4.24}$$

Portanto, com a convergência do algoritmo, tem-se que:

$$\hat{\boldsymbol{\beta}}(u_i, v_i) = \boldsymbol{C}(u_i, v_i)\boldsymbol{z}(u_i, v_i) \tag{4.25}$$

em que,

$$\boldsymbol{C}(u_i, v_i) = [\boldsymbol{X^T W}(u_i, v_i)\boldsymbol{A}(u_i, v_i)^{(m)}\boldsymbol{X}]^{-1}\boldsymbol{X^T W}(u_i, v_i)\boldsymbol{A}(u_i, v_i)^{(m)} \tag{4.26}$$

A matriz de covariância com as estimativas dos parâmetros da RLGP pode ser obtida da forma:

$$\widehat{Cov}\left(\hat{\boldsymbol{\beta}}(u_i, v_i)\right) = \boldsymbol{C}(u_i, v_i)\boldsymbol{A}(u_i, v_i)^{-1}\boldsymbol{C^T}(u_i, v_i) \tag{4.27}$$

Com isso, a estimativa do erro padrão do $k$-ésimo parâmetro para o local $i$ é dada por:

$$EP\left[\hat{\boldsymbol{\beta}}_{\boldsymbol{k}}(u_i, v_i)\right] = \sqrt{\widehat{Cov}(\hat{\boldsymbol{\beta}}(u_i, v_i))_k} \tag{4.28}$$

onde $\widehat{Cov}(\hat{\boldsymbol{\beta}}(u_i, v_i))_k$ representa o $k-$ésimo elemento da diagonal da matriz.

A estimativa utilizada para a matriz de covariância de $\boldsymbol{z}(u_i, v_i)$ não é ponderada pelos pesos $\boldsymbol{W}(u_i, v_i)$, sendo assim, os erros padrão estimados se caracterizam diferemente dos que se resultam das regressões locais ponderadas.

Na escolha da matriz de ponderação espacial $\boldsymbol{W}(u_i, v_i)$ é necessário determinar o parâmetro de suavização para a minimização da estatística AICc. A correção no AIC foi proposta por Hurvich e Tsai (1989) para selecionar os modelos de regressão com tamanhos de amostra pequenos, sendo assim a expressão é dada por:

$$AIC_c = -2\left(\sum_{j=1}^{n} y_j \log(\mu_j) + (1-y_j)\log(1-\mu_j)\right) + 2tr(\boldsymbol{S}) + \frac{2tr(\boldsymbol{S})(tr(\boldsymbol{S})+1)}{n-1-tr(\boldsymbol{S})} \tag{4.29}$$

onde $tr(\boldsymbol{S})$ é o número efetivo de parâmetros da RLGP e a matriz $\boldsymbol{S}$ é a que relaciona as matrizes $\hat{\boldsymbol{\eta}}$ e $\boldsymbol{z}$, sendo $\hat{\boldsymbol{\eta}} = \boldsymbol{S}\boldsymbol{z}$. Portanto, as linhas $\boldsymbol{s_j}$ da matriz $\boldsymbol{S}$ são dadas da forma:

$$\boldsymbol{s_j} = \boldsymbol{X}_j[\boldsymbol{X^T}\boldsymbol{W}(i)\boldsymbol{A}(i)\boldsymbol{X}]^{-1}\boldsymbol{X^T}\boldsymbol{W}(i)\boldsymbol{A}(i) \tag{4.30}$$

sendo que $\boldsymbol{X_j}$ é a $j-$ésima linha da matriz do modelo $\boldsymbol{X}$.

## 4.4 Regressão Poisson geograficamente ponderada

Nakaya et al. (2005) desenvolveram a Regressão Poisson Geograficamente Ponderada - RPGP (do inglês, *Geographically Weighted Poisson Regression - GWPR*), com o objetivo de avaliar a relação entre a taxa de mortalidade e alguns fatores sócio-econômicos na área metropolitana de Tóquio.

Considerando o modelo de Poisson, com a taxa $\mu_j/t_j$, onde $t_j$ reflete uma variável *offset*, que pode ser o tempo de exposição ou a área de interesse do evento, e ainda considerando a regressão de Poisson que foi apresentada anteriormente, tem-se:

$$\log(\mu_j) = \sum_k \beta_k x_{jk} + \log(t_j) \tag{4.31}$$

Portanto,

$$\log\left(\frac{\mu_j}{t_j}\right) = \sum_k \beta_k x_{jk} \tag{4.32}$$

e

$$\mu_j = t_j \exp\left(\sum_k \beta_k x_{jk}\right) \tag{4.33}$$

Por meio da variação espacial aos parâmetros $\beta_k$, o modelo de Regressão de Poisson Geograficamente Ponderado pode ser expresso da forma:

$$y_j \sim Poisson\left[t_j \exp\left(\sum_k \beta_k(u_i, v_i) x_{jk}\right)\right] \tag{4.34}$$

Com base na função local de log-verossimilhança, pode ser feita uma maximização para calibragem do modelo em (4.34). A função de log-veromilhança global para a Poisson é dada da forma:

$$L(\boldsymbol{\beta}(u,v)|\boldsymbol{B}) = \sum_{j=1}^{n} -\mu_j + y_j \log(\mu_j) \tag{4.35}$$

onde $\mu_j$ é expresso em (4.33) e depende de $\boldsymbol{\beta}(u,v)$; $\boldsymbol{B}$ representa os dados $\{x_{jk}\}$ e $\{y_j\}$. Ao considerar a hipótese da superfície de $\beta_k$, sendo aproximadamente plana próximo de um ponto $i$, a log-verossimilhança local pode ser expressa da forma:

$$L(\boldsymbol{\beta}(u_i,v_i)|\boldsymbol{B}) = \sum_{i=1}^{n} \{-\mu_j(\boldsymbol{\beta}(i)) + y_j \log[\mu_j(\boldsymbol{\beta}(i))]\} w(d_{ij}) \tag{4.36}$$

onde $\mu_j(\boldsymbol{\beta}(i))$ representa o valor esperado de $y$ no ponto $j$, baseando-se nos parâmetros do ponto $i$, sendo assim:

$$\mu_j(\boldsymbol{\beta}(i)) = t_j \exp\left(\sum_k \beta_k(u_i, v_i) x_{jk}\right) \tag{4.37}$$

A estimativa do valor para $y$ no ponto $j$ é calculada levando em conta os parâmetros também estimados para o ponto $j$, sendo $\mu_j(\hat{\boldsymbol{\beta}}(i))$. A média que é calculada em (4.37) é calculada como passo intermediário para estimar os parâmetros $\boldsymbol{\beta}(i)$.

Assim como na RLGP, o Método Escore de Fisher também passa a ser utilizado para a maximização de (4.36), incluindo a ponderação geográfica dada pela matriz de proximidade espacial $\boldsymbol{W}(i)$. Sendo assim, é feita uma multiplicação da matriz dos pesos do Método Escore de Fisher pela matriz de pesos da RGP (Fotheringham et al., 2002). Tal solução é dada por:

$$\boldsymbol{\beta}(u_i, v_i)^{(m+1)} = [\boldsymbol{X^T W}(u_i, v_i)\boldsymbol{A}(u_i, v_i)^{(m)}\boldsymbol{X}]^{-1}\boldsymbol{X^T W}(u_i, v_i)\boldsymbol{A}(u_i, v_i)^{(m)}\boldsymbol{z}(u_i, v_i)^{(m)} \tag{4.38}$$

onde $\boldsymbol{X}$ representa a matriz do modelo; $\boldsymbol{W}(u_i, v_i)$ representa a matriz diagonal de pesos da RGP e $\boldsymbol{A}(u_i, v_i)^{(m)}$ representa a matriz diagonal do MLG na interação $m$ para a localidade $i$. Com base na diagonal de $\boldsymbol{A}(u_i, v_i)^{(m)}$, os elementos $a_{ij}^{(m)}$, onde $j = 1, ..., n$, são dados por:

$$a_{ij}^{(m)} = \frac{1}{V(\mu_j)}\left(\frac{\partial \mu_j}{\partial \eta_j}\right)^2 = \mu_j(\boldsymbol{\beta}(u_i, v_i)^{(m)}) \tag{4.39}$$

E $\boldsymbol{z}$ representa o vetor da variável dependente ajustada no algoritmo do Método Escore de Fisher, então para a RPGP, a expressão é dada por:

$$z_j(\boldsymbol{\beta}(i))^{(m)} = \boldsymbol{X}\boldsymbol{\beta}(i)^{(m)} + \frac{y_j - \mu_j(\boldsymbol{\beta}(i)^{(m)})}{\mu_j(\boldsymbol{\beta}(i)^{(m)})} \tag{4.40}$$

Portanto, com a convergência do algoritmo, tem-se que:

$$\hat{\boldsymbol{\beta}}(u_i, v_i) = \boldsymbol{C}(u_i, v_i)\boldsymbol{z}(u_i, v_i) \tag{4.41}$$

em que,

$$\boldsymbol{C}(u_i, v_i) = [\boldsymbol{X^T W}(u_i, v_i)\boldsymbol{A}(u_i, v_i)\boldsymbol{X}]^{-1}\boldsymbol{X^T W}(u_i, v_i)\boldsymbol{A}(u_i, v_i) \tag{4.42}$$

A matriz de covariância com as estimativas dos parâmetros da RPGP pode ser obtida da mesma forma em (4.27), assim como a estimativa do erro padrão do $k$-ésimo parâmetro para o local $i$ é dada por (4.28)

O AIC que é utilizado por Nakaya et al. (2005) é dado como $AIC = D + 2k$, onde $D$, definido em (2.23), é a função *deviance*. Na RPGP, os dois conceitos de AIC são equivalentes, sendo assim, com base em (4.29) e os resultados da RPGP, tem-se:

$$AIC_c = -2\left(\sum_{j=1}^{n} -\mu_j + y_i \log \mu_j\right) + 2tr(\boldsymbol{S}) + \frac{2tr(\boldsymbol{S})(tr(\boldsymbol{S}) + 1)}{n - 1 - tr(\boldsymbol{S})} \tag{4.43}$$

onde $tr(\boldsymbol{S})$ é o número efetivo de parâmetros da RPGP e a matriz $\boldsymbol{S}$ é a que relaciona as matrizes $\hat{\boldsymbol{\eta}}$ e $\boldsymbol{z}$, sendo $\hat{\boldsymbol{\eta}} = \boldsymbol{Sz}$. Portanto, as linhas $\boldsymbol{s_j}$ da matriz $\boldsymbol{S}$ são dadas da forma:

$$\boldsymbol{s_j} = \boldsymbol{X_j}[\boldsymbol{X^T W}(i)\boldsymbol{A}(i)\boldsymbol{X}]^{-1}\boldsymbol{X^T W}(i)\boldsymbol{A}(i) \tag{4.44}$$

sendo que $\boldsymbol{X_j}$ é a $j-$ésima linha da matriz do modelo $\boldsymbol{X}$.

## 4.5 Regressão binomial negativa geograficamente ponderada

A Regressão Binomial Negativa Geograficamente Ponderada - RBNGP, proposta por Da Silva e Rodrigues (2014), é indicada para modelar dados espaciais de contagem não estacionários e com superdispersão. A RBNGP passa a ser mais robusta que a RPGP devido a flexibilização da hipótese de igualdade entre a média e variância da distribuição de Poisson, devido à presença do parâmetro $\alpha$ da distribuição binomial negativa.

A RBNGP permite a variação espacial aos parâmetros $\beta$ e $\alpha$ do modelo global de regressão binomial negativo, sendo uma extensão desse modelo global. A função de ligação logarítmica é utilizada no modelo de Regressão binomial negativo global, portanto, ao parametrizar este modelo, considerando a taxa $\mu_j/t_j$, tem-se (Da Silva e Rodrigues, 2014):

$$y_j \sim BN\left[t_j \exp\left(\sum_k \beta_k x_{jk}\right), \alpha\right] \tag{4.45}$$

onde $t_j$ representa uma variável *offset*.

A RBNGP produz superfícies não paramétricas para as estimativas dos parâmetros, sendo que o modelo espacial local passa a ser descrito da forma:

$$y_j \sim BN\left[t_j \exp\left(\sum_k \beta_k(u_i, v_i) x_{jk}\right), \alpha(u_i, v_i)\right] \tag{4.46}$$

A combinação dos métodos de Newton-Raphson e do Método Escore de Fisher é feita para a estimação dos parâmetros em (4.45). A log-verossimilhança da RBNGP, com função de $\boldsymbol{\beta}$ é dada pela expressão:

$$\begin{aligned} L(\boldsymbol{\beta}(u,v)|\boldsymbol{B}, \boldsymbol{\alpha}(u,v)) \quad = \quad & \sum_{j=1}^{n} \{y_j \log(\alpha_j \mu_j) - (y_j + 1/\alpha_j) \log(1 + \alpha_j \mu_j) + \\ & \log[\Gamma(y_j + 1/\alpha_j)] - \log[\Gamma(1/\alpha_j)] - \log[\Gamma(y_j + 1)]\} \end{aligned} \tag{4.47}$$

onde $\boldsymbol{B} = \{x_{jk}\}$ e $\{y_j\}$, para $j = 1, ..., n$,

em que:

$$\mu_j = t_j \exp\left(\sum_k \beta_k(u_j, v_j) x_{jk}\right) \tag{4.48}$$

$$\alpha_j = \alpha(u_j, v_j) \tag{4.49}$$

A log-verossimilhança local da RBNGP, considerando a hipótese de superfície de $\beta_k$ plana na vizinhança de um ponto $i$ qualquer, passa a ser escrita na forma (Da Silva e Rodrigues, 2014):

$$\begin{aligned} L(\boldsymbol{\beta}(u_i, v_i)|\boldsymbol{B}, \alpha(i)) \quad = \quad & \sum_{j=1}^{n}\{y_j \log[\alpha(i)\mu_j(\boldsymbol{\beta}(i))] - [y_j + 1/\alpha(i)] \log[1 + \alpha(i)\mu_j(\boldsymbol{\beta}(i))] + \\ & \log[\Gamma(y_j + 1/\alpha(i))] - \log[\Gamma(1/\alpha(i))] - \log[\Gamma(y_j + 1)]\}w(d_{ij}) \end{aligned} \tag{4.50}$$

para $i = 1, ..., N$

em que:

$$\mu_j(\boldsymbol{\beta}(i)) = t_j \exp\left(\sum_k \beta_k(u_j, v_j)x_{jk}\right) \tag{4.51}$$

$$\alpha(i) = \alpha(u_i, v_i) \tag{4.52}$$

Considerando os resultados de Nakaya et al. (2005) para a RPGP, tem-se a solução da maximização da log-verossimilhança local que fornece as estimativas $\hat{\boldsymbol{\beta}}(i)$ dos parâmetros da RBNGP é dada por:

$$\boldsymbol{\beta}(u_i, v_i)^{(m+1)} = [\boldsymbol{X^T W}(u_i, v_i)\boldsymbol{A}(u_i, v_i)^{(m)}\boldsymbol{X}]^{-1}\boldsymbol{X^T W}(u_i, v_i)\boldsymbol{A}(u_i, v_i)^{(m)}\boldsymbol{z}(u_i, v_i)^{(m)} \tag{4.53}$$

Os elementos $a_{ij}^{(m)}(j = 1, ..., n)$ da diagonal da matriz $\boldsymbol{A}(u_i, v_i)^{(m)}$, que é uma matriz diagonal de pesos do MLG na interação $m$ para o local $i$, para a RBNGP são dados como:

$$a_{ij}^{(m)} = \frac{\mu_j(\boldsymbol{\beta}(i)^{(m)})}{1 + \alpha(i)\mu_j(\boldsymbol{\beta}(i)^{(m)})} + \frac{[y_j - \mu_j(\boldsymbol{\beta}(i)^{(m)})][\alpha(i)\mu_j(\boldsymbol{\beta}(i)^{(m)})]}{1 + 2\alpha(i)\mu_j(\boldsymbol{\beta}(i)^{(m)}) + \alpha^2(i)\mu_j^2(\boldsymbol{\beta}(i)^{(m)})} \tag{4.54}$$

Finalmente, os elementos da variável dependende $\boldsymbol{z}(u_i, v_i)^{(m)}$ são:

$$z_j(\boldsymbol{\beta}(i)^{(m)}) = \boldsymbol{X\beta}(i)^{(m)} + \frac{[y_j - \mu_j(\boldsymbol{\beta}(i)^{(m)})]}{a_{ij}^{(m)}(1 + \alpha(i) \times \mu_j(\boldsymbol{\beta}(i)^{(m)}))} \tag{4.55}$$

Seguindo a ideia da RPGP, as estimativas dos parâmetros da matriz de covariância podem ser obtidas:

$$\widehat{Cov} = (\hat{\boldsymbol{\beta}}(u_i, v_i)) = \boldsymbol{C}(u_i, v_i)\boldsymbol{A}(u_i, v_i)^{-1}\boldsymbol{C}^{\boldsymbol{T}}(u_i, v_i) \tag{4.56}$$

onde

$$\boldsymbol{C}(u_i, v_i) = (\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{W}(u_i, v_i)\boldsymbol{A}(u_i, v_i)\boldsymbol{X})^{-1}\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{W}(u_i, v_i)\boldsymbol{A}(u_i, v_i) \tag{4.57}$$

e $\boldsymbol{A}(u_i, v_i)$ e $\boldsymbol{z}(u_i, v_i)$ definidos em (4.54) e (4.55), respectivamente.

Para estimação do parâmetro $\alpha(i)$ é utilizado o método de Newton-Raphson, baseado na log-verossimilhança local e supondo que $\boldsymbol{\beta}(u_i, v_i)$ sejam conhecidos. Ao reescrever (4.50), considerando a estimação dos parâmetros $k(i)$, onde $k(i) = 1/\alpha(i)$, tem-se:

$$\begin{aligned} L((k(i))|\boldsymbol{B}, \boldsymbol{\beta}(i)) \;=\; & \sum_{j=1}^{n}\{y_j \log[\mu_j(\boldsymbol{\beta}(i))] - [y_j + k(i)]\log[k(i) + k(i)\log[k(i)] + \\ & \log[\Gamma(y_j + k(i))] - \log[\Gamma(k(i))] - \log[\Gamma(y_j + 1)]\}w(d_{ij}) \end{aligned} \tag{4.58}$$

Pelo método de Newton-Raphson, pode-se maximizar a log-verossimilhança local em (4.58), portanto tem-se a expressão:

$$k(i)^{(m+1)} = k(i)^{(m)} - [H(i)^{(m)}]^{-1}U(i)^{(m)} \tag{4.59}$$

onde $U(i)^{(m)}$ e $H(i)^{(m)}$ são derivadas de primeira e segunda ordem da log-verossimilhança local, com respeito a $k(i)^{(m)}$, sendo:

$$\begin{aligned} U(i)^{(m)} \;=\; & \frac{\partial L(k(i))}{\partial k(i)} = \left(\sum_{j=1}^{n}\psi(k(i)^{(m)} + y_j) - \psi(k(i)^{(m)}) + \log(k(i)^{(m)}) + 1 - \log[k(i)^{(m)} + \mu_j(\boldsymbol{\beta}(i))]\right. \\ & \left. - \frac{k(i)^{(m)} + y_j}{k(i)^{(m)} + \mu_j(\boldsymbol{\beta}(i))}\right)w(d_{ij}) \end{aligned} \tag{4.60}$$

$$
\begin{aligned}
H(i)^{(m)} &= \frac{\partial^2 L(k(i))}{\partial k^2(i)} = \left( \sum_{j=1}^{n} \psi'(k(i)^{(m)} + y_j) - \psi'(k(i)^{(m)}) + \frac{1}{(k(i)^{(m)})} - \frac{2}{k(i)^{(m)} + \mu_j(\boldsymbol{\beta}(i))} \right. \\
&\left. + \frac{k(i)^{(m)} + y_j}{[k(i)^{(m)} + \boldsymbol{\mu}_j(\boldsymbol{\beta}(i)))]^2} \right) w(d_{ij}) \qquad (4.61)
\end{aligned}
$$

onde $\psi(.)$ e $\psi'(.)$, são as funções digama e trigama, respectivamente, sendo expressas: $\psi(z) = \frac{\partial \log \Gamma(z)}{\partial z}$ e $\psi'(z) = \frac{\partial \psi(z)}{\partial z} = \frac{\partial^2 \log \Gamma(z)}{\partial z^2}$.

Portanto, a estimativa para a variância de $\alpha(i)$ é dada:

$$\hat{\alpha}(i) = \frac{1}{\hat{k}(i)} \qquad (4.62)$$

$$\widehat{Var}(\hat{\alpha}(i)) = -\frac{1}{H(i)\hat{k}^4(i)} \qquad (4.63)$$

A estimação dos parâmetros de suavização é necessária para o ajuste do modelo. Portanto,pode-se determinar este parâmetro ao minimizar o AICc. Com base na Equação (4.29), tem-se que:

$$AIC_c = -2L(\boldsymbol{\beta}, \boldsymbol{\alpha}) + 2r + \frac{2r(r+1)}{n-r-1} \qquad (4.64)$$

onde $r$ representa o número efetivo de parâmetros e $L(\boldsymbol{\beta}, \boldsymbol{\alpha})$ é a log-verossimilhança da RBNGP. Esse número efetivo de parâmetros pode ser escrito como $r = r_1 + r_2$, em que $r_1$ e $r_2$ são os números efetivos de parâmetros devido a $\boldsymbol{\beta}$ e $\boldsymbol{\alpha}$, respectivamente (Da Silva e Rodrigues, 2014).

# Capítulo 5

# Regressão Binomial Negativa Inflacionada de Zeros Geograficamente Ponderada

## 5.1 Introdução

Com base na proposta dos trabalhos de Atkinson et al. (2003), que introduziram a RLGP, Nakaya et al. (2005), que introduziram a RPGP e Da Silva e Rodrigues (2014), que introduziram a RBNGP, o intuito deste trabalho é propor a regressão binomial negativa inflacionada de zeros geograficamente ponderada (RBNIZGP), sendo que localmente a RBNIZGP pode ser binomial negativa inflacionada de zeros, Poisson inflacionada de zeros, binomial negativa ou Poisson.

## 5.2 Modelo RPIZGP

Como dito anteriormente, a ideia da regressão geograficamente ponderada para o modelo Poisson inflacionado de zeros, chamada de regressão Poisson inflacionada de zeros geograficamente ponderada - RPIZGP (ou do inglês *Geographically weighted zero inflated Poisson regression* - GWZIPR) foi trabalhada por Kalagirou (2016), Purhadi et al. (2015) e Purhadi et al. (2021). Em Purhadi et al. (2015), pode-se perceber que para a estimação dos parâmetros do modelo RPIZGP foi utilizada a derivada por meio da log-verossimilhança (Purhadi et al., 2015):

$$
\begin{aligned}
L(\boldsymbol{\gamma},\boldsymbol{\beta}) &= \left\{\prod_{y_i=0}\left(\frac{e^{\boldsymbol{G_i\gamma}}+\exp(e^{\boldsymbol{X_i\beta}})}{1+e^{\boldsymbol{G_i\gamma}}}\right)+\prod_{y_i>0}\left(\frac{\frac{1}{1+e^{\boldsymbol{G_i\gamma}}}(\exp(-e^{\boldsymbol{X_i\beta}}+y_i\boldsymbol{X_i\beta}))}{y_i!}\right)\right\}w(d_{ij})\\
\log(L(\boldsymbol{\gamma},\boldsymbol{\beta})) &= \sum_{y=0}\log\left(e^{\boldsymbol{G_i\gamma}}+\exp(e^{\boldsymbol{X_i\beta}})\right)w(d_{ij})-\sum_{i=1}^{n}\log\left(1+e^{\boldsymbol{G_i\gamma}}\right)w(d_{ij})\\
&+ \sum_{y>0}\left(-e^{\boldsymbol{X_i\beta}}+y_i\boldsymbol{X_i\beta}\right)w(d_{ij})-\sum_{y>0}\log(y_i!)w(d_{ij}) \qquad (5.1)
\end{aligned}
$$

No entanto, poderia ser mantida estrutura do MLG, utilizando a matriz $\boldsymbol{A}$, como mostrada no Algoritmo 2 para o modelo logístico, e a matriz $\boldsymbol{A}$ no Algoritmo 3 para o modelo Poisson. No caso da Poisson inflacionada de zeros, que é uma combinação da distribuição logística e Poisson, a parte da RLGP seria estimada usando $z_i$ ao invés de $y_i$ e a RPGP seria ponderada por $(1-z_i)$, seguindo a mesma estrutura feita por Nakaya et al. (2005).

Nakaya et al. (2005) e Da Silva e Rodrigues (2014) utilizaram a estrutura do MLG, com a inclusão do termo de ponderação espacial $w(d_{ij})$ na RPGP e na RBNGP, respectivamente. Portanto, essa estrutura também será utilizada neste trabalho para a elaboração da RBNIZGP.

## 5.3 Modelo RBNIZGP

A regressão binomial negativa inflacionada de zeros geograficamente ponderada é uma extensão do modelo global binomial negativo inflacionado de zeros. Dessa forma, utiliza-se o modelo binomial negativo geograficamente ponderado no lugar do modelo binomial negativo, ponderado por $(\boldsymbol{1}-\boldsymbol{z})$ e o modelo logístico geograficamente ponderado no lugar do modelo binomial, trocando $\boldsymbol{y}$ por $\boldsymbol{z}$. Note que $z_i$ não precisa ser ponderado por $\boldsymbol{W}(i)$, pois $\hat{\boldsymbol{\beta}}(i)$ e $\hat{\boldsymbol{\gamma}}(i)$ já são ponderados por $\boldsymbol{W}(i)$.

Para obter as estimativas $\hat{\boldsymbol{\beta}}(i)$ e $\hat{\boldsymbol{\gamma}}(i)$ da regressão binomial negativa inflacionada de zeros geograficamente ponderada, basta usar o algoritmo EM conforme ilustrado no Algoritmo 7.

**Algoritmo 7:** Algoritmo EM - RBNIZGP

**Entrada:** $\boldsymbol{\beta_i}, \boldsymbol{\gamma_i}$

Estimar $\boldsymbol{\beta_i}$ e $k_i$ para $y > 0$, por meio RBNGP.

$\boldsymbol{\gamma_0} = (\sum_{j=1}^{n} I_{(y_j=0)} - \sum_{j=1}^{n}(k_i/(\mu_j + k_i))^{k_i})/n$

Estimar $\boldsymbol{\gamma_i} = \log(\boldsymbol{\gamma_0}/(1 - \boldsymbol{\gamma_0}))$

$DiffD_i = 1, \quad OldD_i = 0$

**enquanto** *($abs(DiffD_i) > 10^{-6}$)* **faça**

  **Passo E:** Estimar a esperança condicional $z_j$

  **se** $y_j = 0$ **então**

    $$z_j = \left(1 + e^{-\boldsymbol{G_j}\gamma_i}\left[\frac{\hat{k}_i}{e^{\boldsymbol{X_j}\hat{\boldsymbol{\beta}}_i} + \hat{k}_i}\right]^{\hat{k}_i}\right)^{-1}$$

  **senão**

    0

  **fim**

  **Passo M para $\boldsymbol{\beta_i}$ e $k_i$:** Estimação da RBNGP ponderado por $(1 - z_j)$, minimizando a $Deviance$ $(D_1)$ :

  { $\boldsymbol{\eta} = \boldsymbol{X\beta_i} + offset$

  $\boldsymbol{\mu} = \exp(\boldsymbol{\eta})$

  $\boldsymbol{M} = \frac{\Gamma(k_i+\boldsymbol{y})}{\Gamma(k_i)\Gamma(\boldsymbol{y}+1)}\left(\frac{\boldsymbol{\mu}}{\boldsymbol{\mu}+k_i}\right)^{\boldsymbol{y}}\left(\frac{k_i}{\boldsymbol{\mu}+k_i}\right)^{k_i}$

  $D_1 = \sum_{j=1}^{n}[(1 - z_j)(\log(M_j))]$

  }

  **Passo M para $\boldsymbol{\gamma_i}$:** Estimação RLGP não ponderada, utilizando $z_j$ como variável resposta, minimizando a $Deviance$ $(D_2)$ :

  { $\boldsymbol{\eta} = \boldsymbol{G\gamma_i}$

  $D_2 = \sum_{j=1}^{n}(z_j\eta_j - \sum_{j=1}^{n}(\log(1 + \exp(\eta_j)))$

  }

  **Maximização:** $OldD_i = D_i$

  $D_i = D_1 + D_2$

  $DiffD_i = OldD_i - D_i$

**fim**

Para a obtenção dos erros padrão dos parâmetros estimados da RBNIZGP, deve-se utilizar as equações das segundas derivadas (3.36), (3.37) (3.39), (3.38), (3.40) e (3.41), multiplicando tais equações pela matriz de ponderação espacial $\boldsymbol{W}(i)$.

## 5.4 Aspectos computacionais

A estrutura geral do algoritmo da RBNIZGP foi apresentada na seção anterior, entretanto, por ser um modelo geral para dados de contagem, existem alguns aspectos computacionais importantes tanto para a convergência do algoritmo quanto para a otimização do tempo de processamento. Tais aspectos serão vistos a seguir.

### 5.4.1 Verificação da quantidade de zeros

Um ponto a se considerar antes do ajuste da regressão RBNIZGP é a verificação da quantidade de zeros existentes da variável dependente, ou seja, verificar se os dados seguem de fato uma distribuição binomial negativa inflacionada de zeros.

Na regressão Poisson inflacionada de zeros, Lambert (1992) sugere comparar a quantidade de zeros da variável dependente com os valores esperados de zeros pelo modelo Poisson, por meio da probabilidade de excesso de 0 média observada $\hat{p}_0$. Fazendo uma adaptação para o modelo binomial negativo inflacionado de zeros, tem-se que:

$$\hat{p}_0 = \frac{\#(y_i = 0) - \sum_{i=1}^{n} \left(\frac{k}{e^{\boldsymbol{X_i\beta}}+k}\right)^k}{n} \tag{5.2}$$

onde $\left(\frac{k}{e^{\boldsymbol{X_i\beta}}+k}\right)^k$ é o número esperado de zeros em um modelo binomial negativo. Note que na distribuição de Poisson $P(X = x) = \frac{e^{-\lambda}\lambda^x}{x!}$. Então $P(X = 0) = \frac{e^{-\lambda}\lambda^0}{0!} = e^{-\lambda}$, sendo $\boldsymbol{\lambda} = e^{\boldsymbol{X\beta}}$ no caso da regressão Poisson. Para a distribuição binomial negativa, $P(Y = y) = \binom{y+k-1}{k-1}p^k(1-p)^y$ (2.37) e $P(Y = 0) = \binom{0+k-1}{k-1}p^k(1-p)^0 = p^k$, onde $p = \frac{k}{\mu+k}$ e $\boldsymbol{\mu} = e^{\boldsymbol{X\beta}}$ no caso de uma regressão binomial negativa. Dessa forma, como foi visto em (2.47), quando $k \to \infty$, esse número esperado de zeros converge para $\exp(-e^{\boldsymbol{X\beta}})$ como sugerido por Lambert (1992) para a regressão Poisson inflacionada de zeros.

Portanto, se a quantidade de zeros $\#(y_i = 0)$ for menor do que a quantidade de zeros esperada pelo modelo binomial negativo, ou seja, $\hat{p}_0 < 0$, ou quando não existirem zeros na variável dependente, então não haveria razão para o ajuste da RBNIZGP, fazendo com que $\gamma = 0$ e tendo o modelo RBNGP utilizado no ajuste.

### 5.4.2 Tamanho do parâmetro de suavização

Outro ponto importante é a verificação do tamanho do parâmetro de suavização. Caso o tamanho do parâmetro de suavização seja pequeno é necessário fazer uma checagem da matriz de ponderação $\boldsymbol{W}(i)$.

Na regressão global, não é possível estimar o modelo caso a quantidade de observações seja inferior à quantidade de variáveis (Neter et al., 1983). Portanto, para que seja possível realizar o ajuste do modelo RBNIZGP, o número de observações medido pela soma de $\boldsymbol{W}(i)$ não pode ser inferior à soma das dimensões de $\boldsymbol{\beta}(p \times 1)$ e $\boldsymbol{\gamma}(l \times 1)$, ou seja $\sum_{j=1}^{n} w_{ij} > (p + l)$.

### 5.4.3 Estimação do parâmetro de superdispersão $\alpha(i)$

No caso do parâmetro $k(i)$ (ou do parâmetro de superdispersão $\alpha(i) = \frac{1}{k(i)}$), não é necessário realizar uma estimação na $i$-ésima localidade partindo de uma estimativa geral. Para este caso, visando uma melhor otimização computacional e visto que o algoritmo de Newton-Raphson converge mais rápido quando a estimativa inicial está próxima do ótimo, pode-se ajustar uma regressão binomial negativa global e utilizar a estimativa de $k$ como valor inicial para as estimativas locais, ou seja, na $i$-ésima localidade.

### 5.4.4 Estimação da variância

Para obter as estimativas das variâncias dos parâmetros $\boldsymbol{\beta}(i), \boldsymbol{\gamma}(i)$ e $k(i)$ da RBNIZGP é necessário fazer essa estimação de forma conjunta, como foi visto no Capítulo 3. Conforme (4.56), a estimativa da matriz de covariância fica da forma

$$\begin{aligned}\boldsymbol{C}(u_i,v_i)\boldsymbol{A}^{-1}(u_i,v_i)\boldsymbol{C}^{\boldsymbol{T}}(u_i,v_i) &= (\boldsymbol{X}^T\boldsymbol{W}(u_i,v_i)\boldsymbol{A}(u_i,v_i)\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{W}(u_i,v_i)\boldsymbol{A}(u_i,v_i)\boldsymbol{A}^{-1}(u_i,v_i) \times \\ &\quad \boldsymbol{A}^T(u_i,v_i)\boldsymbol{W}^T(u_i,v_i)\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{W}(u_i,v_i)\boldsymbol{A}(u_i,v_i)\boldsymbol{X})^{-1}\end{aligned}$$

$$\begin{aligned} \boldsymbol{C}(u_i,v_i)\boldsymbol{A}^{-1}(u_i,v_i)\boldsymbol{C^T}(u_i,v_i) &= (\boldsymbol{X}^T\boldsymbol{W}(u_i,v_i)\boldsymbol{A}(u_i,v_i)\boldsymbol{X})^{-1}(\boldsymbol{X}^T\boldsymbol{W}(u_i,v_i)\boldsymbol{A}(u_i,v_i)\boldsymbol{W}(u_i,v_i)\boldsymbol{X}) \times \\ & \quad (\boldsymbol{X}^T\boldsymbol{W}(u_i,v_i)\boldsymbol{A}(u_i,v_i)\boldsymbol{X})^{-1} \end{aligned} \tag{5.3}$$

Pode-se observar em (5.3) que a matriz de ponderação espacial $\boldsymbol{W}(u_i,v_i)$ aparece duas vezes no termo não invertido, não permitindo que ela se torne (2.22). Caso $w_{ij} = 1, \forall i,j$, ou seja, tornando o modelo local em um modelo global, então (5.3) se torna equivalente à (2.22)

$$\boldsymbol{C}(u_i,v_i)\boldsymbol{A}^{-1}(u_i,v_i)\boldsymbol{C^T}(u_i,v_i) = (\boldsymbol{X}^T\boldsymbol{A}\boldsymbol{X})^{-1}(\boldsymbol{X}^T\boldsymbol{A}\boldsymbol{X})(\boldsymbol{X}^T\boldsymbol{A}\boldsymbol{X})^{-1} = (\boldsymbol{X}^T\boldsymbol{A}\boldsymbol{X})^{-1} \tag{5.4}$$

Note também que $\boldsymbol{A}$ em (2.49) é exatamente o termo $\frac{ke^{\boldsymbol{X\beta}}}{k+e^{\boldsymbol{X\beta}}}\left(\frac{\boldsymbol{y}-e^{\boldsymbol{X\beta}}}{k+e^{\boldsymbol{X\beta}}}+1\right)$ em (3.48). Dessa forma, para que a variância do parâmetro $\boldsymbol{\beta}(i)$ da RBNIZGP seja equivalente à RBNGP, a partir das derivadas segundas e quando $\boldsymbol{\gamma} = 0$, basta fazer

$$I_{11} = -(\boldsymbol{X^T d_{bb} X})(\boldsymbol{X^T d_{bb} A^{-1} d_{bb}^T X})^{-1}(\boldsymbol{X^T d_{bb} X}) \tag{5.5}$$

onde

$$\boldsymbol{d_{bb}} = \boldsymbol{W}(u_i,v_i)\boldsymbol{A}(u_i,v_i) \tag{5.6}$$

visto que a matriz de covariância (3.35) é a inversa da matriz de informação de Fisher, tornando esse termo igual a

$$Var(\hat{\boldsymbol{\beta}}(i)) = (\boldsymbol{X^T d_{bb} X})^{-1}(\boldsymbol{X^T d_{bb} A^{-1} d_{bb}^T X})(\boldsymbol{X^T d_{bb} X})^{-1} \tag{5.7}$$

que é equivalente à (5.3). Note que essa transformação é necessária pois como $\boldsymbol{W}(u_i,v_i)$ está ao quadrado no termo não invertido em (5.3), isso faz com que a variância de $\boldsymbol{\beta}(i)$ seja menor quanto menor for o parâmetro de suavização, quando comparada com o caso de haver apenas 1 $\boldsymbol{W}(u_i,v_i)$.

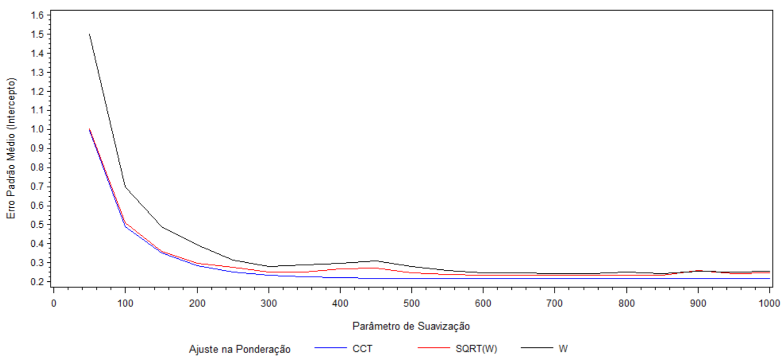
No caso do ajuste de uma RBNIZGP, ou seja, com $\boldsymbol{\gamma} > 0$, uma aproximação razoável para a variância apresentada em (5.3) seria considerar $\sqrt{\boldsymbol{W}(u_i, v_i)}$ ao invés de $\boldsymbol{W}(u_i, v_i)$,

$$I_{11} = -(\boldsymbol{X^T d_{sbb} X}) \tag{5.8}$$

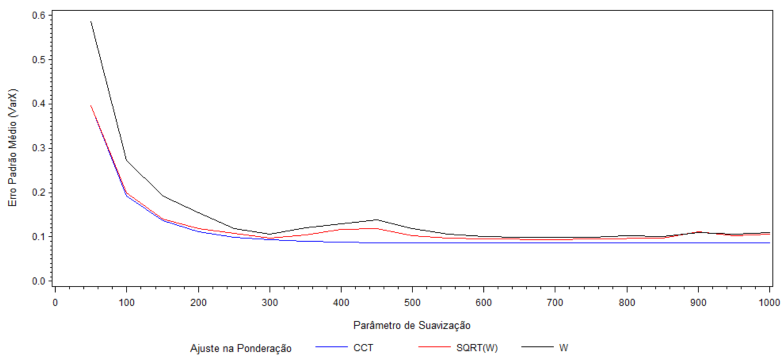
onde

$$\boldsymbol{d_{sbb}} = \sqrt{\boldsymbol{W}(u_i, v_i)}\boldsymbol{A}(u_i, v_i) \tag{5.9}$$

Para ilustrar essa diferença, a Figura 5.1 mostra o efeito do tipo da ponderação nas estimativas dos erros padrão de $\boldsymbol{\beta}(i)$ (intercepto e covariável VarX), em uma base simulada. Note que ao utilizar $\sqrt{\boldsymbol{W}(u_i, v_i)}$ ao invés de $\boldsymbol{W}(u_i, v_i)$ na derivada segunda $\boldsymbol{d_{sbb}}$, o erro padrão fica muito mais próximo do da estrutura dada em (5.3) (curva CCT na Figura 5.1).




(a)




(b)

Figura 5.1: Efeito do ajuste na matriz de ponderação espacial (a) Intercepto (b) VarX

### 5.4.5 Simplificação da *Deviance* $(D_1)$ a ser minimizada

Um problema que pode ocorrer no cálculo da *Deviance* $(D_1)$, expressa no Algoritmo 7, é que se o valor de $k$ for grande, $\Gamma(k)$ será um valor extremamente grande que pode não ser guardado na memória, gerando consequentemente um valor *missing*. No entanto, pode-se observar que $\frac{\Gamma(k+y)}{\Gamma(k)\Gamma(y+1)}$ é uma constante, e não influencia na estimação do parâmetro do modelo $\boldsymbol{\beta}$. Ademais, quando $y = 0$, $\frac{\Gamma(k+y)}{\Gamma(k)\Gamma(y+1)} = 1$ e quando $k \to \infty$, $\frac{\Gamma(k+y)}{\Gamma(k)\Gamma(y+1)} \approx 1$. Portanto, para evitar problemas de convergência, o valor de $\frac{\Gamma(k+y)}{\Gamma(k)\Gamma(y+1)}$ pode ser omitido da *Deviance* $(D_1)$.

### 5.4.6 Solução para o número de parâmetros efetivos referente ao parâmetro de superdispersão

Conforme discutido anteriomente por Da Silva e Rodrigues (2014), o número de parâmetros efetivos da RBNGP foi definido como $r = r_1 + r_2$, onde $r_1 = p_e$, sendo que $p_e$ representa o número efetivo de parâmetros independentes estimados. No entanto, os autores não conseguiram estimar $r_2$, propondo a RBNGPg, ou seja, com o parâmetro de superdispersão $\alpha$ global e fazendo com que $r_2 = 1$.

Uma solução para este problema, seria a utilização da taxa $\frac{p_e}{p}$, proposta por Da Silva e Fotheringham (2016) ao realizar um ajuste no nível de significância dos testes múltiplos, como descrito anteriormente em (4.14). Portanto, pode-se definir $r_2 = \frac{p_e}{p} \geq 1$ e consequentemente $r = p_e + \frac{p_e}{p}$. Quando o tamanho do parâmetro de suavização for considerado grande, então $p_e = p$ e $r_2 = 1$, gerando $r = p + 1$ como na regressão binomial negativa global. Note ainda que a correção no nível de significância $\alpha = \frac{p}{p_e}\xi_m$ (4.14) proposta por Da Silva e Fotheringham (2016) se mantém ao fazer essa correção em $r_2$, pois

$$\frac{p+1}{r} = \frac{p+1}{p_e + \frac{p_e}{p}} = \frac{p+1}{\frac{pp_e+p_e}{p}} = \frac{p(p+1)}{p_e(p+1)} = \frac{p}{p_e} \tag{5.10}$$

### 5.4.7 Gerando modelos Poisson, binomial negativo, binomial e Poisson inflacionada de zeros

O algoritmo da RBNIZGP permite estimar os modelos Poisson, binomial negativo, Poisson inflacionado de zeros ou ainda binomial, apenas atribundo valores 0 para alguns parâmetros.

Como pode ser visto de forma resumida na Tabela 5.1, caso o parâmetro $\boldsymbol{\gamma}$ do modelo binomial negativo inflacionado de zeros seja igual a 0, então o modelo de regressão será o binomial negativo. Caso o parâmetro de superdispersão $\alpha$ do modelo binomial negativo inflacionado de zeros seja igual a 0, então o modelo de regressão será o Poisson inflacionado de zeros. Caso os parâmetros $\alpha$ e $\boldsymbol{\gamma}$ sejam iguais a 0, tal modelo será Poisson. E por fim, caso os parâmetros $\alpha$ e $\boldsymbol{\beta}$ sejam iguais a 0, o modelo será o binomial. Essa ideia já foi descrita com detalhes na Figura 1.8, e mostra que pode-se forçar um modelo específico para os dados.

Tabela 5.1: Mudanças no modelo binomial negativo inflacionado de zeros para se chegar a outros modelos

| **Parâmetros da binomial negativa inflacionada de zeros** | **Modelo gerado** |
|---|---|
| $\boldsymbol{\gamma} = 0$ | *binomial negativo* |
| $\alpha = 0$ | *Poisson inflacionada de zeros* |
| $\alpha = 0, \quad \boldsymbol{\gamma} = 0$ | *Poisson* |
| $\alpha = 0, \quad \boldsymbol{\beta} = 0$ | *binomial* |

No entanto, a maior vantagem ao se utilizar o modelo de regressão binomial negativo inflacionado de zeros geograficamente ponderado (RBNIZGP), é que essas mudanças nos parâmetros ocorrem de forma natural para cada localidade, como foi visto na Figura 1.4, sem interferência do usuário. Ou seja, o usuário sempre pode iniciar a modelagem dos dados com o modelo RBNIZGP, sem precisar verificar a distribuição dos dados antes da análise, pois o algoritmo acomodará naturalmente a estrutura dos dados (a não ser no caso do modelo binomial em que a variável dependente deve ser binária e não uma contagem). Ao se observar a Figura 1.4, pode-se observar os casos em que ocorrem mudanças na distribuição naturalmente: o gráfico número 1 passa a se caracterizar como uma distribuição Poisson; o gráfico número 7 se assemelha a uma distribuição Poisson inflacionada de zeros; e o gráfico número 10 parece se adequar melhor a uma distribuição binomial negativa.

Da Silva e Rodrigues (2014) criaram o algoritmo da RBNGP, mostrando que o modelo binomial negativo retorna para o modelo Poisson, caso o parâmetro de superdispersão $\alpha$ seja igual a 0. Neste algoritmo da RBNIZGP, está sendo desenvolvido um algoritmo geral que pode retornar para outras distribuições, como pode ser visto na Tabela 5.1.

### 5.4.8 Estimação da *Deviance* e $R^2$

A *Deviance* para o modelo binomial negativo inflacionado de zeros é dada por (Martin e Hall, 2016)

$$D = 2\sum_{i=1}^{n}[L(\hat{\boldsymbol{z}}, \boldsymbol{y}, \hat{k}; \boldsymbol{y}) - L(\hat{\boldsymbol{\pi}}, \hat{\boldsymbol{\mu}}, \hat{k}; \boldsymbol{y})] \tag{5.11}$$

onde $\hat{\boldsymbol{\pi}} = \frac{e^{\boldsymbol{G}\hat{\boldsymbol{\gamma}}}}{1+e^{\boldsymbol{G}\hat{\boldsymbol{\gamma}}}}$, $\hat{\boldsymbol{\mu}} = e^{\boldsymbol{X}\hat{\boldsymbol{\beta}}}$ e

$$\begin{aligned} L(\hat{\boldsymbol{\pi}}, \hat{\boldsymbol{\mu}}, \hat{k}; \boldsymbol{y}) \quad = \quad & -\sum_{i=1}^{n}\log(1+e^{\boldsymbol{G_i}\hat{\boldsymbol{\gamma}}}) + \sum_{y_i>0}\left[k\log\left(\frac{k}{k+\hat{\mu}_i}\right) + y_i\log\left(\frac{\hat{\mu}_i}{k+\hat{\mu}_i}\right)\right] + \\ & \sum_{y_i=0}\log\left[e^{\boldsymbol{G_i}\hat{\boldsymbol{\gamma}}} + \left(\frac{k}{k+\hat{\mu}_i}\right)^k\right] + \\ & \sum_{y_i>0}[-\log\Gamma(y_i+1) - \log\Gamma(k) + \log\Gamma(y_i+k)] \end{aligned} \tag{5.12}$$

$$\begin{aligned} L(\hat{\boldsymbol{z}}, \boldsymbol{y}, \hat{k}; \boldsymbol{y}) \quad = \quad & -\sum_{i=1}^{n}\log(1+z_i) + \sum_{y_i>0}\left[k\log\left(\frac{k}{k+y_i}\right) + y_i\log\left(\frac{y_i}{k+y_i}\right)\right] + \\ & \sum_{y_i=0}\log\left[z_i + \left(\frac{k}{k+y_i}\right)^k\right] + \\ & \sum_{y_i>0}[-\log\Gamma(y_i+1) - \log\Gamma(k) + \log\Gamma(y_i+k)] \end{aligned} \tag{5.13}$$

Como foi visto em (3.32), quando $\boldsymbol{\gamma} = 0$, então $e^{\boldsymbol{G}\hat{\boldsymbol{\gamma}}} = 1$, e para que a *Deviance* do modelo

RBNIZGP seja o mesmo dos modelos Poisson ou binomial negativo, a Equação (5.12) deve ser

$$\begin{aligned} L(\hat{\boldsymbol{\pi}}, \hat{\boldsymbol{\mu}}, \hat{k}; \boldsymbol{y}) &= -\sum_{i=1}^{n} \log(0 + e^{\boldsymbol{G_i}\hat{\boldsymbol{\gamma}}}) + \sum_{y_i>0} \left[ k \log\left(\frac{k}{k+\hat{\mu}_i}\right) + y_i \log\left(\frac{\hat{\mu}_i}{k+\hat{\mu}_i}\right)\right] + \\ & \sum_{y_i=0} \log\left[0 \times e^{\boldsymbol{G_i}\hat{\boldsymbol{\gamma}}} + \left(\frac{k}{k+\hat{\mu}_i}\right)^k\right] + \sum_{y_i>0} [-\log\Gamma(y_i+1) - \log\Gamma(k) + \log\Gamma(y_i+k)] \end{aligned} \tag{5.14}$$

Ou seja, é uma forma computacional de fazer $e^{\boldsymbol{G}\hat{\boldsymbol{\gamma}}} = 0$, ou em outra palavras, eliminar esse termo do cálculo da *Deviance* para a regressão Poisson ou binomial negativa. A Equação (5.13) não precisa ser alterada pois quando $\boldsymbol{\gamma} = 0$, então $\boldsymbol{z} = 0$. Ademais, quando não existirem zeros na variável dependente, então toda a parte $\sum_{y_i=0}$ deve ser excluída e a parte $\sum_{y_i>0}$ deve ser $\sum_{y_i\geq 0}$, ou seja

$$\begin{aligned} L(\hat{\boldsymbol{\pi}}, \hat{\boldsymbol{\mu}}, \hat{k}; \boldsymbol{y}) &= -\sum_{i=1}^{n} \log(0 + e^{\boldsymbol{G_i}\hat{\boldsymbol{\gamma}}}) + \sum_{i=1}^{n} \left[ k \log\left(\frac{k}{k+\hat{\mu}_i}\right) + y_i \log\left(\frac{\hat{\mu}_i}{k+\hat{\mu}_i}\right)\right] + \\ & \sum_{i=1}^{n} [-\log\Gamma(y_i+1) - \log\Gamma(k) + \log\Gamma(y_i+k)] \end{aligned} \tag{5.15}$$

$$\begin{aligned} L(\hat{\boldsymbol{z}}, \boldsymbol{y}, \hat{k}; \boldsymbol{y}) &= -\sum_{i=1}^{n} \log(1 + z_i) + \sum_{i=1}^{n} \left[ k \log\left(\frac{k}{k+y_i}\right) + y_i \log\left(\frac{y_i}{k+y_i}\right)\right] + \\ & \sum_{i=1}^{n} [-\log\Gamma(y_i+1) - \log\Gamma(k) + \log\Gamma(y_i+k)] \end{aligned} \tag{5.16}$$

Uma forma de calcular o coeficiente de determinação $(R^2)$, nos mesmos moldes do desenvolvido a partir da *Deviance* para as regressões Poisson e binomial negativa da forma (Cameron e Windmeijer, 1997)

$$R_D^2 = 1 - \frac{D(\hat{\boldsymbol{\mu}}; \boldsymbol{y})}{D(\bar{y}; \boldsymbol{y})} \tag{5.17}$$

é dador por Martin e Hall (2016) como

$$R^2_{ZINB} = 1 - \frac{L(\hat{\boldsymbol{z}}, \boldsymbol{y}, \hat{k}; \boldsymbol{y}) - L(\hat{\boldsymbol{\pi}}, \hat{\boldsymbol{\mu}}, \hat{k}; \boldsymbol{y})}{L(\hat{\boldsymbol{z}}, \boldsymbol{y}, \hat{k}; \boldsymbol{y}) - L(0, \bar{y}, \hat{k}; \boldsymbol{y})} \tag{5.18}$$

onde

$$\begin{aligned} L(0, \bar{y}, \hat{k}; \boldsymbol{y}) \quad = \quad & -\sum_{i=1}^{n} \log(1+0) + \sum_{y_i>0} \left[ k \log\left(\frac{k}{k+\bar{y}}\right) + y_i \log\left(\frac{\bar{y}}{k+\bar{y}}\right) \right] + \\ & \sum_{y_i=0} \log\left[0 + \left(\frac{k}{k+\bar{y}}\right)^{k}\right] + \sum_{y_i>0} [-\log\Gamma(y_i+1) - \log\Gamma(k) + \log\Gamma(y_i+k)] \end{aligned} \tag{5.19}$$

E para o $(R^2_{adj})$, nos mesmos moldes do desenvolvido a partir da *Deviance* para as regressões Poisson e binomial negativa da forma (Mittlböck e Waldhör, 2000)

$$R^2_{D,adj} = 1 - \frac{[D(\hat{\boldsymbol{\mu}}; \boldsymbol{y}) + p/2]}{D(\bar{y}; \boldsymbol{y})} \tag{5.20}$$

é dador por Martin e Hall (2016) como

$$R^2_{ZINB,adj} = 1 - \frac{L(\hat{\boldsymbol{z}}, \boldsymbol{y}, \hat{k}; \boldsymbol{y}) - L(\hat{\boldsymbol{\pi}}, \hat{\boldsymbol{\mu}}, \hat{k}; \boldsymbol{y}) + p + l + 1,5}{L(\hat{\boldsymbol{z}}, \boldsymbol{y}, \hat{k}; \boldsymbol{y}) - L(0, \bar{y}, \hat{k}; \boldsymbol{y})} \tag{5.21}$$

onde $p$ e $l$ são as dimensões de $\boldsymbol{\beta}$ e $\boldsymbol{\gamma}$, respectivamente.

# Capítulo 6

# Materiais e Métodos

## 6.1 Introdução

O objetivo deste Capítulo é apresentar os materias e métodos utilizados no trabalho. Primeiramente serão utilizados dados simulados de distribuições Poisson, binomial negativa, Poisson inflacionado de zeros e binomial negativa inflacionada de zeros, sem variação espacial a fim de verificar se o modelo RBNIZGP é capaz de se ajustar. A seguir, serão utilizados os mesmos dados de Weinstein et al. (2021) sobre os casos de COVID-19 na Coréia do Sul, para verificar a qualidade do ajuste em dados reais.

## 6.2 Materiais

O intuito desta seção é descrever os principais materiais a serem utilizados. Primeiramente as análises serão feitas baseando-se nos dados simulados para as distribuições comentadas anteriormente e na sequência serão analisados os dados reais.

### 6.2.1 Dados simulados

A Tabela 6.1 traz os detalhes sobre os dados simulados, sendo que os dados simulados para as distribuições inflacionadas de zeros foram gerados segundo Erdman et al. (2008). Note que as variáveis explicativas $\boldsymbol{x}, \boldsymbol{x_1}, \boldsymbol{x_2}, \boldsymbol{x_3}$ foram geradas de uma distribuição normal e que as variáveis resposta $\boldsymbol{y}$ foram geradas segundo os parâmetros descritos seguindo uma distribuição específica para cada caso (Poisson, binomial negativa, Poisson inflacionado de zeros e binomial negativa inflacionada de zeros). O dados não possuem dependência espacial, sendo assim, nesta etapa tem-se por objetivo avaliar a potencialidade do algoritmo em se ajustar a dados não espaciais.

Tabela 6.1: Parâmetros dos dados simulados

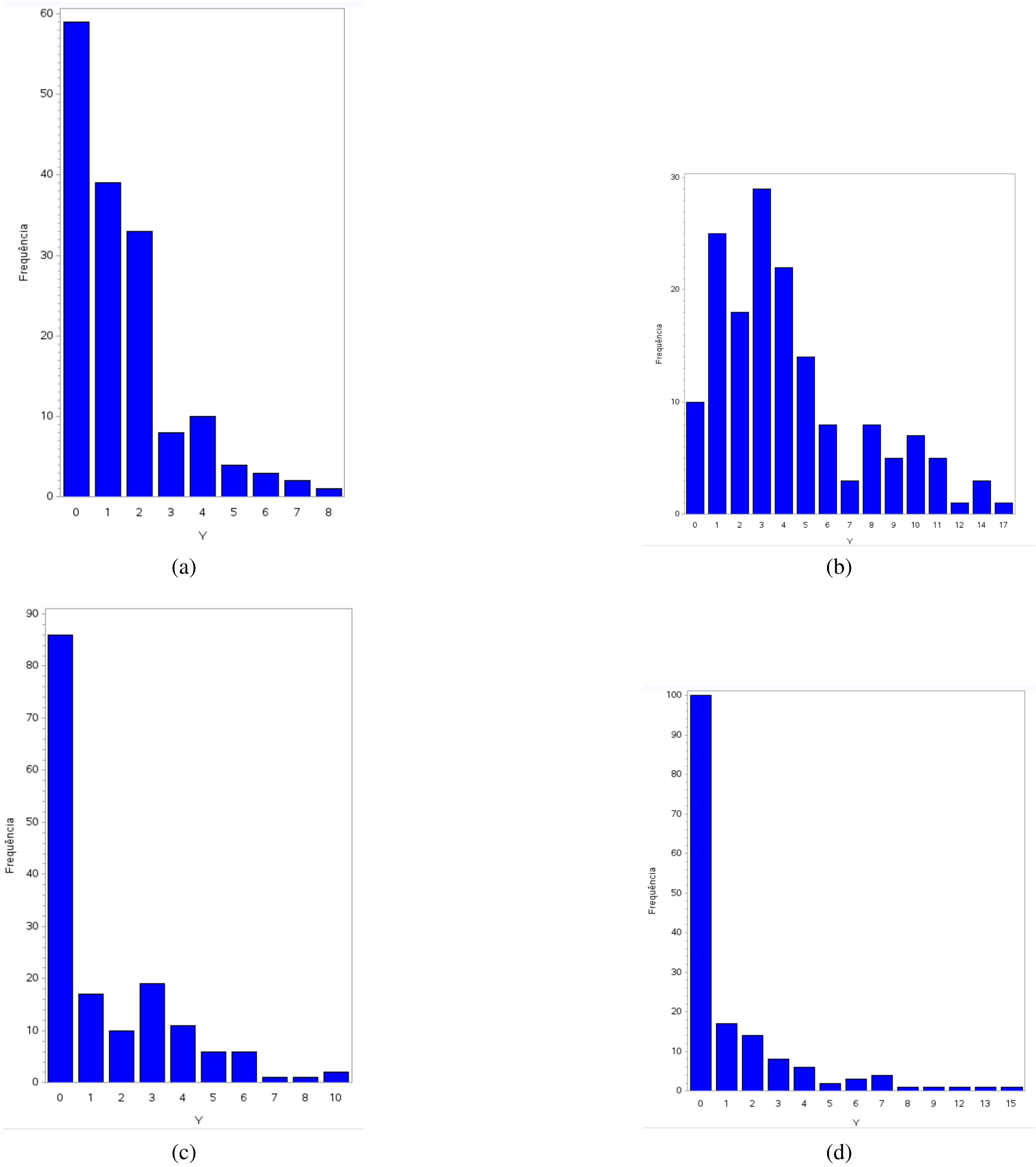
| **Poisson** | **Binomial negativa** | **Poisson inflacionado de zeros** | **Binomial negativa inflacionada de zeros** |
|---|---|---|---|
| $\boldsymbol{x} \sim Normal(3,1)$ | $\boldsymbol{x} \sim Normal(3,1)$ | $\boldsymbol{x_1} \sim Normal(0,1)$ | $\boldsymbol{x_1} \sim Normal(0,1)$ |
| | | $\boldsymbol{x_2} \sim Normal(0,1)$ | $\boldsymbol{x_2} \sim Normal(0,1)$ |
| | | $\boldsymbol{x_3} \sim Normal(0,1)$ | $\boldsymbol{x_3} \sim Normal(0,1)$ |
| | $k = 3$ | | $k = 1$ |
| $\boldsymbol{b} = 2,2 - 0,8\boldsymbol{x}$ | $\boldsymbol{b} = 2,2 - 0,3\boldsymbol{x}$ | $\boldsymbol{\mu} = \exp(1 + 0,3\boldsymbol{x_1} + 0,3\boldsymbol{x_2})$ | $\boldsymbol{\mu} = \exp(1 + 0,3\boldsymbol{x_1} + 0,3\boldsymbol{x_2})$ |
| $\boldsymbol{p} = \exp(\boldsymbol{b})$ | $\boldsymbol{p} = k/(\exp(\boldsymbol{b}) + k)$ | | $\boldsymbol{p} = 1/(1 + \boldsymbol{\mu}/k)$ |
| $\boldsymbol{y} \sim Poisson(\boldsymbol{p})$ | $\boldsymbol{y} \sim BN(\boldsymbol{p}, k)$ | $\boldsymbol{y}_{poi} \sim Poisson(\boldsymbol{\mu})$ | $\boldsymbol{y}_{bn} \sim BN(\boldsymbol{p}, k)$ |
| | | $\boldsymbol{y}_{zero} \sim logit(2\boldsymbol{x_3})$ | $\boldsymbol{y}_{zero} \sim logit(2\boldsymbol{x_3})$ |

Para poder utilizar o modelo RBNIZGP (ou qualquer modelo espacial) é necessário que as observações possuam coordenadas geográficas, e para isso foram utilizados os condados da Georgia, EUA, analisado por (Da Silva e Fotheringham, 2016), sendo este um conjunto de dados bastante clássico em análises espaciais. O conjunto de dados é composto por 159 condados e apresenta como variável resposta a taxa de escolaridade e como variáveis preditoras algumas características da população. Assim, os dados da Tabela 6.1 foram atribuídos igualmente a todos os 159 condados, gerando dessa forma uma distribuição sem variação espacial.

A fim de verificar questões de processamento, acurácia e convergência do algoritmo, serão simulados $n = 100$ conjuntos de dados para cada variável resposta $\boldsymbol{y}$ descrita na Tabela 6.1. As distribuições estão na Figura 6.1.

### 6.2.2 Dados reais

Os dados reais a serem utilizados referem-se aos casos de COVID-19 na Coréia do Sul em 2020, sendo os mesmos dados utilizados por Weinstein et al. (2021). O conjunto de da-




(a) (b) (c) (d)

Figura 6.1: Distribuições simuladas (a) Poisson, (b) binomial negativa, (c) Poisson inflacionada de zeros e (d) binomial negativa inflacionada de zeros

dos é composto por 244 observações, sendo que a variável resposta utilizada na modelagem é o número de casos de COVID-19 na fase inicial da pandemia e as variáveis explicativas seriam: *MORBIDITY* (Comorbidade), *HIGH_SCH_P* (Proporção de pessoas com 2° grau completo), *HEALTHCARE_ACCESS* (Acesso à saúde), *DIFF_SD* (Dificuldade de distanciamento social), *CROWDING* (Aglomeração), *MIGRATION* (Migração) e *HEALTH_BEHAVIOR* (Comportamento de saúde).

A COVID-19 teve início na cidade Wuhan, na China, em dezembro de 2019 (Wikipédia, 2020). A porta de entrada em outros países ocorreu por meio dos aeroportos e em pouco tempo, o vírus se espalhou pelo mundo, devastando e matando também muitas pessoas (Wikipédia, 2020). A maior quantidade de casos esteve presente nas regiões de Seoul e Daegu, além dessas regiões serem bastante populosas, elas também estão próximas de grandes aeroportos e portanto, isso facilitou a circulação do vírus, resultando em número expressivo de casos da doença (Wikipédia, 2021).

Na Figura 6.2, que representa o mapa da Coréia do Sul, pode-se verificar a distribuição espacial desses dados de COVID-19 nas fases da pandemia e observa-se a grande concentração de zeros, caracterizando uma possível distribuição binomial negativa inflacionada de zeros. No entanto, analisaremos somente os dados referentes ao *Early Phase*, ou seja, a primeira fase da doença antes da quarentena, sendo uma fase em que a concentração de zeros foi muito maior do que nas outras fases.

A título de complementação, a Figura 6.3 mostra o mapa da Coréia do Sul com destaque para as regiões com a maior quantidade de casos: Seoul e Daegu.

## 6.3 Métodos

Como foi descrito na Seção 5.3 sobre o modelo binomial negativo inflacionado de zeros, a partir dos dados não espacialmente dependentes, o objetivo é verificar a potencialidade do algoritmo RBNIZGP em se ajustar às distribuições Poisson, binomial negativa e Poisson inflacionado de zeros. Para isso, o método a ser utilizado está descrito na Figura 6.4, ou seja, serão comparados os resulados do modelo RBNIZGP com a distribuição especificada (Poisson,

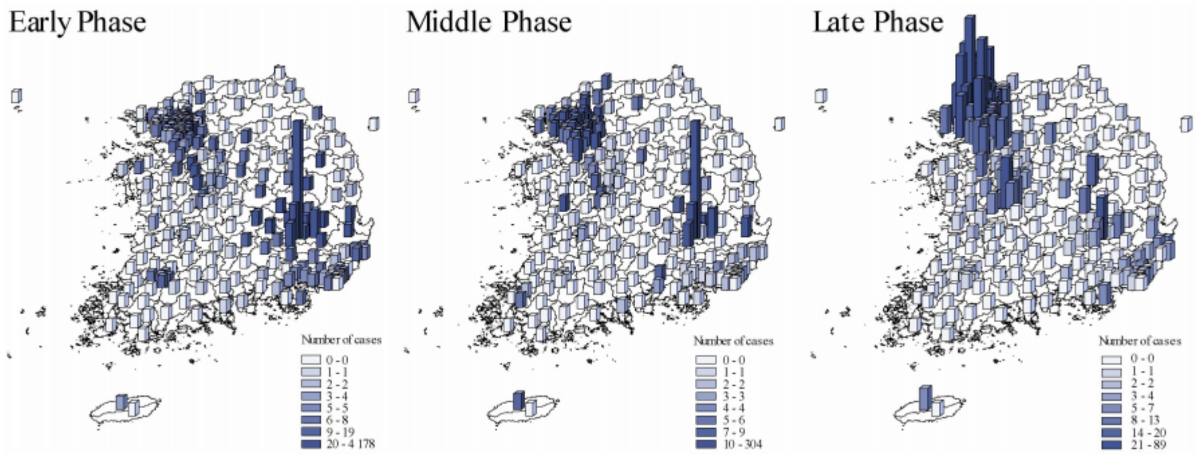
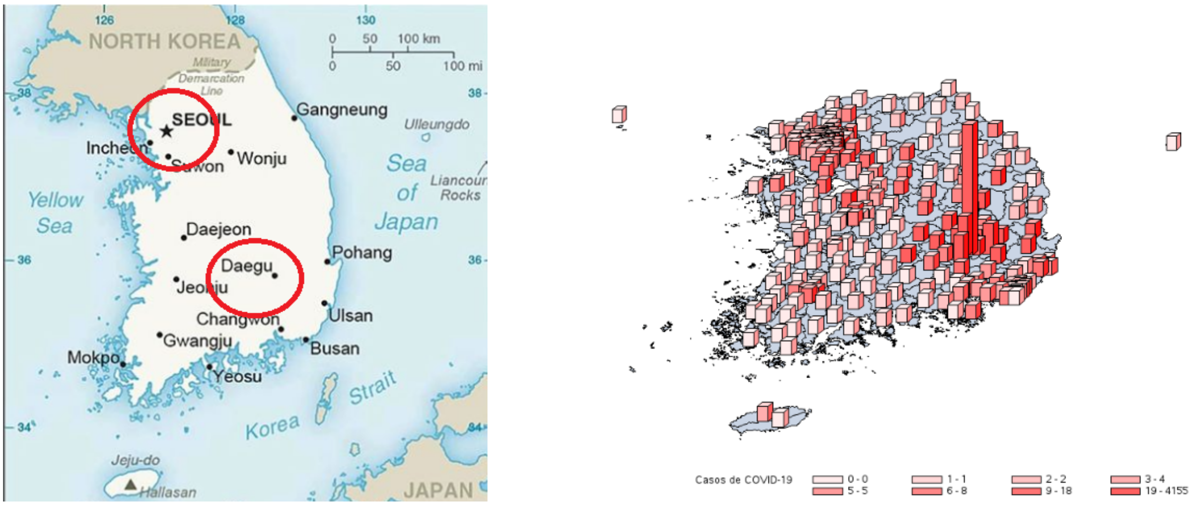



Figura 6.2: Distribuição espacial dos casos COVID-19 nas fases da pandemia na Coréia do Sul, 2020

Fonte: Weinstein et al. (2021)




(a) Fonte: Washington (2015) (b)

Figura 6.3: (a) Mapa da Coréia do Sul e (b) Mapa da Coréia do Sul (representação do número de casos de COVID-19 - Fase $Early$)

binomial negativa, Poisson inflacionada de zeros ou binomial negativa inflacionada de zeros), geradas a partir do modelo RBNIZGP ajustando os parâmetros da Figura 6.4.

Note que sabendo que os dados seguem uma distribuição de Poisson com os parâmetros especificados na Tabela 6.1, espera-se que o modelo RBNIZGP estime os parâmetros de superdispersão $\alpha$ e da parte inflacionada de zeros $\gamma$ como zeros (ou não significativos caso sejam

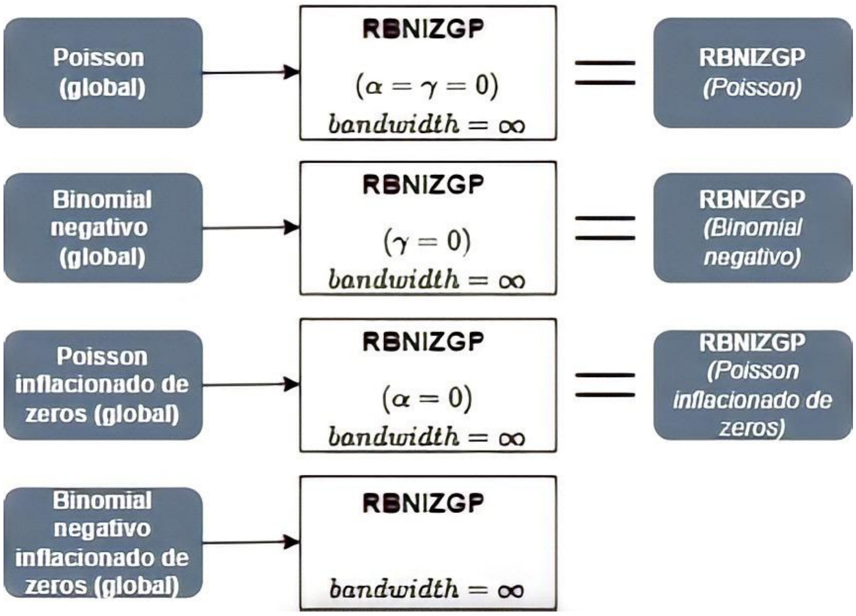



Figura 6.4: Estrutura dos parâmetros na RBNIZGP (Global)

diferentes de zero) e que o parâmetro de suavização (*bandwidth*) seja grande. Esse resultado será comparado com o modelo Poisson gerado a partir do modelo RBNIZGP forçando os parâmetros $\alpha = 0$ e $\gamma = 0$. O mesmo vale para as outras distribuições e parâmetros descritos na Figura 6.4.

### 6.3.1 Estudo de caso

A análise seguirá com um estudo de caso com dados reais, conforme o método ilustrado na Figura 6.5. Como foi descrito anteriormente nas Figuras 1.3 e 1.4, dependendo da localidade e do tamanho do parâmetro de suavização, o modelo local poderá ser Poisson, binomial negativo, Poisson inflacionado de zeros ou binomial negativo inflacionado de zeros.

Sendo assim, o objetivo principal deste estudo de caso consiste em verificar e comparar as diferenças que existem no modelo que ja foi publicado por Weinstein et al. (2021) (onde foi utilizada a distribuição binomial negativa), com a modelagem feita neste trabalho.

A título de ilustração, veja na Figura 6.5 que os parâmetros estimados na $i-$ésima observação pela RBNIZGP (com $\alpha = 0$ e $\gamma = 0$) podem ser comparados com os parâmetros estimados pela RPGP, assim como os parâmetros estimados na $j-$ésima observação pela RBNIZGP ($\gamma = 0$), podem ser comparados com os parâmetros estimados pela RBNGP, e assim sucessivamente para os demais modelos. No entanto, os parâmetros estimados na $j-$ésima observação pela RBNIZGP (com $\gamma = 0$) não podem ser comparados com os parâmetros estimados pela RPGP e nem pela RPIZGP (regressão Poisson inflacionada de zeros geograficamente ponderada).

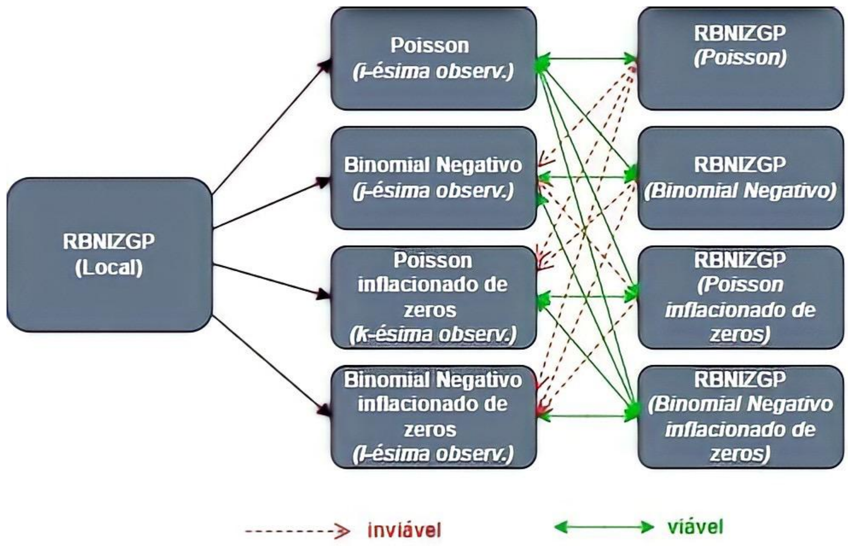



Figura 6.5: Relação entre os modelos na RBNIZGP (Local)

Na Figura 6.6, pode-se observar a relação que existe entre um determinado modelo e a estimação de um dado específico. Sendo assim, partindo de um modelo de Poisson, é possível fazer a estimação de um dado Poisson, no entanto, a estimação de dados que seguem outras distribuições (binomial negativa, Poisson inflacionada de zeros e binomial negativa inflacionada de zeros) não é possível ser feita. A mesma ideia pode ser observada nos outros casos, por exemplo, partindo de um modelo binomial negativo, a estimação para um dado que seja binomial negativo é viável, assim como a estimação de um dado Poisson, neste último caso considerando que o parâmetro de superdispersão seja $\alpha = 0$.

No caso do modelo Poisson inflacionado de zeros, nota-se que a estimação pode ocorrer para dados, cuja distribuição seja Poisson inflacionada de zeros ou simplesmente Poisson, neste último caso considerando que $\gamma = 0$. Por último, observa-se que a partir de um modelo binomial negativo inflacionado de zeros, a estimação de um dado, cuja distribuição seja binomial negativa inflacionada de zeros pode ocorrer, assim como é possível que ocorra a estimação de um dado Poisson inflacionado de zeros (se $\alpha = 0$), de um dado binomial negativo (se $\gamma = 0$) ou ainda de um dado Poisson (se $\gamma = 0$ e $\alpha = 0$). Portanto, pode-se concluir a partir da Figura 6.6 que um modelo menos geral não estima um modelo mais geral.

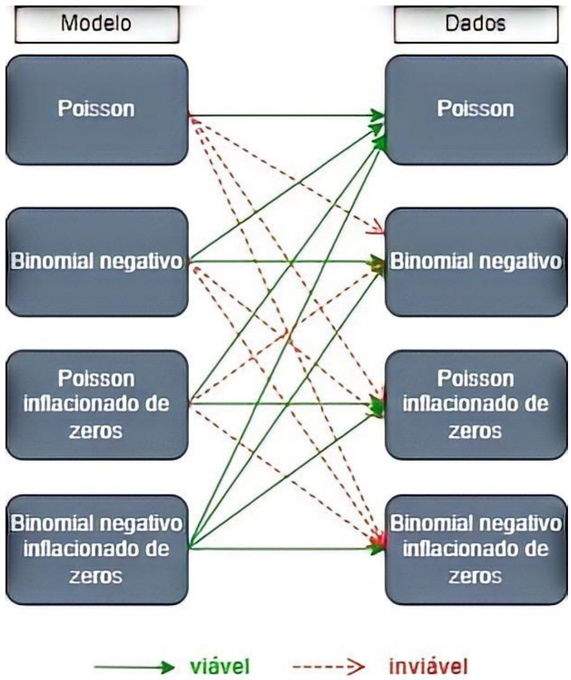



Figura 6.6: Relação entre modelo de regressão e estimação de dados

# Capítulo 7

# Resultados

## 7.1 Introdução

Este Capítulo apresenta os resultados decorrentes do método apresentado no Capítulo anterior, seguindo a estrutura apresentada nas Figuras 6.4 e 6.5. O *software* SAS 9.4 foi utilizado para a construção do algortimo, para a simulação dos dados e para o processamento dos resultados, sendo a *PROC GENMOD* utilizada como parâmetro de comparação das estimativas obtidas pela RBNIZGP.

## 7.2 Dados simulados

A Figura 7.1 mostra a distribuição dos dados simulados no condado da Georgia/EUA. Pode-se observar que os dados estão aleatoriamente distribuídos na região, visto que os mesmos não são espacialmente dependentes. Para fins de comparação com os resultados apresentados, a distância máxima calculada entre os centróides das regiões foi de aproximadamente 690 km. Nos mapas que representam os dados simulados para as distribuições Poisson inflacionada de zeros e binomial negativa inflacionada de zeros (Figura 7.1(c) e (d)), percebe-se uma quantidade maior de zeros espalhados pelas regiões, caracterizando dessa forma distribuições inflacionadas de zeros e não espacialmente dependentes.

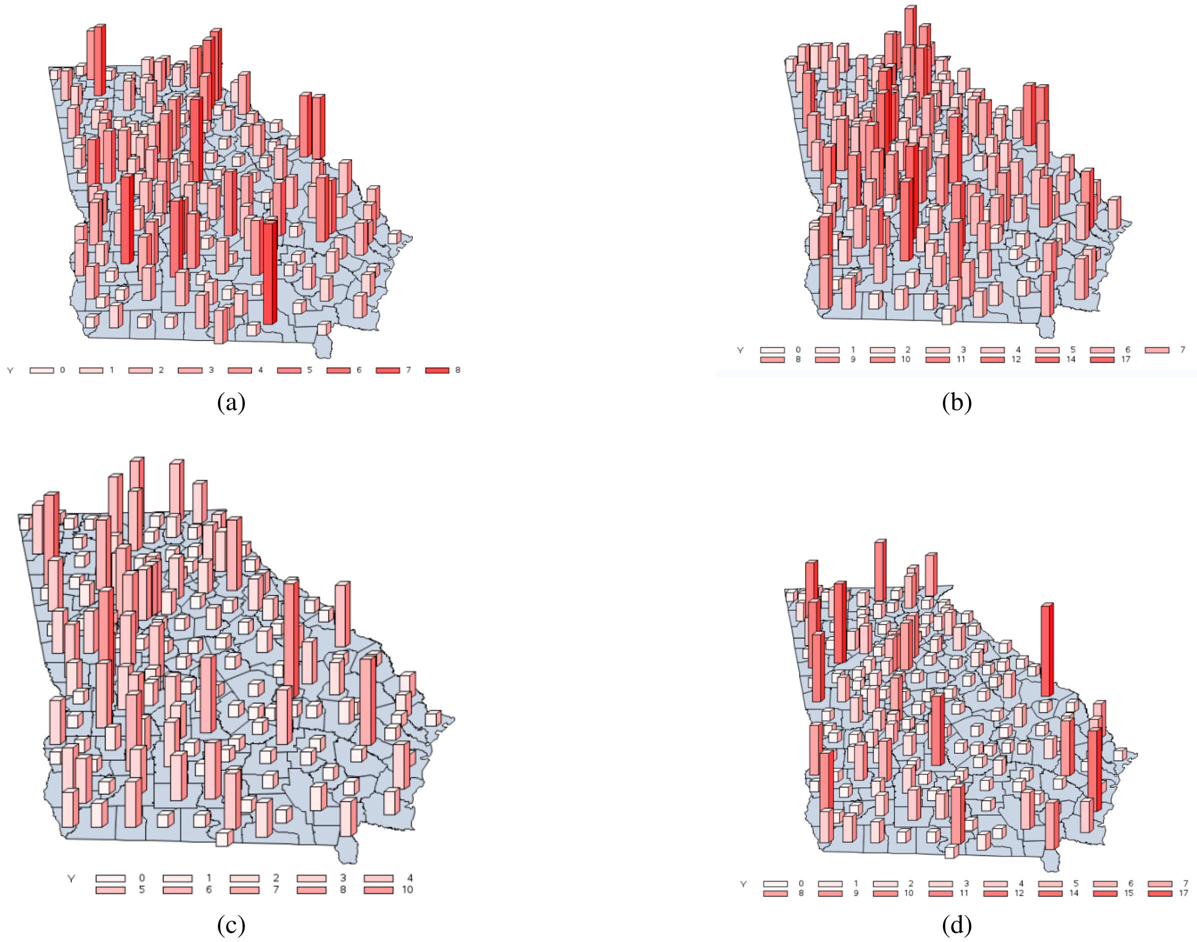


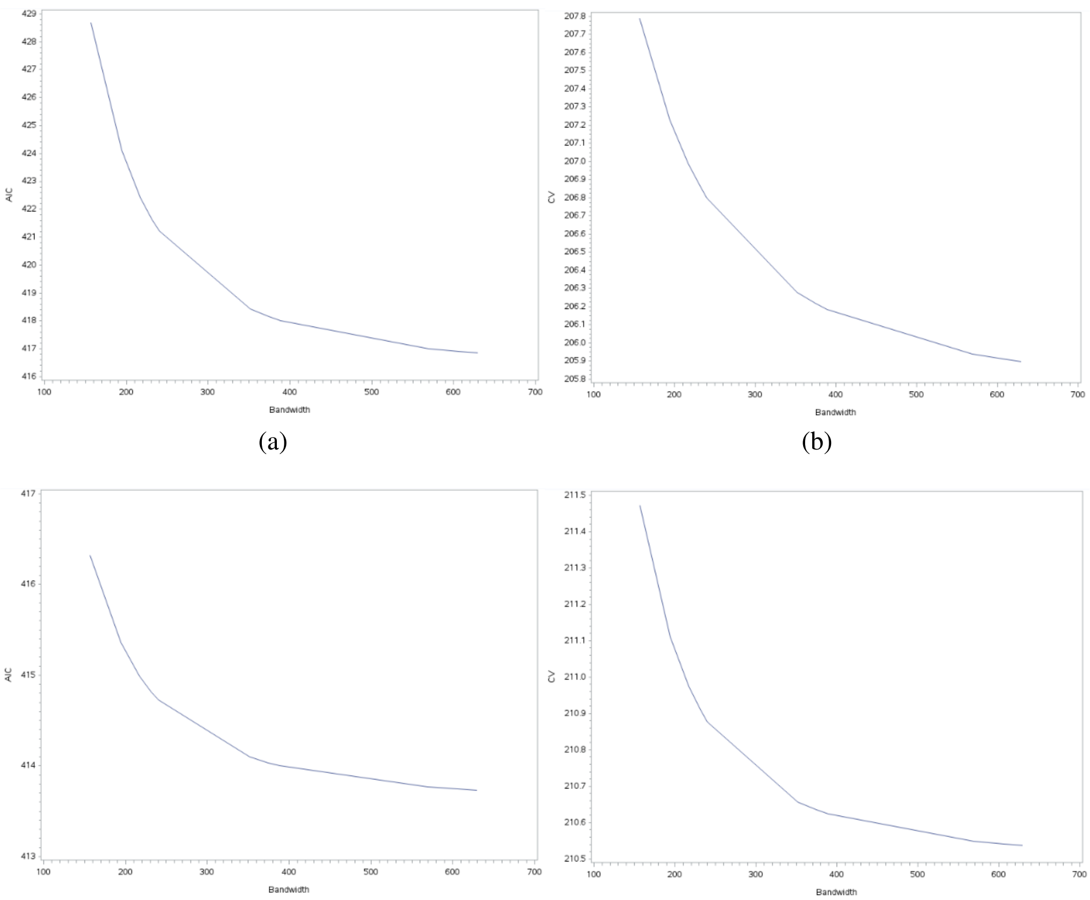
Figura 7.1: Distribuição dos dados simulados (a) Poisson, (b) binomial negativo, (c) Poisson inflacionado de zeros e (d) binomial negativo inflacionado de zeros

### 7.2.1 Simulação com dados Poisson

Ao testar os dados simulados a partir de uma distribuição Poisson, foram utilizados três algoritmos: *PROC GENMOD* do SAS 9.4, Algoritmo RBNIZGP (utilizando a binomial negativa inflacionada de zeros) e Algoritmo RBNIZGP (utilizando a Poisson, ou seja, fazendo $\alpha = 0$ e $\gamma = 0$).

O primeiro passo para utilizar os modelos geograficamente ponderados é encontrar o parâmetro de suavização, que indicará se os dados apresentam dependência espacial ou não. Quanto maior ele for, mais indícios tem-se sobre a não existencia de dependência espacial. A Figura 7.2

mostra as funções $AIC$ e $CV$ em função da distância. Note que em todos os casos, o parâmetro de suavização (distância que gera o menor $AIC/CV$) tende à maior distância entre os dados, caracterizando dessa forma a não dependência espacial, e que o mesmo só não é maior por estar limitado pela máxima distância entre os pontos, conforme o algoritmo *Golden Section Search*. Assim, recomenda-se usar um grande valor para o parâmetro de suavização a fim do algoritmo RBNIZGP gerar todas as estimativas iguais nas regiões. Caso use esse parâmetro de suavização igual a 628.8850 km, podem aparecer pequenas diferenças nos parâmetros entre as diferentes regiões.




(a) (b) (c) (d)

Figura 7.2: Esboço da Função (a) $AIC$ - RBNIZGP (binomial negativa inflacionada de zeros), (b) $CV$ - RBNIZGP (binomial negativa inflacionada de zeros), (c) $AIC$ - RBNIZGP (Poisson) e (d) $CV$ - RBNIZGP (Poisson)

A Tabela 7.1 mostra, para fins ilustrativos, os resultados de uma das 100 bases simuladas, e nota-se que as estimativas das variáveis preditoras (Intercepto e $X$) foram exatamente iguais entre a *PROC GENMOD* e o Algoritmo RBNIZGP (Poisson), e percebe-se ainda que tais estimativas foram consideradas significativas a um nível de significância de $\alpha = 5\%$. Os valores de $AIC$, *Deviance* e Log-verossimilhança também são os mesmos. Pode-se observar também que os resultados estão muito próximos do que foi definido nas simulações da Tabela 6.1, ou seja, Intercepto igual a 2,2 e parâmetro referente à variável $X$ igual -0,8.

Tabela 7.1: Dados simulados - distribuição Poisson

| **Variáveis/ Estatísticas** | **Poisson** | | | | | |
|---|---|---|---|---|---|---|
| | ***PROC GENMOD*** | | ***RBNIZGP (binomial negativo inflacionado de zeros)*** | | ***RBNIZGP (Poisson)*** | |
| | *Estimativa* | *Erro padrão* | *Estimativa* | *Erro padrão* | *Estimativa* | *Erro padrão* |
| Intercepto | $2,3039^{*}$ | $0,1503$ | $2,1800^{*}$ | $0,1752$ | $2,3039^{*}$ | $0,1503$ |
| $X$ | $-0,7891^{*}$ | $0,0663$ | $-0,7099^{*}$ | $0,0878$ | $-0,7890^{*}$ | $0,0663$ |
| $\alpha$ | - | - | $0,0000^{\text{NS}}$ | $0,0001$ | - | - |
| Intercepto (inflacionado) | - | - | $-11,8970^{**}$ | $6,4578$ | - | - |
| $X$ (inflacionado) | - | - | $2,7570^{**}$ | $1,4794$ | - | - |
| Parâmetro de suavização | - | | $628,8850$ | | $628,8850$ | |
| $AIC$ | $413,4930$ | | $415,1466$ | | $413,4930$ | |
| $Deviance$ | $152,2006$ | | $148,4620$ | | $152,2004$ | |
| Log-verossimilhança | $-204,7465$ | | $-202,8774$ | | $-204,7465$ | |

$(*)$ Significativo a 5%
$(**)$ Significativo a 10%
NS Não Significativo a 10%

No caso do algoritmo RBNIZGP (binomial negativo inflacionado de zeros), percebe-se que a estimativa para $\alpha$ foi zero, como era esperado, e as estimativas para a parte inflacionada de zeros foram consideradas significativas para um nível de significância de $\alpha = 10\%$, mas não para $\alpha = 5\%$. Isso influenciou um pouco nas estimativas das variáveis preditoras (Intercepto e $X$). Para verificar isso, veja que para um registro da covariável $\boldsymbol{G}$, $\exp(G_i\boldsymbol{\gamma})/(1+\exp(G_i\boldsymbol{\gamma})) = \exp(1\times(-11,897)+4,3118\times 2,757)/(1+\exp(1\times(-11,897)+4,3118\times 2,757)) = 0.4977$, que é diferente de zero. Ainda assim, veja que tanto a $Deviance$ quanto a Log-verossimilhança tiveram resultados mais satisfatórios do que na regressão Poisson. Mas considerando que as variáveis inflacionadas foram não significativas para $\alpha = 5\%$, elas podem ser removidas do modelo, gerando um resultado igual à regressão Poisson.

A Figura 7.3 mostra os *box-plot* das estimativas das $n = 100$ simulações para as variáveis preditoras (Intercepto e $X$) de cada modelo gerado pelos três algoritmos. Com os resultados, pode-se observar que nos Interceptos gerados na *PROC GENMOD* houve uma amplitude um pouco superior aos demais algoritmos, trazendo um valor máximo maior e um valor mínimo menor que os demais, entretanto, observa-se que a média e mediana são muito próximas de 2,2. Os resultados para as estimativas dos coeficientes de $X$ seguem a mesma tendência com média e mediana próximas à -0,8 nos três algoritmos.

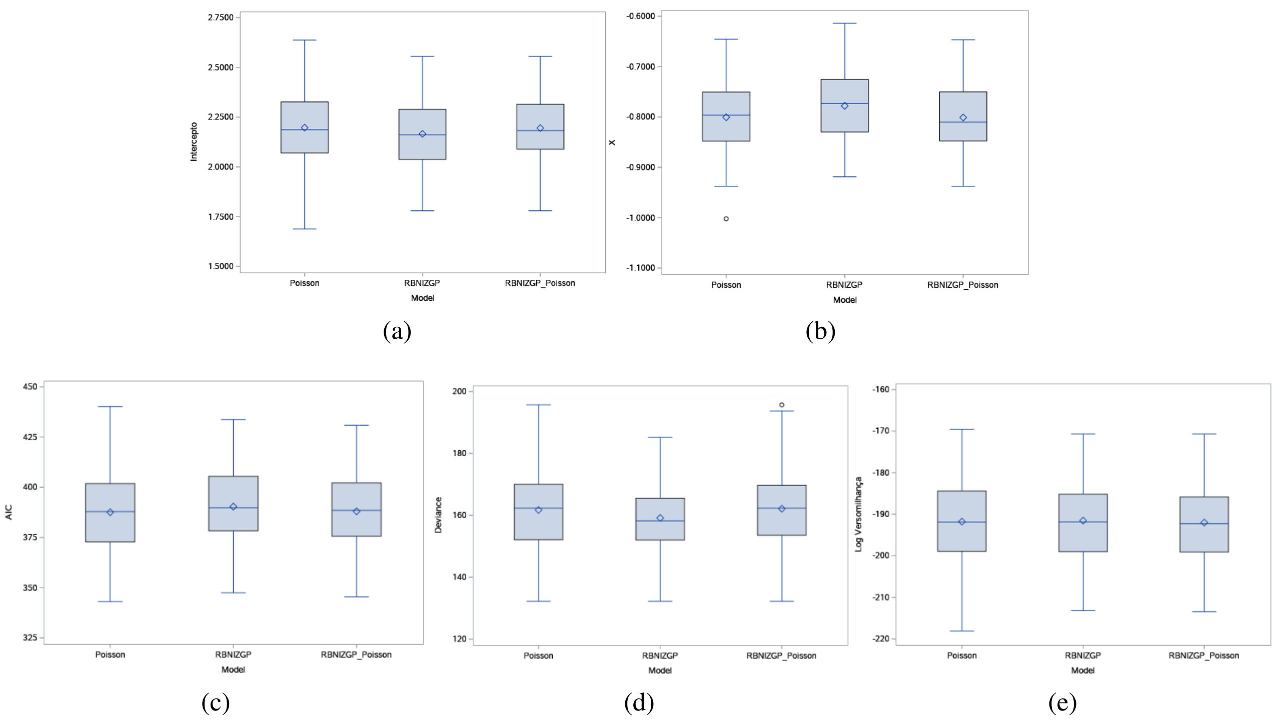
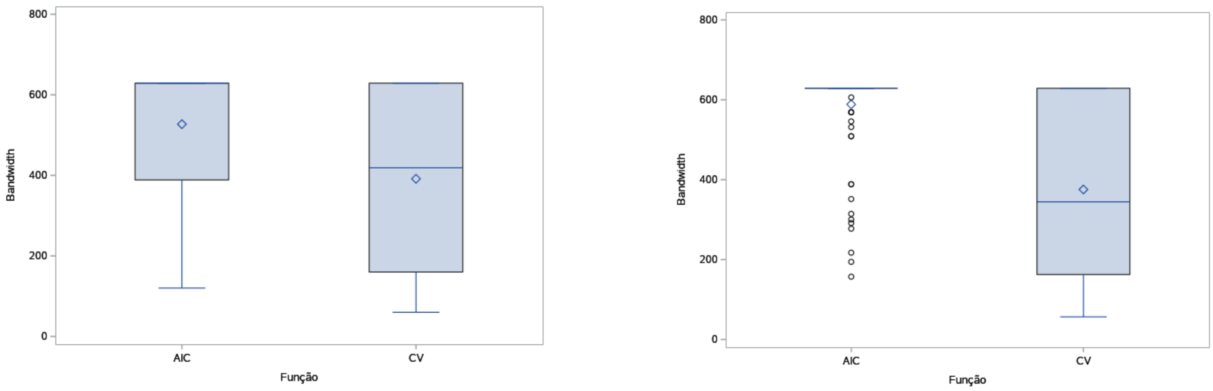



Figura 7.3: *box-plot* - Resultados da modelagem com base na simulação dados Poisson: (a) Intercepto, (b) $X_1$, (c) $AIC$, (d) *Deviance* e (e) Log-verossimilhança

Pode-se verificar também os resultados das estimativas de $AIC$, *Deviance* e Log-verossimilhança. Observa-se que não existe muita diferença entre os três algoritmos nas três medidas.

A Figura 7.4 mostra os *box-plot* dos parâmetros de suavização estimados pelos Algoritmos RBNIZGP (utilizando a binomial negativa inflacionada de zeros) e RBNIZGP (Poisson) minimizando $AIC$ e $CV$. Nos resultados obtidos via algoritmo RBNIZGP (Poisson), pode-se verificar que houve uma diferença entre os resultados da função $AIC$ e $CV$; no caso do $AIC$, é possível notar que a média do valor mínimo está um pouco acima de 500 km, entretanto, no caso da $CV$, a média está próxima de 400 km e há uma variabilidade maior nos dados, comparando com o $AIC$.

Já no caso dos resultados obtidos pelo algoritmo RBNIZGP (binomial negativa inflacionada de zeros), no caso do $AIC$ é possível verificar que praticamente todos os dados se concentraram no valor 628.885 km, com exceção de pontos discrepantes existentes na distribuição, diferentemente da $CV$ que houve uma variação maior de valores, mostrando que $25\%$ dos dados se concentram em uma faixa inferior a 200 km e a sua média é inferior a 400 km.




(a) (b)

Figura 7.4: *box-plot* do parâmetro de suavização (Dados simulados Poisson) para as funções (a) $AIC$ e $CV$ - RBNIZGP (Poisson) e (b) $AIC$ e $CV$ - RBNIZGP (binomial negativa inflacionada de zeros)

### 7.2.2 Simulação com dados binomial negativo

Como no caso anterior, ao testar os dados simulados a partir de uma distribuição binomial negativa, foram utilizados três algoritmos, sendo: *PROC GENMOD* do SAS 9.4, Algoritmo

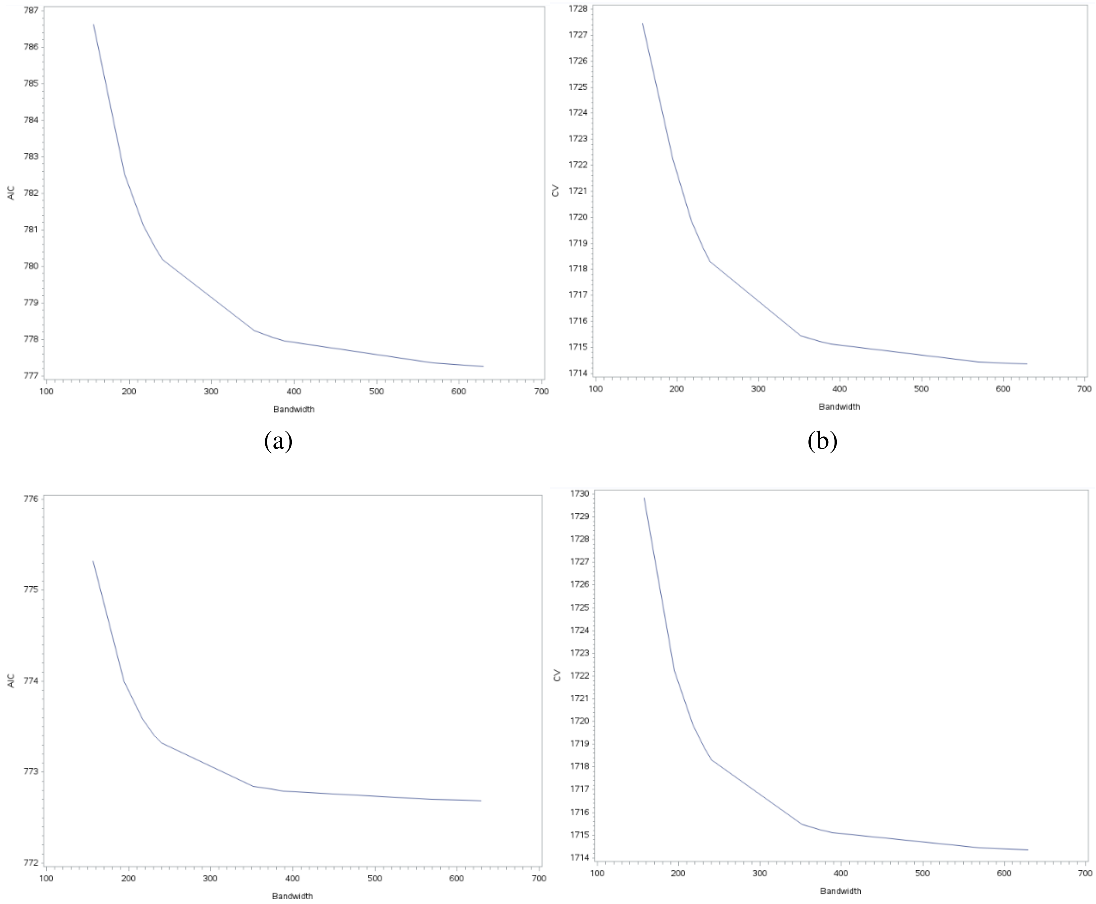
RBNIZGP (utilizando a binomial negativa inflacionada de zeros) e o Algoritmo RBNIZGP (utilizando a binomial negativa, ou seja, fazendo $\boldsymbol{\gamma} = 0$).

Assim como no caso da Poisson, pode-se observar na Figura 7.5 que tanto minimizando os critérios $AIC$ ou $CV$, para um parâmetro de suavização fixo ou adaptável, o parâmetro de suavização tende à maior distância entre os dados e ainda tal parâmetro só não é maior por estar limitado pela distância máxima entre os pontos, em conformidade com o algoritmo *Golden Section Search*. Portanto, é possível concluir que não existe dependência espacial e é recomendável utilizar também um grande valor para o parâmetro de suavização, para que o algoritmo RBNIZGP possa gerar todas estimativas iguais nas regiões.

Os resultados de uma das 100 bases simuladas estão resumidos na Tabela 7.2 e pode-se observar que as estimativas das variáveis preditoras (Intercepto, $X$ e $\alpha$) estão condizentes com os dados que foram definidos para a simulação, na Tabela 6.1, sendo o Intercepto igual a 2,2, o parâmetro referente à variável $X$ igual a 0,3 e $\alpha = \frac{1}{3}$. As estimativas das variáveis preditoras (Intercepto, $X$ e $\alpha$) ficaram muito próximas, sendo praticamente iguais nos três algoritmos e ainda foram consideradas significativas a um nível de significância $\alpha = 5\%$. Os valores das medidas de ajuste também ficaram iguais, com exceção do $AIC$ no Algoritmo RBNIZGP (binomial negativo inflacionado de zeros).

No caso do Algoritmo RBNIZGP (binomial negativo inflacionado de zeros), percebe-se que foram geradas estimativas para a parte inflacionada de zeros, no entanto, não foram consideradas significativas a um nível de significância $\alpha = 5\%$. Percebe-se ainda que esses valores da parte inflacionada não influenciaram nas estimativas das outras variáveis preditoras (Intercepto, $X$ e $\alpha$). Como exemplo, utilizando um registro da covariável $\boldsymbol{G}$, tem-se: $\exp(G_i\boldsymbol{\gamma})/(1 + \exp(G_i\boldsymbol{\gamma})) = \exp(1 \times (-18,0443) + 4,3118 \times (-0.1393))/(1 + \exp(1 \times (-18,0443) + 4,3118 \times (-0.1393)) = 0$, portanto, é notável que tais estimativas praticamente não fizeram efeito nas outras variáveis preditoras.

Os *box-plot* das estimativas das $n = 100$ variáveis preditoras (Intercepto, $X$ e $\alpha$) de cada modelo gerado pelos três algoritmos estão ilustrados na Figura 7.6. Com os resultados, pode-se




(a) (b) (c) (d)

Figura 7.5: Esboço da Função (a) $AIC$ - RBNIZGP (binomial negativa inflacionada de zeros), (b) $CV$ - RBNIZGP (binomial negativa inflacionada de zeros), (c) $AIC$ - RBNIZGP (binomial negativa) e (d) $CV$ - RBNIZGP (binomial negativa)

observar que nos Interceptos gerados pelo Algoritmo RBNIZGP (binomial negativa inflacionada de zeros) houve uma amplitude um pouco superior aos demais algoritmos, trazendo um valor máximo maior que os demais, entretanto, observa-se que a média e mediana são muito próximas de 2,2. É notável também a presença de pontos discrepantes nos três algoritmos (tanto nos resultados de Intercepto e da variável $X$). Os resultados para as estimativas dos coeficientes de $X$ seguem a mesma tendência com média e mediana próximas à -0,3 nos três algoritmos.

Já nos resultados das estimativas de $AIC$, *Deviance* e Log-verossimilhança é perceptível que existe uma tendência similar nos três algoritmos para os casos do $AIC$ e Log-verossimilhança.

Tabela 7.2: Dados simulados - modelo binomial negativo

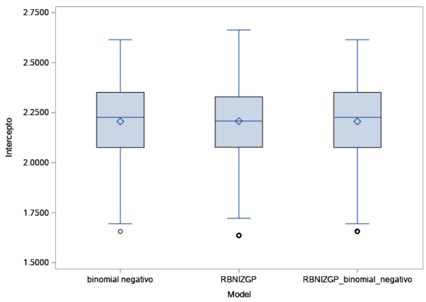
| **Variáveis/ Estatísticas** | **binomial negativa** | | | | | |
|---|---|---|---|---|---|---|
| | ***PROC GENMOD*** | | ***RBNIZGP (binomial negativo inflacionado de zeros)*** | | ***RBNIZGP (binomial negativo)*** | |
| | *Estimativa* | *Erro padrão* | *Estimativa* | *Erro padrão* | *Estimativa* | *Erro padrão* |
| Intercepto | $2,1340^*$ | $0,1709$ | $2,1339^*$ | $0,1708$ | $2,1339^*$ | $0,1708$ |
| $X$ | $-0,2462^*$ | $0,0576$ | $-0,2462^*$ | $0,0575$ | $-0,2462^*$ | $0,0575$ |
| $\alpha$ | $0,3173^*$ | $0,0645$ | $0,3172^*$ | $0,0645$ | $0,3173^*$ | $0,0645$ |
| Intercepto (inflacionado) | - | - | $-18,0443^{NS}$ | $4512,8199$ | - | - |
| $X$ (inflacionado) | - | - | $-0,1393^{NS}$ | $1490,7306$ | - | - |
| Parâmetro de suavização | - | | $628,8850$ | | $628,8850$ | |
| $AIC$ | $772,4940$ | | $776,4942$ | | $772,4940$ | |
| $Deviance$ | $171,3346$ | | $171,3346$ | | $171,3346$ | |
| Log-verossimilhança | $-383,2470$ | | $-383,2470$ | | $-383,2470$ | |

$(*)$ Significativo a 5%
(NS) Não Significativo a 5%

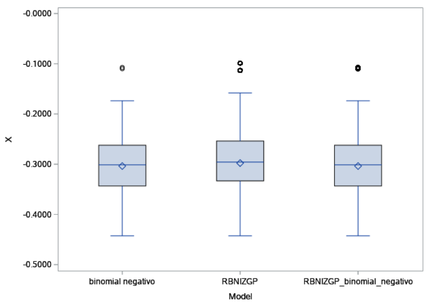
No entanto, na *Deviance*, é possível observar que no resultado do Algoritmo RBNIZGP (binomial negativa inflacionada de zeros) houve uma amplitude muito maior, gerando um valor máximo maior que os demais.

A Figura 7.7 mostra os *box-plot* dos parâmetros de suavização estimados pelos Algoritmos RBNIZGP (utilizando a binomial negativa inflacionada de zeros) e RBNIZGP (binomial negativo) minimizando $AIC$ e $CV$. Nos resultados obtidos via algoritmo RBNIZGP (binomial negativo), pode-se verificar que houve uma diferença entre os resultados da função $AIC$ e $CV$; no caso do $AIC$, é possível verificar a existência de *outliers* abaixo do limite inferior e possível notar também que a média do valor mínimo está um pouco acima de 500 km, entretanto, no caso da $CV$, a média está próxima de 400 km e há uma variabilidade maior nos dados, comparando com o $AIC$.

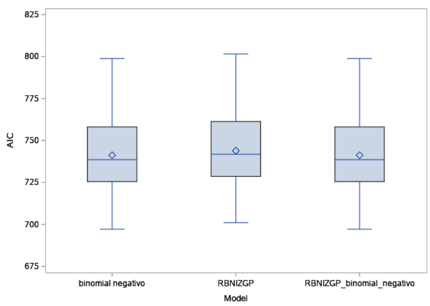
No caso dos resultados obtidos pelo algoritmo RBNIZGP (binomial negativa inflacionada de zeros), no caso do $AIC$ é possível verificar que há uma variabilidade menor nos dados, comparando com o $CV$ e também é possível verificar a existência de *outliers* abaixo do limite



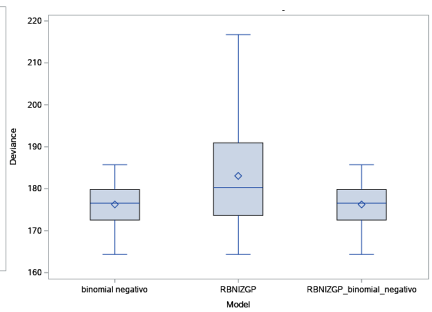

(a)



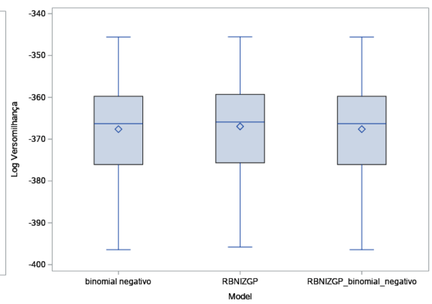

(b)




(c)

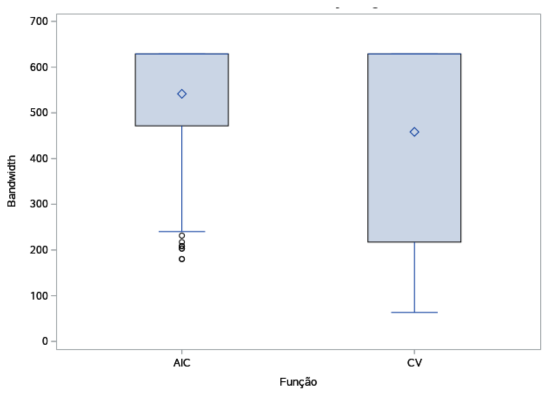



(d)

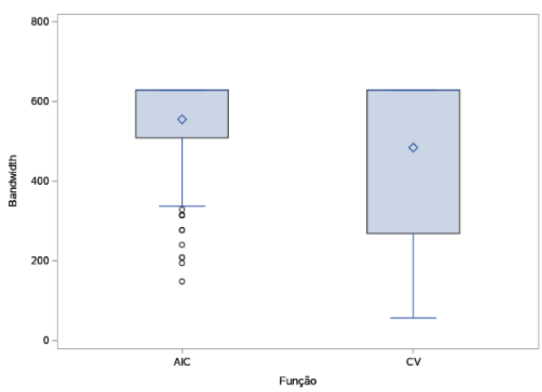



(e)

Figura 7.6: *box-plot* - Resultados da modelagem com base na simulação dados da binomial negativa: (a) Intercepto, (b) $X_1$, (c) $AIC$, (d) *Deviance* e (e) Log-verossimilhança




(a)




(b)

Figura 7.7: *box-plot* do parâmetro de suavização (Dados simulados binomial negativa) para as funções (a) $AIC$ e $CV$ - RBNIZGP (binomial negativa) e (b) $AIC$ e $CV$ - RBNIZGP (binomial negativa inflacionada de zeros)

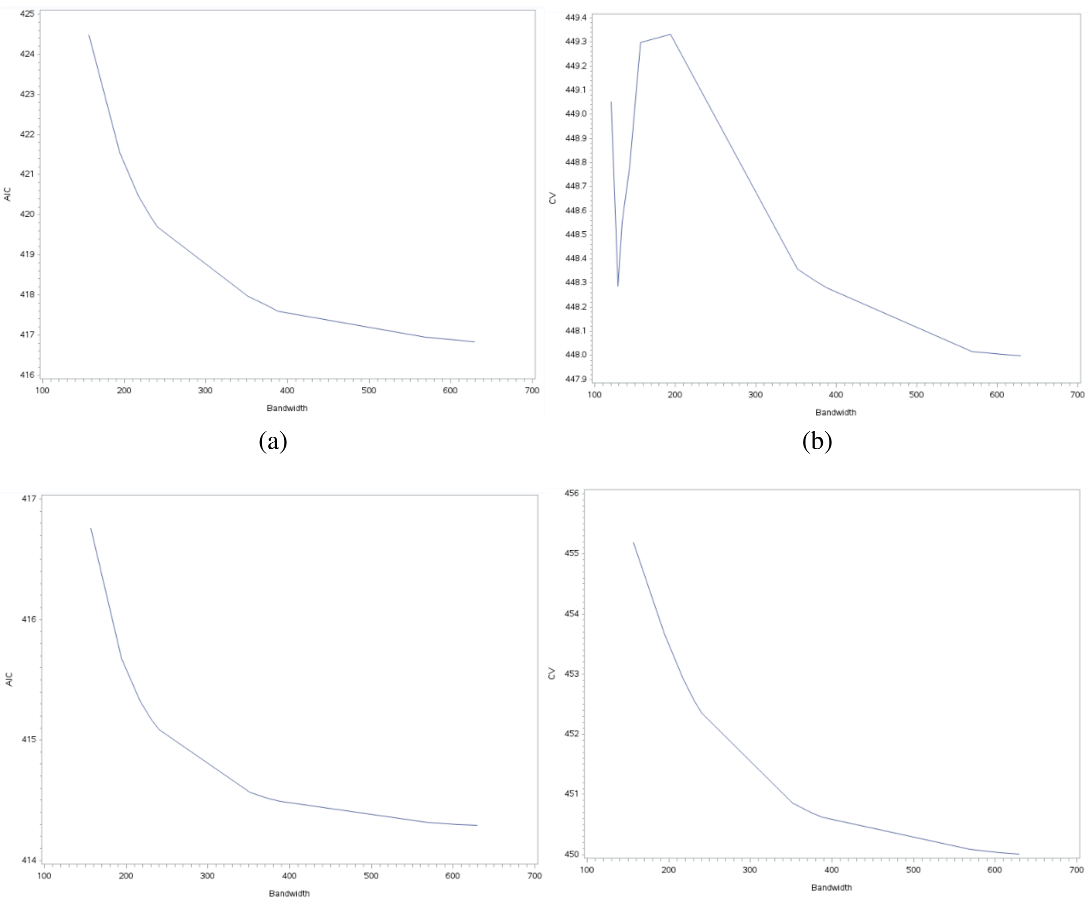
inferior. Já no caso dos resultados da função $CV$, pode-se perceber que houve uma variação maior de valores, mostrando que $25\%$ dos dados se concentram em uma faixa superior a 200 km e a sua média é superior a 400 km.

### 7.2.3 Simulação com dados da Poisson inflacionada de zeros

Para os dados simulados a partir de uma distribuição Poisson inflacionada de zeros, foram utilizamos três algoritmos, sendo: *PROC GENMOD* do SAS, Algoritmo RBNIZGP (utilizando a binomial negativa inflacionada de zeros) e o Algoritmo RBNIZGP (utilizando a Poisson inflacionada de zeros, ou seja, fazendo $\alpha = 0$).

Como foi feito anteriormente para a Poisson e binomial negativa, o passo inicial para utilização dos modelos geograficamente ponderados é encontrar o parâmetro de suavização. Quanto maior ele for, mais indícios tem-se sobre a não existencia de dependência espacial. A Figura 7.8 mostra as funções $AIC$ e $CV$ em função da distância. É possível perceber que em todos os casos, o parâmetro de suavização tende à maior distância entre os dados, mostrando que não existe dependência espacial, e que o mesmo só não é maior por estar limitado pela máxima distância entre os pontos, conforme o algoritmo *Golden Section Search*. Portanto, também recomenda-se usar um grande valor para o parâmetro de suavização a fim do algoritmo RBNIZGP gerar todas as estimativas iguais nas regiões.

Os resultados de uma das 100 bases simuladas estão ilustrados na Tabela 7.3. Nota-se que as estimativas das variáveis preditoras (Intercepto, $X_1$, $X_2$, Intercepto (inflacionado) e $X_3$ (inflacionado)) foram exatamente iguais entre a *PROC GENMOD* e o Algoritmo RBNIZGP (Poisson inflacionado de zeros), e percebe-se ainda que tais estimativas foram consideradas significativas a um nível de significância de $\alpha = 5\%$, com exceção do Intercepto (inflacionado). Tais resultados estão muito próximos do que foi definido nas simulações da Tabela 6.1, ou seja, Intercepto igual a 1,0, o parâmetro referente à variável $X_1$ igual 0,3, o parâmetro referente à variável $X_2$ igual 0,3 e o parâmetro referente à variável $X_3$ (inflacionada) igual 2,0. O motivo para que o Intercepto (inflacionado) não ter sido considerado significativo a um nível de significância de $\alpha = 5\%$ se deve ao fato de não ter sido definido na simulação, como pode ser visto na Tabela 6.1.




(a) (b) (c) (d)

Figura 7.8: Esboço da Função (a) $AIC$ - RBNIZGP (binomial negativa inflacionada de zeros), (b) $CV$ - RBNIZGP (binomial negativa inflacionada de zeros), (c) $AIC$ - RBNIZGP (Poisson inflacionada de zeros) e (d) $CV$ - RBNIZGP (Poisson inflacionada de zeros)

No resultado gerado pelo algoritmo RBNIZGP (binomial negativo inflacionado de zeros), pode-se observar que a estimativa de $\alpha = 0.0178$, não foi significativa para um nível de significância de $5\%$. No entanto, como o valor de $\alpha$ não é exatamente zero, ele influência nas outras estimativas e por isso que essas estimativas não foram exatamente iguais aos resultados dos outros algoritmos.

Os valores de $AIC$ e a Log-verossimilhança também são os mesmos nos casos da *PROC GENMOD* e do Algoritmo RBNIZGP (Poisson inflacionado de zeros). Já a *Deviance* gerada pela *PROC GENMOD* é o dobro da Log-verossimilhança, visto que o SAS 9.4 não faz o cálculo da *Deviance*, e sim atribui o dobro da Log-verossimilhança para essa estatística (SAS, 2011).

Tabela 7.3: Dados simulados - modelo Poisson inflacionado de zeros

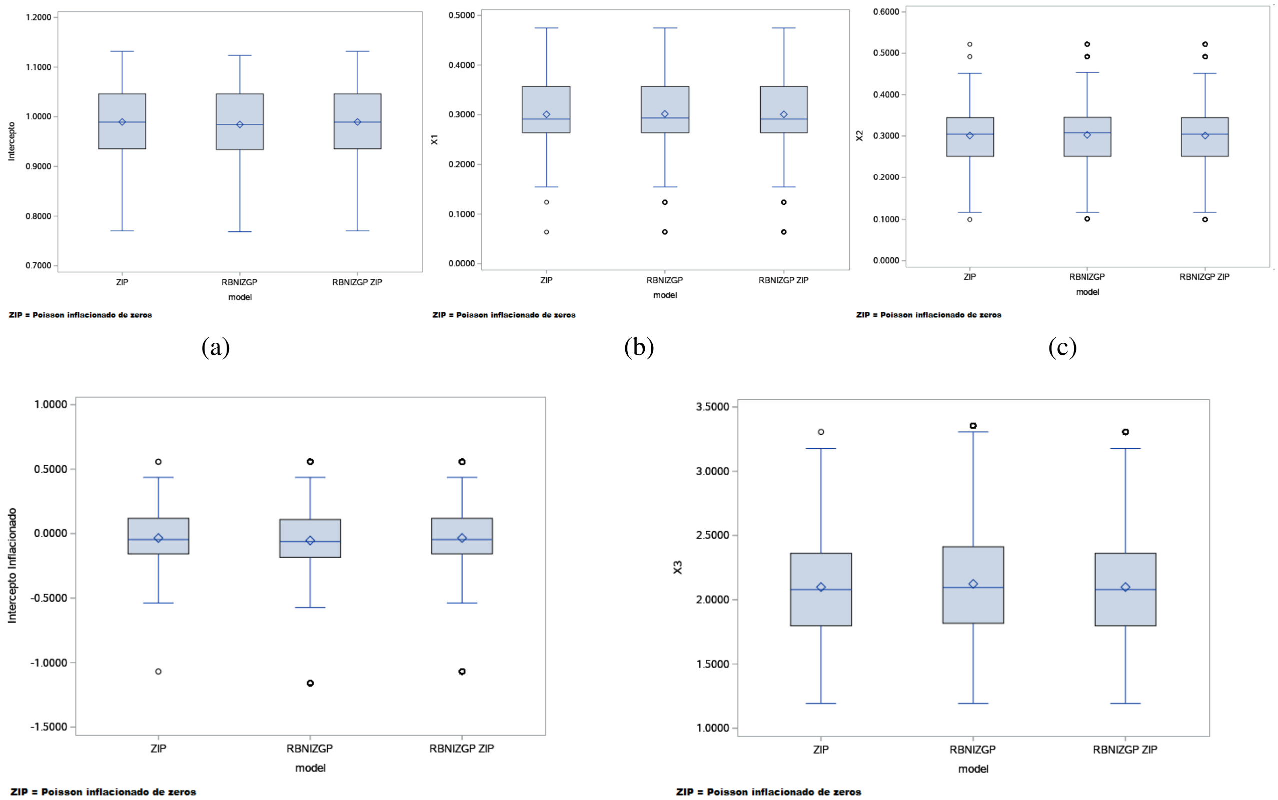
| **Variáveis/ Estatísticas** | **Poisson inflacionado de zeros** | | | | | |
|---|---|---|---|---|---|---|
| | ***PROC GENMOD*** | | ***RBNIZGP (binomial negativo inflacionado de zeros)*** | | ***RBNIZGP (Poisson inflacionado de zeros)*** | |
| | *Estimativa* | *Erro padrão* | *Estimativa* | *Erro padrão* | *Estimativa* | *Erro padrão* |
| Intercepto | $0,9709^*$ | $0,0823$ | $0,9654^*$ | $0,0831$ | $0,9709^*$ | $0,0823$ |
| $X_1$ | $0,3202^*$ | $0,0698$ | $0,3241^*$ | $0,0714$ | $0,3202^*$ | $0,0698$ |
| $X_2$ | $0,3178^*$ | $0,0721$ | $0,3219^*$ | $0,0741$ | $0,3178^*$ | $0,0721$ |
| $\alpha$ | - | - | $0,0178^{\text{NS}}$ | $0,0352$ | - | - |
| Intercepto (inflacionado) | $0,5004^{\text{NS}}$ | $0,2740$ | $0,5198^{\text{NS}}$ | $0,2778$ | $0,5003^{\text{NS}}$ | $0,2740$ |
| $X_3$ (inflacionado) | $2,3408^*$ | $0,4092$ | $2,3533^*$ | $0,4138$ | $2,3407^*$ | $0,4092$ |
| Parâmetro de suavização | - | | $628,8850$ | | $628,8850$ | |
| $AIC$ | $413,7961$ | | $415,6859$ | | $413,7962$ | |
| $Deviance$ | $403,7961$ | | $187,7819$ | | $192,0025$ | |
| Log-verossimilhança | $-201,8980$ | | $-201,8430$ | | $-201,8980$ | |

$(*)$ Significativo a 5%
(NS) Não Significativo a 5%

A Figura 7.9 mostra os *box-plot* das estimativas das $n = 100$ variáveis preditoras de cada modelo gerado pelos três algoritmos. Pode-se observar que os Interceptos, no três algoritmos, possuem um padrão similar, mostrando também que a média e mediana estão muito próximas de 1,0. O mesmo ocorre com as variáveis preditoras ($X_1$, $X_2$), se concentrando em valores próximos à 0,3. Já nos casos das variáveis preditoras inflacionadas de zeros (Intercepto e $X_3$) também é possível verificar um comportamento similar entre os três algoritmos.

Nas estimativas do $AIC$ e Log-verossimilhança ilustradas na Figura 7.10, os resultados são similares entre as três soluções, tendo uma diferença na *Deviance* na *PROC GENMOD* como já discutido anteriormente.

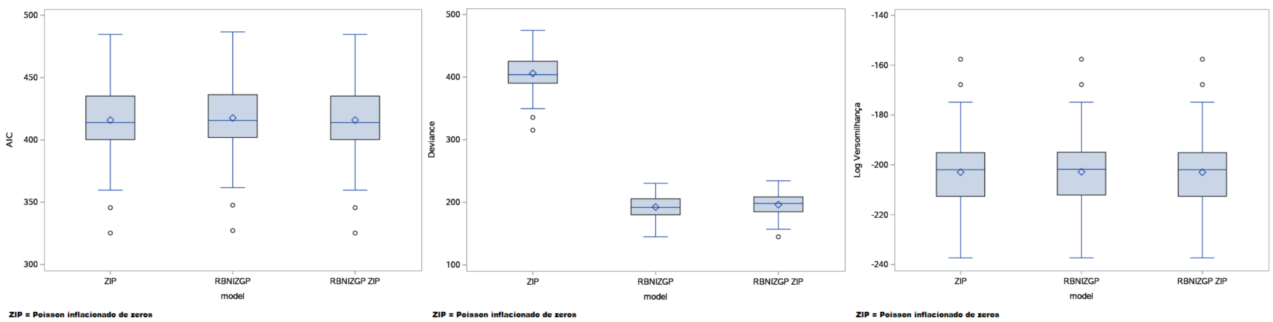
A Figura 7.11 mostra os *box-plot* dos parâmetros de suavização estimados pelos Algoritmos RBNIZGP (utilizando a binomial negativa inflacionada de zeros) e RBNIZGP (Poisson inflacionada de zeros) minimizando $AIC$ e $CV$. Nos resultados obtidos via algoritmo RBNIZGP (Poisson inflacionada de zeros), pode-se verificar que houve uma diferença entre os resultados




(a) (b) (c)

(d) (e)

Figura 7.9: *box-plot* - Resultados da modelagem com base na simulação dados Poisson inflacionado de zeros: (a) Intercepto, (b) $X_1$, (c) $X_2$, (d) Intercepto inflacionado e (e) $X_3$

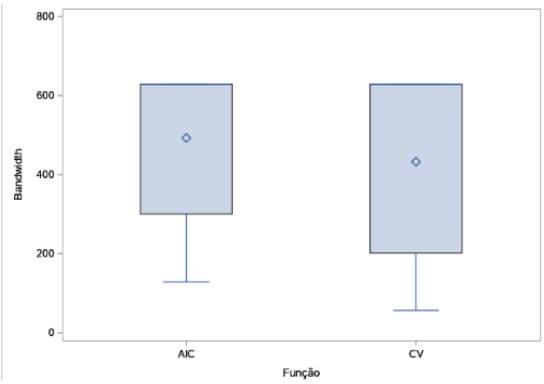



(a) (b) (c)

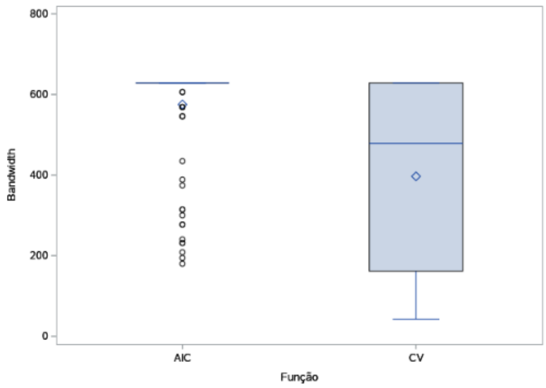
Figura 7.10: *box-plot* - Resultados da modelagem com base na simulação dados Poisson inflacionado de zeros (a) $AIC$, (b) *Deviance* e (c) Log-verossimilhança

da função $AIC$ e $CV$; no caso do $AIC$, é possível notar que a média do valor mínimo está próxima de 500 km, entretanto, no caso da $CV$, a média está próxima de 400 km e há uma variabilidade maior nos dados, comparando com o $AIC$.

Já no caso dos resultados obtidos pelo algoritmo RBNIZGP (binomial negativa inflacionada de zeros), no caso do $AIC$ é posível verificar que praticamente todos os dados se concentraram no valor 628.885 km, com exceção de pontos discrepantes existentes na distribuição, diferentemente da $CV$ que houve uma variação maior de valores, mostrando que $25\%$ dos dados se concentram em uma faixa inferior a 200 km e a sua média é próxima de 400 km.




(a)



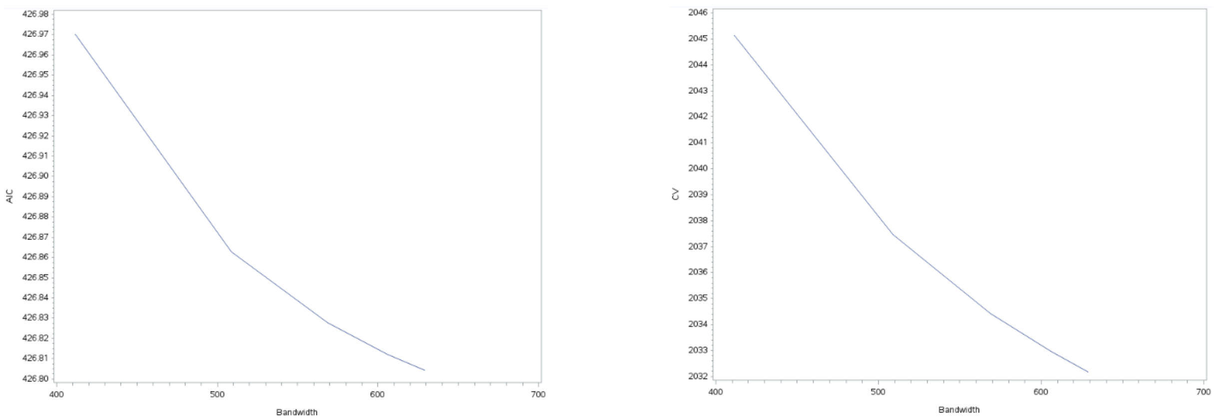

(b)

Figura 7.11: *box-plot* do parâmetro de suavização para as funções (a) $AIC$ e (b) $CV$ - Dados simulados Poisson inflacionada de zeros - RBNIZGP (Poisson inflacionada de zeros) e RBNIZGP (binomial negativa inflacionada de zeros)

### 7.2.4 Simulação com dados binomial negativo inflacionado de zeros

Para os dados simulados a partir de uma distribuição binomial negativa inflacionada de zeros, foram utilizados dois algoritmos: *PROC GENMOD* do SAS 9.4 e Algoritmo RBNIZGP (utilizando a binomial negativa inflacionada de zeros).

Como feito anteriormente através das outras simulações, o primeiro passo para utilização dos modelos geograficamente ponderados é a busca do parâmetro de suavização, que indicará se os dados apresentam dependência espacial ou não. Pode-se perceber na Figura 7.12 que o parâmetro de suavização tende à maior distância entre os dados, caracterizando dessa forma a não dependência espacial, tanto na função $AIC$ e $CV$ e esse parâmetro também não é maior devido a limitação da distância máxima entre os pontos.




(a) (b)

Figura 7.12: Esboço da Função (a) $AIC$ e (b) $CV$ - Dados simulados binomial negativa inflacionada de zeros - RBNIZGP (binomial negativa inflacionada de zeros)

Na Tabela 7.4 estão os resultados de uma das 100 bases simuladas, e verifica-se que as estimativas das variáveis preditoras foram exatamente iguais entre a *PROC GENMOD* e o Algoritmo RBNIZGP (binomial negativa inflacionada de zeros) sendo significativas a um nível de significância de $\alpha = 5\%$. Já as estimativas de erro padrão apresentaram uma diferença. O motivo dessas diferenças pode ser devido aos algoritmos utilizados bem como aos critérios de parada da otimização. Mas note que os valores estão próximos.

Os valores de $AIC$ e Log-verossimilhança também são os mesmos, mas a diferença existente na *Deviance* se deve ao mesmo motivo da Poisson inflacionada de zeros, em que na *PROC GENMOD*, tal medida é o dobro da Log-verossimilhança. Pode-se observar também que os resultados estão muito próximos do que foi definido nas simulações da Tabela 6.1, lembrando que no caso do Intercepto (inflacionado), assim como na Poisson inflacionada de zeros, o motivo para que esta estimativa não ter sido considerada significativa a um nível de significância de $\alpha = 5\%$ se deve ao fato de não ter sido definido na simulação.

A Figura 7.13 mostra os *box-plot* das estimativas das $n = 100$ variáveis preditoras de cada modelo gerado pelos dois algoritmos. Com os resultados, pode-se observar que em todos os

Tabela 7.4: Dados simulados - modelo binomial negativo inflacionado de zeros

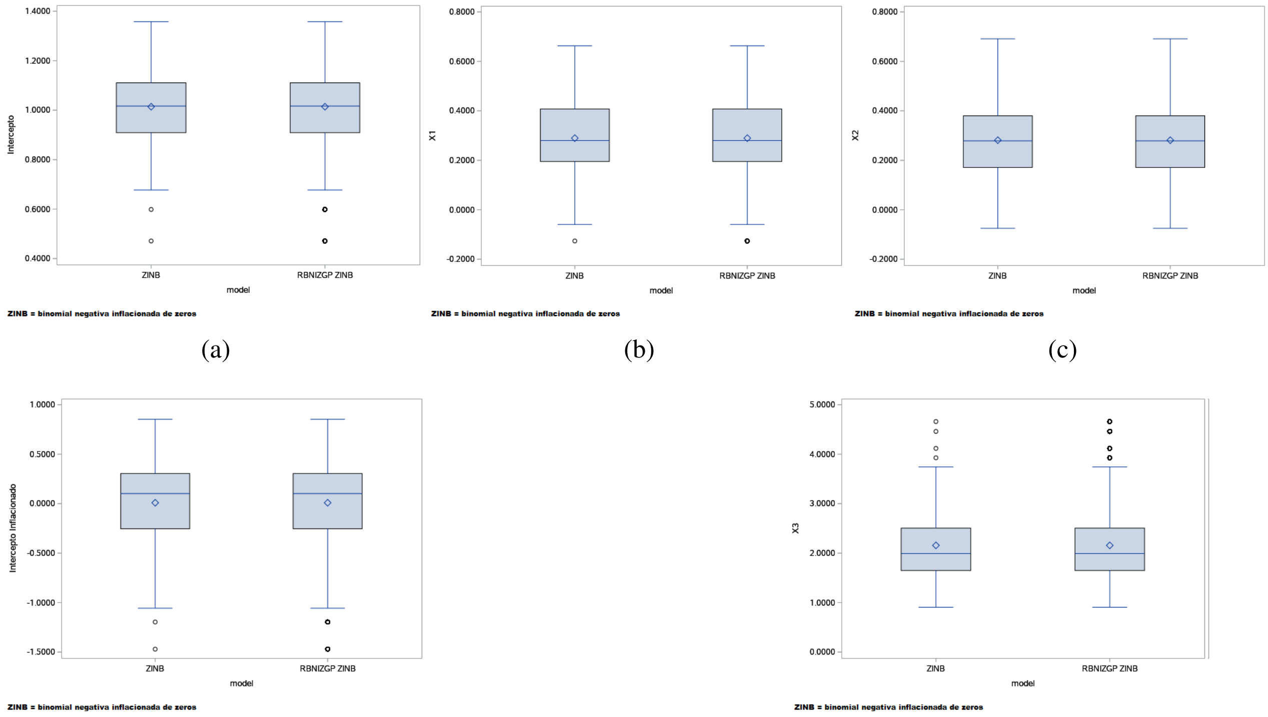
| **Variáveis/ Estatísticas** | **binomial negativa inflacionada de zeros** | | | |
|---|---|---|---|---|
| | ***PROC GENMOD*** | | ***RBNIZGP (binomial negativo inflacionado de zeros)*** | |
| | *Estimativa* | *Erro padrão* | *Estimativa* | *Erro padrão* |
| Intercepto | $1,1679^{*}$ | $0,1714$ | $1,1679^{*}$ | $0,1753$ |
| $X_1$ | $0,3430^{*}$ | $0,1425$ | $0,3430^{*}$ | $0,1427$ |
| $X_2$ | $0,3718^{*}$ | $0,6727$ | $0,3718^{*}$ | $0,1634$ |
| $\alpha$ | $1,1032^{*}$ | $0,3378$ | $1,1032^{*}$ | $0,3220$ |
| Intercepto (inflacionado) | $-0,3467^{\text{NS}}$ | $0,4711$ | $-0,3467^{\text{NS}}$ | $0,4355$ |
| $X_3$ (inflacionado) | $3,0298^{*}$ | $0,7411$ | $3,0298^{*}$ | $0,7027$ |
| Parâmetro de suavização | - | | $628,8850$ | |
| $AIC$ | $426,1669$ | | $426,1688$ | |
| $Deviance$ | $414,1689$ | | $142,2856$ | |
| Log-verossimilhança | $-207,0844$ | | $-207,0844$ | |

$(*)$ Significativo a 5%
(NS) Não Significativo a 5%

resultados, o comportamento das $n = 100$ estimativas, de todas as variáveis preditoras são muito similares. Nos Interceptos, a média e mediana estão muito próximas de 1.0, mostrando também a presença de *outliers* abaixo do limite inferior, já nas variáveis preditoras ($X_1$, $X_2$), as médias e medianas se concentram em valores próximos de 0,3. Já nos casos das variáveis preditoras inflacionadas de zeros (Intercepto e $X_3$) também é perceptível a presença de *outliers* e as médias e medianas estão próximas de 0,0 e 2,0 (Intercepto e $X_3$, respectivamente).

Nas estimativas do $AIC$ e Log-verossimilhança ilustradas na Figura 7.14, os resultados são similares entre os algoritmos. No entanto, na *Deviance* pode-se observar que através do resultado da *PROC GENMOD*, as estimativas estão em uma escalar superior à das outras soluções, pois esta estimativa é calculada como o dobro da Log-verossimilhança, tal fato pode ser visto também nos resultados do modelo gerado na Tabela 7.4, assim como na Poisson inflacionada de zeros.

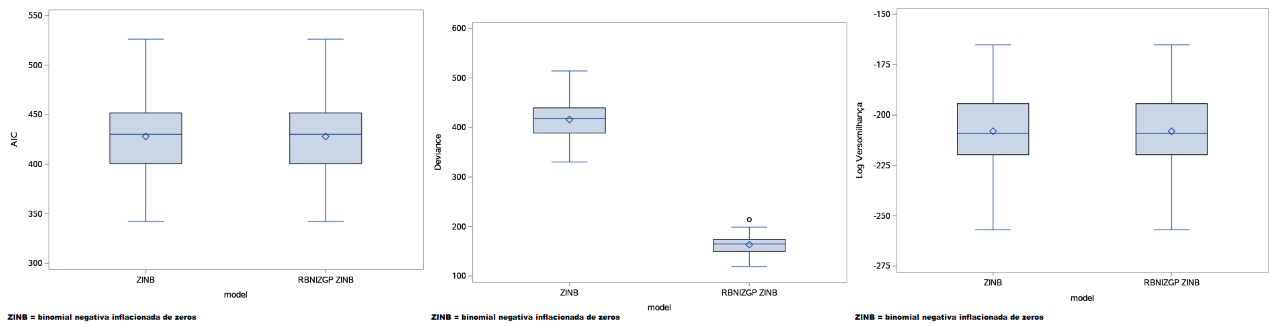
A Figura 7.15 mostra os *box-plot* dos parâmetros de suavização estimados pelo Algoritmo




(a) (b) (c)

(d) (e)

Figura 7.13: *box-plot* - Resultados da modelagem com base na simulação dados binomial negativa inflacionado de zeros: (a) Intercepto, (b) $X_1$, (c) $X_2$, (d) Intercepto inflacionado e (e) $X_3$




(a) (b) (c)

Figura 7.14: *box-plot* - Resultados da modelagem com base na simulação dados binomial negativo inflacionado de zeros (a) $AIC$, (b) *Deviance* e (c) Log-verossimilhança

RBNIZGP (utilizando a binomial negativa inflacionada de zeros) minimizando $AIC$ e $CV$. Pode-se verificar que houve uma grande diferença entre os resultados da função $AIC$ e $CV$. No caso da função $AIC$, é possível verificar a concentração de valores em torno de 200km

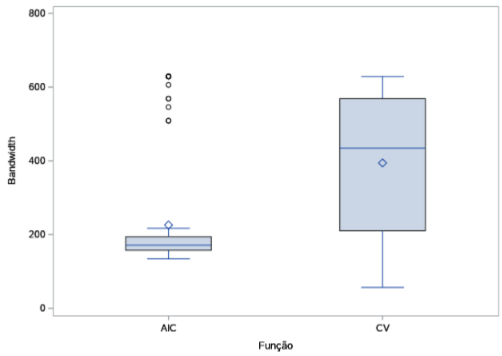



Figura 7.15: *box-plot* do parâmetro de suavização para as funções $AIC$ e $CV$ - Dados simulados binomial negativa inflacionada de zeros - RBNIZGP (binomial negativa inflacionada de zeros)

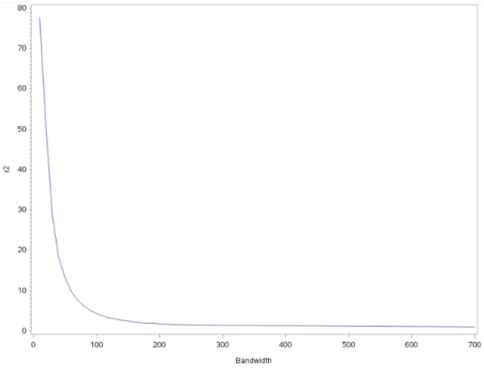
e devido a presença de pontos discrepantes acima do limite superior (próximos de 600km), a média se concentra em um valor acima de 200km. Já na função $CV$, a média se concentra em um valor próximo a 400km e pode-se verificar que existe uma maior variação nesse conjunto de dados.

### 7.2.5 Avaliação do parâmetro $r_2$

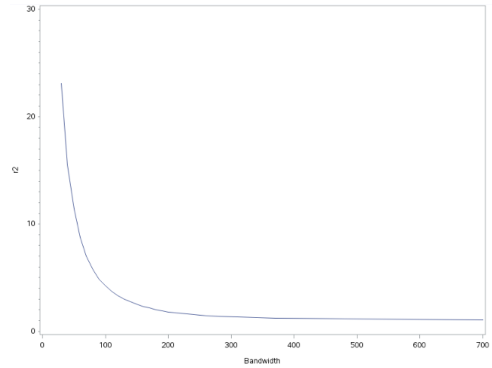
A Figura 7.16 mostra a ilustração gráfica da função $r_2$ em uma base simulada, que representa o número efetivo de parâmetros de $\alpha$, conforme definido no Capítulo 4. Em ambos os casos (binomial negativo e binomial negativo inflacionado de zeros), nota-se o rápido decaimento que existe na função, mostrando inclusive que as funções são estritamente decrescentes. Sendo assim, quanto menor o valor do parâmetro de suavização, maior será o valor $r_2$, seguindo a mesma ideia do número de parâmetros efetivos do modelo RGP. No caso da binomial negativa é possível verificar que foram necessários valores maiores do parâmetro $r_2$ e a partir de 200km (valor do parâmetro de suavização) a função se estabiliza.

### 7.2.6 Comparação entre os modelos

Antes de partir para a análise com os dados reais, a ideia que foi ilustrada na Figura 6.5 será comentada nesta Seção. A Tabela 7.5 apresenta algumas medidas de ajuste dos modelos




(a)




(b)

Figura 7.16: Esboço da função $r_2$ (número efetivos de parâmetros de $\alpha$) $\times$ parâmetro de suavização - (a) RBNIZGP (binomial negativa) e (b) RBNIZGP (binomial negativa inflacionada de zeros)

Poisson, binomial negativa, Poisson inflacionada de zeros e binomial negativa inflacionada de zeros em relação aos dados seguindo os mesmos modelos. Sendo assim, partindo da ideia da Figura 6.5, verifica-se que um modelo de probabilidade mais geral permite a estimação dos parâmetros de um modelo de probabilidade mais simples, por exemplo, a partir de um modelo binomial negativo, com parâmetro de superdispersão $(\alpha = 0)$, o que está sendo estimado é um modelo Poisson. Da mesma forma, a partir de um modelo Poisson inflacionado de zeros, quando os parâmetros inflacionados são todos iguais a zero ($\boldsymbol{\gamma} = 0$), o que está sendo estimado é novamente um modelo Poisson.

Na Tabela 7.5, é possível verificar que o modelo mais geral sempre se ajusta melhor aos dados. No caso dos dados Poisson, observa-se que os resultados da *Deviance* e a Log-verossimilhança foram praticamente os mesmos em todos os modelos, mostrando que a simplicidade de um dado Poisson. Já no conjunto de dados binomial negativo, o melhor ajuste foi da própria binomial negativa e na binomial negativo inflacionado de zeros.

Nos casos dos conjuntos de dados Poisson inflacionado de zeros e binomial negativo inflacionado de zeros, respectivamente, verifica-se que o ajuste do modelo Poisson foi considerado o pior em ambos os casos, e no caso de dados seguindo uma distribuição binomial negativa inflacionada de zeros, verifica-se que o modelo modelo binomial negativo inflacionado de zeros foi o com melhor ajuste, seguido do modelo binomial negativo.

Tabela 7.5: Comparação entre os modelos gerados pelo Algoritmo RBNIZGP, utilizando as medidas de ajuste

| **Dados** | **Algoritmo - RBNIZGP** | **Medidas de ajuste** | |
|---|---|---|---|
| | | **Log-verossimilhança** | ***Deviance*** |
| **Poisson** | *Poisson* | $-204,7465$ | $152,2006$ |
| | *binomial negativo* | $-204,7465$ | $152,2006$ |
| | *Poisson inflacionado de zeros* | $-202,8774$ | $148,4621$ |
| | *binomial negativo inflacionado de zeros* | $-202,8774$ | $148,4621$ |
| **binomial negativo** | *Poisson* | $-421,5317$ | $373,3839$ |
| | *binomial negativo* | $-383,2470$ | $171,3346$ |
| | *Poisson inflacionado de zeros* | $-415,5634$ | $361,4476$ |
| | *binomial negativo inflacionado de zeros* | $-383,2470$ | $171,3346$ |
| **Poisson inflacionado de zeros** | *Poisson* | $-307,8650$ | $403,9365$ |
| | *binomial negativo* | $-253,2164$ | $149,8475$ |
| | *Poisson inflacionado de zeros* | $-201,8980$ | $192,0025$ |
| | *binomial negativo inflacionado de zeros* | $-201,8430$ | $187,7819$ |
| **binomial negativo inflacionado de zeros** | *Poisson* | $-457.3507$ | $739,9017$ |
| | *binomial negativo* | $-240,1336$ | $118,0894$ |
| | *Poisson inflacionado de zeros* | $-262,7694$ | $350,7392$ |
| | *binomial negativo inflacionado de zeros* | $-207,0844$ | $142,2856$ |

## 7.3 Estudo de caso: dados da COVID-19 na Coréia do Sul

Como foi comentado na Seção 6.2.2, os casos de COVID-19 foram analisados por Weinstein et al. (2021) e a maior quantidade de casos esteve presente nas regiões de Seoul e Daegu. Além dessas regiões serem bastante populosas, elas também estão próximas de grandes aeroportos e portanto, isso facilitou a circulação do vírus, resultando em número expressivo de casos da doença (Wikipédia, 2021).

Para este estudo de caso, as variáveis utilizadas foram o número de casos de COVID-19 na fase inicial da pandemia como variável dependente, e as variáveis explicativas: *MORBIDITY* (Comorbidade), *HIGH_SCH_P* (Proporção de pessoas com 2° grau), *HEALTHCARE_ACCESS* (Acesso à saúde), *DIFF_SD* (Dificuldade de distanciamento social), *CROWDING* (Aglomeração), *MIGRATION* (Migração) e *HEALTH_BEHAVIOR* (Comportamento de saúde). A maioria desssas variáveis foi criada a partir da Análise Fatorial, e mais detalhes podem ser vistos em Weinstein et al. (2021).

Portanto, o modelo global da parte não inflacionada de zeros é definido como

$$\text{Casos de COVID-19} = \exp(\beta_0 + \beta_1 \textit{MORBIDITY} + ... + \beta_7 \textit{HEALTH\_BEHAVIOR}) \quad (7.1)$$

onde $\beta_0, \beta_1, ..., \beta_7$ representam os coeficientes da regressão.

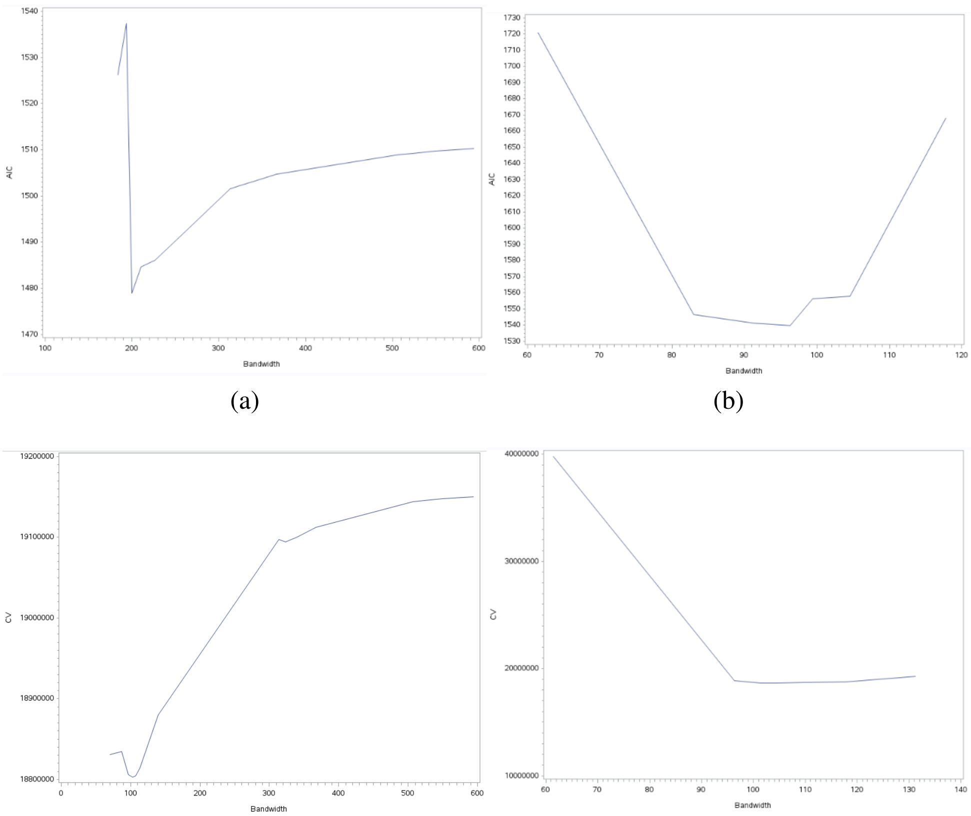
Já o modelo global da parte inflacionada de zeros é definido como

$$\text{Prob de COVID-19} = \frac{\exp(\gamma_0 + \gamma_1 \textit{MORBIDITY} + ... + \gamma_7 \textit{HEALTH\_BEHAVIOR})}{1 + \exp(\gamma_0 + \gamma_1 \textit{MORBIDITY} + ... + \gamma_7 \textit{HEALTH\_BEHAVIOR})} \quad (7.2)$$

onde $\gamma_0, \gamma_1, ..., \gamma_7$ representam os coeficientes da regressão.

Baseando-se no método proposto, o modelo RBNIZGP foi ajustado aos dados, buscando o melhor parâmetro de suavização pela minimização do $AIC$ ou $CV$ (usando o parâmetro de suavização fixo) e $AIC$ ou $CV$ (usando o parâmetro de suavização adaptável). Sendo assim, ao encontrar o parâmetro de suavização para cada uma dessas funções, foi possível verificar a diferença de comportamento existente em cada caso. Já a distância máxima calculada entre os centróides das regiões foi de aproximadamente 690 km. Na Figura 7.17 estão as curvas das funções $AIC$ (Fixo e Adaptável) e $CV$ (Fixo e Adaptável) encontradas.

Na função $AIC$ (Fixo), nota-se que o mímino global do parâmetro de suavização é igual a 199,96 km, gerando um valor de $AIC$ próximo de 1480. Já na função CV (Fixo), nota-se que o mímino global do parâmetro de suavização é igual a 103,14 km, gerando um valor de $CV$ próximo de 18,800,000. Portanto, com base nas curvas da Figura 7.17 pode-se concluir que os dados possuem dependência espacial, visto que os mínímos encontrados estão longe da máxima distância possível entre os pontos.




(a) (b) (c) (d)

Figura 7.17: Esboço da função (a) $AIC$ (Fixo), (b) $AIC$ (Adaptável), (c) $CV$ (Fixo) e (d) $CV$ (Adaptável) dos dados de COVID-19 na Coréia do Sul (modelo binomial negativo inflacionado de zeros)

As medidas de ajuste do modelo são essenciais na escolha de qual o melhor modelo a ser utilizado. Portanto, na Tabela 7.6 constam os resultados dessas medidas e o valor mínimo do parâmetro de suavização para cada função. Pode-se observar que o menor $AIC$ provém do modelo em que foi utilizada a função $AIC$ (adaptável), tendo $AIC = 1461,2036$. Os pseudos coeficientes de determinação ($R^2$ e $R^2$ ajustado) tiveram um resultado simular entre as funções $AIC$ e $CV$ (Adaptável) sendo (0,9943 e 0,9936, respectivamente), diferente das funções $AIC$ e $CV$ (Fixo) sendo (0,5333 e 0,9899, respectivamente).

Já na função *Deviance*, é perceptível que o melhor resultado foi na função $AIC$ (Fixo), sendo *Deviance* $= 227,9770$ e o maior valor desta função foi na função $CV$ (Fixo), sendo

*Deviance* $= 527,3921$. Portanto, conclui-se que o melhor modelo foi o que resultou por meio da função $AIC$ (Adaptável), obtendo a maior estimativa na Log-verossimilhança, a menor estimativa no $AIC$ e os maiores $R^2$ e $R^2$ ajustado.

Tabela 7.6: Medidas de ajuste do modelo local binomial negativo inflacionado de zeros utilizando o algoritmo da RBNIZGP, segundo as funções: $AIC$ e $CV$ (Fixo) e $AIC$ e $CV$ (Adaptável)

| **RBNIZGP** | | | | |
|---|---|---|---|---|
| **Medidas de Ajuste** | **Função** $AIC$ | | **Função** $CV$ | |
| | **Fixo** | **Adaptável** | **Fixo** | **Adaptável** |
| Log-verossimilhança | $-712,3219$ | $-650,9548$ | $-819,4793$ | $-672,9266$ |
| *Deviance* | $227,9770$ | $297,2286$ | $527,3921$ | $336,2705$ |
| $AIC$ | $1473,3000$ | $1461,2036$ | $1721,6389$ | $1493,0188$ |
| $AICc$ | $1478,9357$ | $1539,8468$ | $1738,9984$ | $1557,8058$ |
| $R^2$ | $0,5333$ | $0,9943$ | $0,9899$ | $0,9936$ |
| $R^2$ ajustado | $0,4316$ | $0,9912$ | $0,9883$ | $0,9908$ |
| Parâmetro de suavização | $199,96$ | $96$ | $103,14$ | $104$ |
| Nº de parâmetros estimados | $24,3281$ | $79,6470$ | $41,3402$ | $73,5828$ |

Como foi comentado no Capítulo 6, o principal objetivo deste estudo de caso consiste em verificar e comparar as diferenças entre os parâmetros estimados e na qualidade de ajuste no modelo binomial negativo (já analisado por Weinstein et al. (2021)), com o modelo binomial negativo inflacionado de zeros. Portanto, o mesmo passo para encontrar o melhor parâmetro de suavização foi feito utilizando a minimização do $AIC$ ou $CV$ (usando o parâmetro de suavização fixo) e $AIC$ ou $CV$ (usando o parâmetro de suavização adaptável). Na Figura 7.18 estão as curvas das funções $AIC$ (Fixo e Adaptável) e $CV$ (Fixo e Adaptável) encontradas.

As medidas de ajuste do modelo binomial negativo estão ilustradas na Tabela 7.7. Pode-se verificar que o menor $AIC$ provém do modelo em que foi utilizada a função $AIC$ (adaptável), tendo $AIC = 1393,4701$. Como no modelo binomial negativo inflacionado de zeros, os coeficientes de determinação ($R^2$ e $R^2$ ajustado) tiveram um resultado similar entre as funções $AIC$ e $CV$ (Fixo) tendo valores 0,4625 e 0,4304, respectivamente. No entanto, na função $AIC$ adptável este valor foi o maior, sendo 0,9344. Pode-se concluir que o melhor modelo

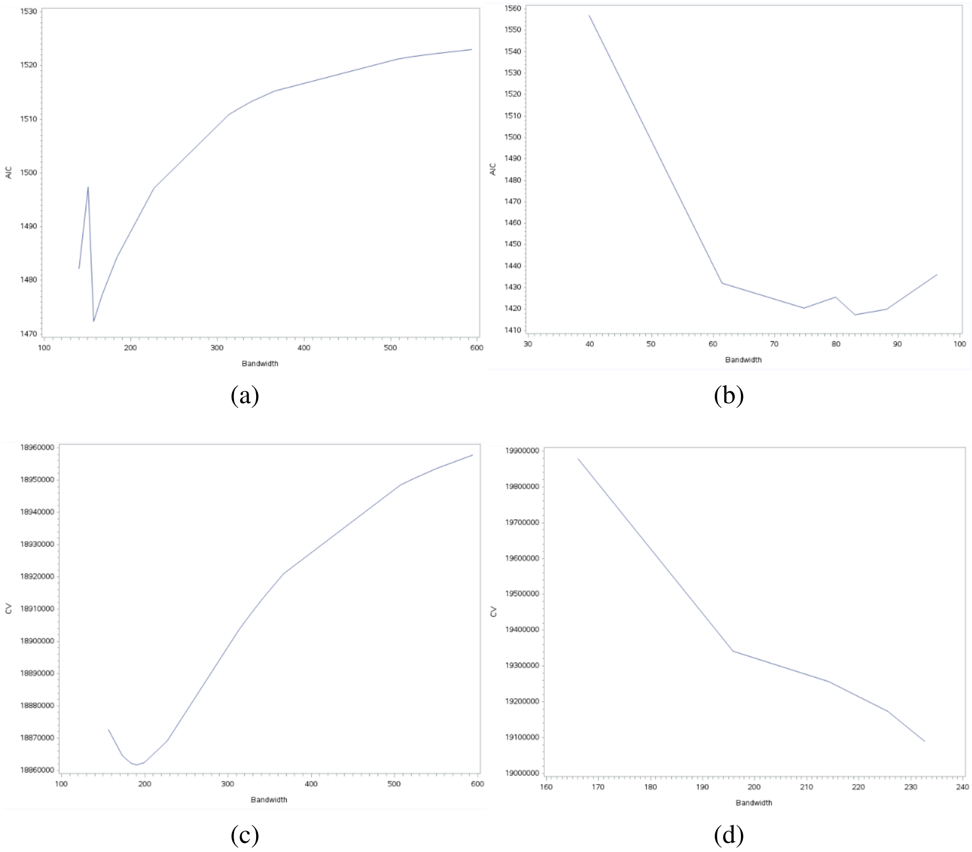



Figura 7.18: Esboço da função (a) $AIC$ (Fixo), (b) $AIC$ (Adaptável), (c) $CV$ (Fixo) e (d) $CV$ (Adaptável) dos dados de COVID-19 na Coréia do Sul (modelo binomial negativo)

foi o que resultou por meio da função $AIC$ (Adaptável), obtendo a maior estimativa na Log-verossimilhança, a menor estimativa no $AIC$ e os maiores valores de $R^2$ e $R^2$ ajustado.

Comparando os resultados entre os dois modelos, verifica-se que as medidas de ajuste geradas nas funções $AIC$ e $CV$ (fixo) foram sempre bem próximas. A Tabela 7.8 apresenta as estimativas obtidas pelo modelo global binomial negativo inflacionado de zeros e pelo modelo binomial negativo. Pode-se perceber que as estimativas de alguns dos parâmetros são similares, no entanto, devido a parte inflacionada de zeros, tais estimativas não ficam totalmente iguais. Todas as estimativas (não inflacionadas) foram consideradas significativas a um nível de $5\%$ de significância, com exceção da variável CROWDING no modelo binomial negativo, que foi considerada significativa a um nível de $12\%$ de significância.

Tabela 7.7: Medidas de ajuste do modelo local binomial negativo utilizando o algoritmo da RBNIZGP, segundo as funções: $AIC$ e $CV$ (Fixo) e $AIC$ e $CV$ (Adaptável)

| **RBNIZGP** | | | | |
|---|---|---|---|---|
| **Medidas de Ajuste** | **Função $AIC$** | | **Função $CV$** | |
| | **Fixo** | **Adaptável** | **Fixo** | **Adaptável** |
| Log-verossimilhança | $-720,8734$ | $-649,0485$ | $-729,7504$ | $-715,9926$ |
| *Deviance* | $224,9974$ | $277,0861$ | $232,3174$ | $222,4450$ |
| $AIC$ | $1470,4079$ | $1393,4701$ | $1484.9994$ | $1463,8808$ |
| $AICc$ | $1472,3294$ | $1417,2444$ | $1486.5221$ | $1466,2615$ |
| $R^2$ | $0,4625$ | $0,9344$ | $0,4304$ | $0,4768$ |
| $R^2$ ajustado | $0,3916$ | $0,9117$ | $0,3654$ | $0,3994$ |
| Parâmetro de suavização | $156,66$ | $82$ | $189,74$ | $232$ |
| Nº de parâmetros estimados | $14,3305$ | $47,6866$ | $12,7493$ | $15,9477$ |

Tabela 7.8: Estimativas globais dos modelos binomial negativo inflacionado de zeros e binomial negativo

| | **Variáveis / Estatísticas** | **binomial negativo inflacionado de zeros** | | | **binomial negativo** | | |
|---|---|---|---|---|---|---|---|
| | | **Estimativas** | **Erro padrão** | **P-valor** | **Estimativas** | **Erro padrão** | **P-valor** |
| NÃO INFLACIONADO | Intercepto | $-15,0587^{*}$ | $1,8461$ | $< 0,0001$ | $-12,1913^{*}$ | $1,8745$ | $< 0,0001$ |
| | MORBIDITY | $0,0466^{*}$ | $0,0082$ | $< 0,0001$ | $0,0469^{*}$ | $0,0080$ | $< 0,0001$ |
| | HIGH_SCH_P | $-0,1120^{*}$ | $0,0385$ | $0,0026$ | $-0,1100^{*}$ | $0,0387$ | $0,0049$ |
| | HEALTHCARE_ACCESS | $-0,1437^{*}$ | $0,0328$ | $< 0,0001$ | $-0,1374^{*}$ | $0,0308$ | $< 0,0001$ |
| | DIFF_SD | $0,0671^{*}$ | $0,2952$ | $0,0239$ | $0,0717^{*}$ | $0,0310$ | $0,0218$ |
| | CROWDING | $0,4326^{*}$ | $0,1261$ | $0,0007$ | $0,2220^{**}$ | $0,1164$ | $0,0578$ |
| | MIGRATION | $-0,3664^{*}$ | $0,0855$ | $< 0,0001$ | $-0,3770^{*}$ | $0,0904$ | $< 0,0001$ |
| | HEALTH_BEHAVIOR | $0,0520^{*}$ | $0,0192$ | $0,0073$ | $0,0425^{*}$ | $0,0193$ | $0,0288$ |
| | $\alpha$ | $3,1492^{*}$ | $0,1477$ | $< 0,0001$ | $3,6817^{*}$ | $0,3555$ | $< 0,0001$ |
| INFLACIONADO | Intercepto | $-18,5334^{**}$ | $9,4789$ | $0,0517$ | - | - | - |
| | MORBIDITY | $-0,0109^{NS}$ | $0,0312$ | $0,7251$ | - | - | - |
| | HIGH_SCH_P | $-0,0346^{NS}$ | $0,2067$ | $0,8673$ | - | - | - |
| | HEALTHCARE_ACCESS | $0,1365^{**}$ | $0,0853$ | $0,1110$ | - | - | - |
| | DIFF_SD | $-0,1401^{NS}$ | $0,1550$ | $0,3671$ | - | - | - |
| | CROWDING | $0,6198^{*}$ | $0,3044$ | $0,0428$ | - | - | - |
| | MIGRATION | $0,1181^{NS}$ | $0,4040$ | $0,7702$ | - | - | - |
| | HEALTH_BEHAVIOR | $0,1626^{NS}$ | $0,1620$ | $0,3163$ | - | - | - |
| GLOBAL | $AIC$ | $1528,9484$ | | | $1527,1617$ | | |
| | $Deviance$ | $282,67674$ | | | $263,5412$ | | |
| | Log-verossimilhança | $-747,4742$ | | | $-754,5808$ | | |

$(*)$ Significativo a 5%
$(**)$ Significativo a 12%
(NS) Não Significativo a 12%

No modelo binomial negativo inflacionado de zeros, para fins de comparação entre as medidas de ajuste local (Tabela 7.6) e global (Tabela 7.8), verifica-se que a Log-verossimilhança e o AIC ficaram melhores no ajuste local, no entanto, a *Deviance* ficou melhor no ajuste global. Já no modelo binomial negativo, percebe-se o mesmo comportamento da binomial negativo inflacionado de zeros, as medidas de ajuste no modelo local (Tabela 7.7): Log-verossimilhança e o AIC ficaram melhores, já a *Deviance* ficou melhor no ajuste global.

Tabela 7.9: Sumário das estimativas dos parâmetros utilizando Algoritmo RBNIZGP (binomial negativo inflacionado de zeros) - Resultados da função $AIC$ (Adaptável)

| RBNIZGP (binomial negativa inflacionada de zeros) | | | | | | | |
|---|---|---|---|---|---|---|---|
| | **Variáveis** | ***Estimativas*** | | | | | |
| | | **Mínimo** | **Q1** | **Média** | **Mediana** | **Q3** | **Máximo** |
| NÃO INFLACIONADO | Intercepto | $-37,6100$ | $-12,9812$ | $-11,8597$ | $-12,2736$ | $-9,8222$ | $6,0197$ |
| | MORBIDITY | $-0,0058$ | $0,0183$ | $0,0407$ | $0,03335$ | $0,0462$ | $0,2738$ |
| | HIGH_SCH_P | $-0,3722$ | $-0,1467$ | $-0,1120$ | $-0,0994$ | $-0,0764$ | $0,1901$ |
| | HEALTHCARE_ACCESS | $-0,4118$ | $-0,1409$ | $-0,1032$ | $-0,1185$ | $-0,0555$ | $0,1373$ |
| | DIFF_SD | $-0,1875$ | $-0,0042$ | $0,0507$ | $0,0022$ | $0,1479$ | $0,3533$ |
| | CROWDING | $-0,7188$ | $0,0367$ | $0,1674$ | $0,2331$ | $0,3270$ | $0,7441$ |
| | MIGRATION | $-1,9930$ | $-0,2708$ | $-0,1217$ | $-0,0202$ | $0,0427$ | $0,3684$ |
| | HEALTH_BEHAVIOR | $-0.1308$ | $0,0189$ | $0,0289$ | $0,0267$ | $0,0422$ | $0,1773$ |
| | $\alpha$ | $0,0000$ | $0,0000$ | $1,2653$ | $0,5699$ | $2,7078$ | $3,8126$ |
| INFLACIONADO | Intercepto | $-83271,3400$ | $-2273,6020$ | $-3285,2150$ | $-1341,102$ | $-21,1250$ | $0,0000$ |
| | MORBIDITY | $-26,0522$ | $0,0000$ | $9,9208$ | $0,2363$ | $2,0558$ | $335,9579$ |
| | HIGH_SCH_P | $-158,8096$ | $-0,1471$ | $26,8064$ | $6,8805$ | $26,4502$ | $621,0657$ |
| | HEALTHCARE_ACCESS | $-47,20052$ | $0,0000$ | $13,7097$ | $0,1261$ | $13,4269$ | $334,4453$ |
| | DIFF_SD | $-145,4886$ | $-0,2221$ | $19,7863$ | $0,0484$ | $14,0479$ | $729,3378$ |
| | CROWDING | $-271,4220$ | $0,7237$ | $43,0420$ | $9,5809$ | $46,5500$ | $1045,813$ |
| | MIGRATION | $-2600,5440$ | $-25,2981$ | $-66,4887$ | $-1,4696$ | $0,3592$ | $132,5150$ |
| | HEALTH_BEHAVIOR | $-83,9257$ | $-4,1299$ | $2,2334$ | $0,0315$ | $0,2597$ | $203,4262$ |

Por meio dos resultados que foram ajustados no modelo global da binomial negativa inflacionada de zeros, Tabela 7.8, pode-se perceber que as estimativas das variáveis preditoras em geral estão próximas das respectivas médias/medianas dessas variáveis nas estimativas dos modelos locais na Tabela 7.9. Na parte inflacionada de zeros, as estimativas do modelo global ficaram ainda mais distantes das médias e medianas das estimativas locais.

Para fins de verificação do melhor ajuste do modelo, no caso do modelo inflacionado de zeros, o algoritmo foi executado novamente, mas desta vez considerando apenas as variáveis preditoras inflacionadas significativas. Na busca do melhor parâmetro de suavização pela minimização do $AIC$ ou $CV$ (usando o parâmetro de suavização fixo) e $AIC$ ou $CV$ (usando o parâmetro de suavização adaptável), conclui-se que o modelo gerado por meio da função $AIC$ adaptável gerou um melhor ajuste, como pode-se verificar na Tabela 7.10. Ainda é possível verificar através destes resultados que as medidas de ajuste $AIC$ e Log-Verossimilhança ficaram muito próximas e melhores do que o modelo binomial negativo. Um ponto interessante é que agora ambos os modelos possuem o mesmo parâmetro de suavização igual a 82.

Tabela 7.10: Estimativas globais do modelos reduzido binomial negativo inflacionado de zeros e do modelo binomial negativo

| | **Variáveis / Estatísticas** | **binomial negativo inflacionado de zeros** | | | **binomial negativo** | | |
|---|---|---|---|---|---|---|---|
| | | **Estimativas** | **Erro padrão** | **P-valor** | **Estimativas** | **Erro padrão** | **P-valor** |
| NÃO INFLACIONADO | Intercepto | $-15,1281^*$ | $1,8349$ | $< 0,0001$ | $-12,1913^*$ | $1,8745$ | $< 0,0001$ |
| | MORBIDITY | $0,0470^*$ | $0,0081$ | $< 0,0001$ | $0,0469^*$ | $0,0080$ | $< 0,0001$ |
| | HIGH_SCH_P | $-0,1107^*$ | $0,0359$ | $0,0023$ | $-0,1100^*$ | $0,0387$ | $0,0049$ |
| | HEALTHCARE_ACCESS | $-0,1446^*$ | $0,0326$ | $< 0,0001$ | $-0,1374^*$ | $0,0308$ | $< 0,0001$ |
| | DIFF_SD | $0,0680^*$ | $0,0293$ | $0,0211$ | $0,0717^*$ | $0,0310$ | $0,0218$ |
| | CROWDING | $0,4354^*$ | $0,1250$ | $0,0006$ | $0,2220^{**}$ | $0,1164$ | $0,0578$ |
| | MIGRATION | $-0,3671^*$ | $0,0849$ | $< 0,0001$ | $-0,3770^*$ | $0,0904$ | $< 0,0001$ |
| | HEALTH_BEHAVIOR | $0,0513^*$ | $0,0192$ | $0,0082$ | $0,0425^*$ | $0,0193$ | $0,0288$ |
| | $\alpha$ | $3,1147^*$ | $0,2049$ | $< 0,0001$ | $3,6817^*$ | $0,3555$ | $< 0,0001$ |
| INFLAC | Intercepto | $-13,8643^*$ | $3,6194$ | $0,0002$ | - | - | - |
| | HEALTHCARE_ACCESS | $0,1307^*$ | $0,0602$ | $0,0311$ | - | - | - |
| | CROWDING | $0,5694^*$ | $0,2228$ | $0,0112$ | - | - | - |
| GLOBAL | $AIC$ | $1520,5892$ | | | $1527,1617$ | | |
| | $Deviance$ | $286,6230$ | | | $263,5412$ | | |
| | Log-verossimilhança | $-748,2946$ | | | $-754,5808$ | | |
| LOCAL | Parâmetro de suavização | 82 | | | 82 | | |
| | $AIC$ | $1417,8852$ | | | $1393,4701$ | | |
| | $Deviance$ | $285,8302$ | | | $277,0861$ | | |
| | Log-verossimilhança | $-641,7630$ | | | $-649,0485$ | | |

$(*)$ Significativo a 5%
$(**)$ Significativo a 12%
(NS) Não Significativo a 12%

Partindo do modelo escolhido (função $AIC$ adaptável), o próximo passo é mostrar a ideia que foi descrita na Figura 6.5. Ao executar o algoritmo da RBNIZGP (binomial negativa inflacionada de zeros), foi possível identificar as localidades em que o parâmetro $\gamma = 0$ (Figura 7.19b), e que o parâmetro de superdispersão $\alpha$ foi considerado significativo (Figura 7.19a). Para fins de ilustração, pode-se observar na Figura 7.19c que as observações (ID=124 e ID=125) marcadas na cor verde são os casos em que o parâmetro $\gamma = 0$ e que o parâmetro de superdispersão $\alpha$ foi considerado significativo. Portanto, ao observar esses resultados, supõe-se que nessas linhas a distribuição seja uma binomial negativa. Os erros padrão permanecem inalterados nessas linhas específicas.

Para verificar isso, o algoritmo foi executado novamente, mas desta vez forçando o algoritmo RBNIZGP a modelar uma distribuição binomial negativa (ou seja, fazendo $\gamma = 0$). Desta vez, as mesmas linhas foram destacadas e é possível perceber que as estimativas na Figura 7.19e são as mesmas da Figura 7.19c, confirmando a ideia descrita na Figura 6.5. Ainda é possível verificar na Figura 7.19c as outras linhas marcadas em vermelho, que são os casos em que o parâmetro $\gamma \neq 0$, confirmando que nessas localidades os dados não seguem uma distribuição binomial negativa, e sim uma distribuição binomial negativa inflacionada de zeros. Ao comparar as estimativas dessas linhas com as estimativas das linhas na Figura 7.19e, percebe-se que tais valores são diferentes.

Seguindo o mesmo método para o modelo Poisson inflacionado de zeros, ao executar o algoritmo da RBNIZGP (binomial negativa inflacionada de zeros), desta vez o número de variáveis foi reduzido (modelo da Tabela 7.10), considerando as variáveis inflacionadas: Intercepto, *CROWDING* e *HEALTHCARE_ACCESS* (sendo as variáveis preditoras que foram consideradas significativas a um nível de significância de $12\%$), foram observadas as linhas em que $\gamma \neq 0$ e $\alpha = 0$.

Na Figura 7.20a, observa-se que a observação (ID=72) possui parâmetro de superdispersão $\alpha = 0$, e como os parâmetros inflacionados da Figura 7.20c para o mesmo ID=72 não são considerados significativos, pode-se concluir que a distribuição seja Poisson. Dessa forma, ao executar o algoritmo RBNIZGP novamente, mas forçando uma Poisson (ou seja, fazendo $\gamma = 0$ e $\alpha = 0$), é possível perceber que as estimativas do ID=72 (Figura 7.20d) ficaram próximas das estimativas dos parâmetros da Figura 7.20b. Ainda é possível verificar que a estimativa do intercepto inflacionado em 7.20c traz influencias nas estimativas da parte não inflacionada, geradas em 7.20d. Naturalmente, os valores dos erros padrão nas linhas tiveram uma redução ao forçar a distribuição Poisson.

| | id | alpha | std | tstat | probt | sig_alpha |
|---|---|---|---|---|---|---|
| 122 | 122 | 2.4878024623 | 0.6534816099 | 3.8069968988 | 0.0001406645 | significant at 99% |
| 123 | 123 | 2.8591072222 | 0.5048454152 | 5.6633320536 | 1.4846147E-8 | significant at 99% |
| 124 | 124 | 2.6266327874 | 0.5716356643 | 4.5949421135 | 4.3286942E-6 | significant at 99% |
| 125 | 125 | 2.2302076949 | 0.5145285268 | 4.3344685059 | 0.0000146113 | significant at 99% |
| 126 | 126 | 2.8275427807 | 0.7208751669 | 3.9223750668 | 0.0000876804 | significant at 99% |

(a)

| id | Inf_Intercept | Inf_MORBIDITY | Inf_HIGH_SCH_P | Inf_HEALTHCARE_ACCESS | Inf_DIFF_SD | Inf_CROWDING | Inf_MIGRATION | Inf_HEALTH_BEHAVIOR |
|---|---|---|---|---|---|---|---|---|
| 122 | -3563.490235 | 8.9153733503 | 20.86579459 | 21.049853123 | -80.64284045 | -228.7730362 | 34.533728566 | 113.89138293 |
| 123 | -2297.935321 | 3.4327699924 | 19.582654834 | 21.647468352 | 37.121627538 | 21.798915361 | 1.3334991052 | -4.220162901 |
| 124 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 125 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 126 | -149.9124939 | -0.464361599 | -5.371104462 | 0.4446627234 | -2.434364006 | 10.02283689 | 4.1274079926 | 4.1768398938 |

(b)

| id | Intercept | MORBIDITY | HIGH_SCH_P | HEALTHCARE_ACCESS | DIFF_SD | CROWDING | MIGRATION | HEALTH_BEHAVIOR |
|---|---|---|---|---|---|---|---|---|
| 122 | 3.1265492264 | 0.0474082795 | -0.233102154 | -0.153036851 | -0.006798446 | -0.543969611 | 0.2159996598 | 0.0632224597 |
| 123 | -12.40793006 | 0.0175903244 | -0.085469759 | -0.054893425 | 0.1975375501 | 0.236674619 | -0.303749382 | 0.0307693993 |
| 124 | -2.918449626 | 0.0357278328 | -0.252136461 | -0.124315238 | 0.1252175061 | -0.118960736 | 0.0109765555 | 0.0527556475 |
| 125 | -11.41414119 | 0.0377847669 | -0.083144404 | -0.098979954 | 0.1728174921 | 0.0831493621 | -0.210875602 | 0.0445280179 |
| 126 | -13.59496763 | 0.0522098782 | -0.082649244 | -0.132127868 | 0.1873629111 | 0.2013880942 | -0.161823743 | 0.0275207553 |

(c)

| | id | alpha | std | tstat | probt | sig_alpha |
|---|---|---|---|---|---|---|
| 122 | 122 | 2.8655302505 | 0.7576775876 | 3.7819915717 | 0.0001555786 | significant at 99% |
| 123 | 123 | 3.1597795563 | 0.5556082672 | 5.6870636076 | 1.2924233E-8 | significant at 99% |
| 124 | 124 | 2.6266327874 | 0.5716356643 | 4.5949421134 | 4.3286942E-6 | significant at 99% |
| 125 | 125 | 2.2302076947 | 0.5145285268 | 4.3344685058 | 0.0000146113 | significant at 99% |
| 126 | 126 | 2.9797819394 | 0.7542807048 | 3.9504947169 | 0.0000779898 | significant at 99% |

(d)

| id ▲ | Intercept | MORBIDITY | HIGH_SCH_P | HEALTHCARE_ACCE... | DIFF_SD | CROWDING | MIGRATION | HEALTH_BEHAVIOR |
|---|---|---|---|---|---|---|---|---|
| 122 | 2.8091363351 | 0.0507575327 | -0.306041626 | -0.164827798 | 0.0497268214 | -0.422132128 | 0.2505709697 | 0.0579726274 |
| 123 | -11.62858468 | 0.0229511612 | -0.099166064 | -0.077310313 | 0.1621788702 | 0.2209234083 | -0.356907318 | 0.0381735187 |
| 124 | -2.918449626 | 0.0357278328 | -0.252136461 | -0.124315238 | 0.1252175061 | -0.118960736 | 0.0109765555 | 0.0527556475 |
| 125 | -11.41414119 | 0.0377847669 | -0.083144404 | -0.098979954 | 0.1728174921 | 0.0831493621 | -0.210875602 | 0.0445280179 |
| 126 | -12.94008484 | 0.05593706 | -0.077432515 | -0.133248399 | 0.1845140645 | 0.1456578553 | -0.166578623 | 0.0181466108 |

(e)

Figura 7.19: Visualização de algumas observações da base de dados com estimativas dos parâmetros: (a) $\alpha$ da RBNIZGP; (b) Parte inflacionada ($\gamma$) da RBNIZGP; (c) parte não-inflacionada ($\beta$) da RBNIZGP; (d) $\alpha$ da RBNIZGP (binomial negativa); (e) RBNIZGP (binomial negativa)

A próxima análise será feita utilizando o risco relativo (ou do inglês, *Relative risk* - RR) para a parte não inflacionada e a razão de chances (ou do inglês, *Odds Ratio* - OR) para a parte inflacionada. Neste caso, o RR traz como informação do quanto que tal variável pode trazer de acréscimo ou decréscimo no número de casos de COVID-19, entretanto, a OR traz como informação do risco em que determinado fator pode trazer em ter ou não ter COVID-19 (Agresti, 2003).

| | id | alpha | std | tstat | probt | sig_alpha |
|---|---|---|---|---|---|---|
| 70 | 70 | 3.0675138514 | 0.9442113128 | 3.2487577831 | 0.0011591013 | significant at 99% |
| 71 | 71 | 1.0534500699 | 0.5310790791 | 1.9836030292 | 0.0473001113 | not significant at 90% |
| 72 | 72 | 1E-6 | 0.0000617909 | 0.0161836104 | 0.9870879108 | not significant at 90% |
| 73 | 73 | 1E-6 | 0.0000191791 | 0.0521400205 | 0.9584171246 | not significant at 90% |
| 74 | 74 | 1E-6 | 5.5217565E-6 | 0.1811017926 | 0.8562876775 | not significant at 90% |

(a)

| id ▲ | Intercept | MORBIDITY | HIGH_SCH_P | HEALTHCARE_ACCESS | DIFF_SD | CROWDING | MIGRATION | HEALTH_BEHAVIOR |
|---|---|---|---|---|---|---|---|---|
| 70 | 5.1499163148 | 0.0577461037 | -0.273803699 | -0.162351747 | -0.043091871 | -0.634831994 | 0.32596916 | 0.0552523543 |
| 71 | -5.835281415 | 0.0185566299 | -0.09075913 | 0.0122942596 | -0.047575745 | -0.169408645 | 0.1041208061 | -0.027015591 |
| 72 | -8.889064215 | 0.0211394099 | -0.037520285 | -0.190364749 | 0.0063837835 | -0.077927263 | 0.1640798203 | 0.0346990252 |
| 73 | -4.88149288 | 0.0971355198 | -0.251720945 | 0.0106151918 | -0.14059406 | -0.303282961 | -0.038702895 | -0.038190051 |
| 74 | -9.200926312 | 0.1412114783 | -0.093974139 | -0.015098192 | -0.164387325 | -0.440828593 | -0.725557215 | -0.054677636 |

(b)

| id | Inf_Intercept | Inf_HEALTHCARE_ACC... | Inf_CROWDING | Inf_sig_Intercept | Inf_sig_HEALTHCARE_ACC... | Inf_sig_CROWDING |
|---|---|---|---|---|---|---|
| 70 | 4325.3854789 | 71.515016481 | -521.7354493 | not significant at 90% | not significant at 90% | not significant at 90% |
| 71 | 230.34992596 | 7.02278029 | -35.92794547 | not significant at 90% | not significant at 90% | not significant at 90% |
| 72 | -17.4748332 | 0.1664348367 | 0.7921555314 | not significant at 90% | not significant at 90% | not significant at 90% |
| 73 | -9.85931019 | 0.1079415538 | 0.3628587549 | not significant at 90% | not significant at 90% | not significant at 90% |
| 74 | -9.428113165 | 0.0744360664 | 0.3700300297 | not significant at 90% | not significant at 90% | not significant at 90% |

(c)

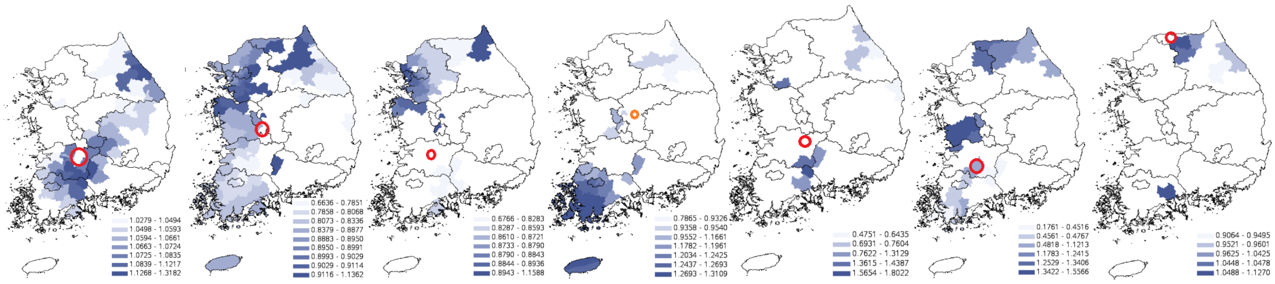
| id ▲ | Intercept | MORBIDITY | HIGH_SCH_P | HEALTHCARE_ACCESS | DIFF_SD | CROWDING | MIGRATION | HEALTH_BEHAVIOR |
|---|---|---|---|---|---|---|---|---|
| 70 | -20.0818934 | 0.2024054507 | -0.053884111 | -0.214056239 | 0.1169297566 | -0.22548122 | -0.254483127 | -0.002593652 |
| 71 | -11.28576421 | 0.0128383895 | -0.011009669 | -0.048624324 | -0.034368561 | 0.0655154978 | 0.0332458366 | 0.0109343375 |
| 72 | -8.714851076 | 0.0207687039 | -0.036676594 | -0.195877595 | 0.0084114729 | -0.090030731 | 0.1669855992 | 0.0353971291 |
| 73 | -3.801517613 | 0.0942539714 | -0.255681455 | 0.0109934509 | -0.132481459 | -0.366633558 | -0.014943557 | -0.040680432 |
| 74 | -8.121021032 | 0.1396561397 | -0.103942657 | -0.020037573 | -0.171146787 | -0.482014967 | -0.678289811 | -0.054715377 |

(d)

Figura 7.20: Visualização de algumas observações da base de dados com estimativas dos parâmetros: (a) $\alpha$ da RBNIZGP (binomial negativa inflacionada de zeros); (b) RBNIZGP (binomial negativa inflacionada de zeros); (c) RBNIZGP (Poisson)

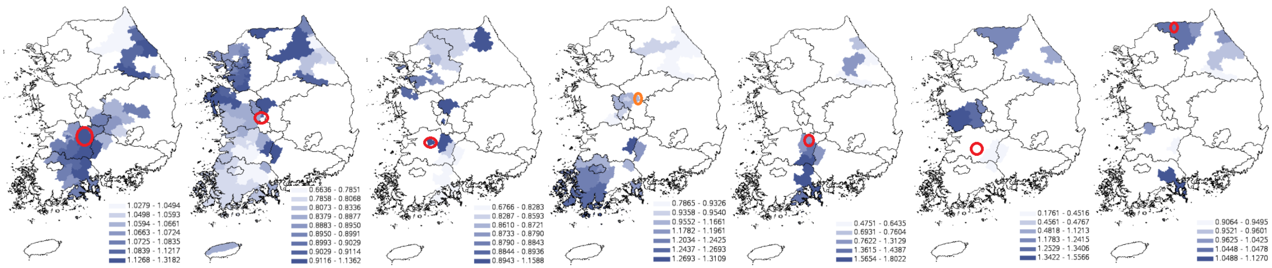
Nas Figuras 7.21 e 7.22 é possível identificar as mudanças espaciais no RR de COVID-19, comparando os dois modelos. Note que foi utilizada a mesma escala em ambas as Figuras, para cada variável, a fim de facilitar a comparação. A comorbidade (*MORBIDITY*) aumentou o risco de COVID-19 na região centro-sul e em alguns pontos da parte nordeste (sendo mais característico no caso do modelo binomial negativo). Já a proporção de pessoas com 2° grau (*HIGH_SCH_P*) aumentou esse risco na região noroeste e em alguns pontos da região norte também, no entanto, percebe-se um risco mais elevado na parte sul, no caso do modelo binomial negativo inflacionado de zeros.

Na variável acesso à saúde (*HEALTHCARE_ACCESS*) é possível identificar também um



(a) (b) (c) (d) (e) (f) (g)

Figura 7.21: Modelo binomial negativo - Variação espacial no risco relativo de COVID-19 nas variáveis (a) *MORBIDITY*, $(b)$*HIGH_SCH_P*, (c) *HEALTHCARE_ACCESS*, (d) *DIFF_SD*, (e) *CROWDING*, (f) *MIGRATION*, (g) *HEALTH_BEHAVIOR*



(a) (b) (c) (d) (e) (f) (g)

Figura 7.22: Modelo binomial negativo inflacionado de zeros - Variação espacial no risco relativo de COVID-19 nas variáveis (a) *MORBIDITY*, $(b)$*HIGH_SCH_P*, (c) *HEALTHCARE_ACCESS*, (d) *DIFF_SD*, (e) *CROWDING*, (f) *MIGRATION*, (g) *HEALTH_BEHAVIOR*

risco mais elevado na região noroeste (em ambos modelos), no entanto, no modelo binomial negativo inflacionado de zeros é possível identificar alguns trechos com destaque na parte central e na parte sul, diferente do modelo binomial negativo que apresenta um maior destaque de risco relativo em uma parte da região norte. A dificuldade em distanciamento social (*DIFF_SD*) e aglomeração (*CROWDING*) aumentaram o risco de COVID-19 em algumas partes na região sul, sendo mais presente no caso do modelo binomial negativo. Já a migração (*MIGRATION*) e o comportamento de saúde (*HEALTH_BEHAVIOR*) refletiram um risco de COVID-19 similar em ambos modelos, afetando mais a região centro-oeste (no caso da migração) e afetando a região sul e alguns picos na região norte (no caso do comportamento de saúde).

O próximo passo desta análise é fazer uma comparação em determinadas localidades, a fim de verificar a potencialidade do algoritmo RBNIZGP em identificar o melhor ajuste.

Baseando-se no modelo reduzido (em que todas as variáveis preditoras da parte inflacionada foram consideradas significativas), para a observação (ID=186), que se refere a região de Jinan, província de Jeollabuk, pode-se dizer que se a taxa de comorbidade (*MORBIDITY*) aumentar em uma unidade, o número esperado de casos de COVID-19 aumentaria por um fator de $\exp(0,0306) = 1,0310$, ou seja espera-se ver um aumento de $3,1\%$ no número de casos de COVID-19 no caso do modelo binomial negativo, diferente do modelo binomial negativo inflacionado de zeros, em que esse aumento seria de $13\%$, pois $\exp(0,1275) = 1,1359$. Na Figura 7.21a, é possível verificar a marcação da localidade, sendo que no modelo binomial negativo inflacionado de zeros esta variável foi considerada significativa, mas no modelo binomial negativo ela não foi considerada significativa. A fim de verificar a influência dos parâmetros estimados da parte inflacionada nos parâmetros estimados na parte não inflacionada, calculou-se a probabilidade predita para as covariáveis $\boldsymbol{G}$ desta observação: $\exp(G_i\boldsymbol{\gamma})/(1+\exp(G_i\boldsymbol{\gamma})) = \exp((1 \times (-15,2106)) + (36,0958 \times 0,1206) + (13,6873 \times 0,7041))/(1+\exp((1\times(-15,2106))+(36,0958\times 0,1206)+(13,6873\times 0,7041))) = 0,2279$. Vale ressaltar que entre as variáveis inflacionadas desta observação, somente o intercepto foi considerado significativo.

No caso da variável proporção de pessoas com 2° grau (*HIGH_SCH_P*), foi utilizada como referência a observação (ID = 52), referente a região de Daedeok , província de Daejeon. Portanto, no caso do modelo binomial negativo, nota-se que o número esperado de casos de COVID-19 para um aumento de uma unidade na proporção de pessoas com 2° grau é $-0,0691$, isso equivale a uma diminuição de $6,67\%$, pois $1 - \exp(-0,0691) = 0,0667$. Já no modelo binomial negativo inflacionado de zeros, o número esperado de casos de COVID-19 para um aumento de uma unidade na proporção de pessoas com 2° grau é $-0,0958$, sendo equivalente a uma diminuição de $9,13\%$, pois $1-\exp(-0,0958) = 0,0913$. Também utilizando a covariável $\boldsymbol{G}$ desta observação, tem-se:$\exp(G_i\boldsymbol{\gamma})/(1+\exp(G_i\boldsymbol{\gamma})) = \exp((1\times(-17,8904))+(17,1109\times$

$0,1274) + (12,4154 \times 0,8751))/(1 + \exp((-17,8904) + (17,1109 \times 0,1274) + (12,4154 \times 0,8751))) = 0,0078$. Como a diferença entre as estimativas dos modelos foi pequena, era esperado que essa probabilidade fosse pequena, como visto, sendo menor do que a probabilidade predita para a variável (*MORBIDITY*) vista anteriormente, que apresentou uma maior diferença entre os modelos. Nessa observação, vale ressaltar que entre as variáveis inflacionadas, somente o intercepto e a variável (*CROWDING*) foram considerados significativas.

Já na variável acesso à saúde (*HEALTHCARE_ACCESS*), foi utilizada como referência a observação (ID = 18), referente a região de Cheongwon, província de Chungcheongbuk. Sendo assim, ao aumentar em uma unidade dessa variável, o número esperado de casos de COVID-19 aumentaria por um fator de $\exp(0,0663) = 1,0685$, ou seja espera-se ver um aumento de $6,8\%$ no número de casos de COVID-19 no caso do modelo binomial negativo, diferente do modelo binomial negativo inflacionado de zeros, que essa esse aumento seria de $8,9\%$, pois $\exp(0,0849) = 1,0886$. Utilizando a covariável $\boldsymbol{G}$ desta observação, tem-se: $\exp(G_i\boldsymbol{\gamma})/(1 + \exp(G_i\boldsymbol{\gamma})) = \exp((1\times(-17,3958)) + (18,0329\times 0,1168) + (14,9292\times 0,8460))/(1+\exp((1\times(-17,3958)) + (18,0329 \times 0,1168) + (14,9292 \times 0,8460))) = 0,0655$, mostrando pouca influência na inclusão dessa parte inflacionada. Nese caso, entre as variáveis inflacionadas, somente a variável aglomeração (*CROWDING*) foi considerada significativa.

Para a variável dificuldade em distanciamento social (*DIFF_SD*), foi utilizada a observação (ID = 28), referente a região de Sangdang, província de Chungcheongbuk. No caso do modelo binomial negativo, nota-se que o número esperado de casos de COVID-19 para um aumento de uma unidade na variável é $-0,0957$, isso equivale a uma diminuição de $9,12\%$, pois $1 - \exp(-0,0957) = 0,0912$. Já no modelo binomial negativo inflacionado de zeros, o número esperado de casos de COVID-19 para um aumento de uma unidade na variável dificuldade em distanciamento social é $-0,0489$, sendo equivalente a uma diminuição de $4,78\%$, pois $1 - \exp(-0,0489) = 0,0478$. Utilizando a covariável $\boldsymbol{G}$ desta observação, tem-se: $\exp(G_i\boldsymbol{\gamma})/(1+\exp(G_i\boldsymbol{\gamma})) = \exp((1\times(-18,1573)) + (1,2602\times 0,1222) + (14,4881\times 0,8882))/(1+\exp((1\times(-18,1573)) + (1,2602\times 0,1222) + (14,4881\times 0,8882))) = 0,0058$, mostrando uma pequena influência da parte inflacionada. Nesse caso, entre as variáveis inflacionadas, somente a variável aglomeração (*CROWDING*) foi considerada significativa.

Na variável aglomeração (*CROWDING*), foi utilizada como referência a observação (ID = 186), que se refere a região de Jinan, província de Jeollabuk. Sendo assim, no caso do modelo binomial negativo, nota-se que o número esperado de casos de COVID-19 para um aumento de uma unidade na aglomeração é $-0,0600$, isso equivale a uma diminuição de $5,82\%$, pois $1 - \exp(-0,0600) = 0,0582$. No modelo binomial negativo inflacionado de zeros, o valor da aglomeração aumentaria o número de casos de COVID-19 em $24,7\%$, pois $\exp(0,2205) = 1,2467$. Tendo como base a utilização da covariável $\boldsymbol{G}$ desta observação, tem-se: $\exp(G_i\boldsymbol{\gamma})/(1+\exp(G_i\boldsymbol{\gamma})) = \exp((1 \times (-15,2106)) + (36,0958 \times 0,1206) + (13,6873 \times 0,7041))/(1+\exp((1\times(-15,2106))+(36,0958\times 0,1206)+(13,6873\times 0,7041))) = 0,2279$, mostrando uma probabilidace não tão pequena. Nesse caso, entre as variáveis inflacionadas desta observação, somente o intercepto foi considerado significativo.

A variável migração (*MIGRATION*), referente à observação (ID = 201), localizada na região de Jeollanam, província de Hwasun, foi considerada significativa no modelo binomial negativo e nota-se que o número esperado de casos de COVID-19 para um aumento de uma unidade na migração é $-0,6696$, isso equivale a uma diminuição de $48,8\%$, pois $1 - \exp(-0,6696) = 0,4880$. Já no modelo binomial negativo inflacionado de zeros essa variável não foi considerada significativa, sendo que o número esperado de casos de COVID-19 para um aumento de uma unidade na migração é $-0,7059$, o que equivale a uma diminuição de $50,6\%$, pois $1 - \exp(-0,7059) = 0,5063$. Utilizando a covariável $\boldsymbol{G}$ desta observação, tem-se: $\exp(G_i\boldsymbol{\gamma})/(1+\exp(G_i\boldsymbol{\gamma})) = \exp((1 \times (-16,8493)) + (26,3799 \times 0,1553) + (15,9508 \times 0,7400))/(1 + \exp((1 \times (-16,8493)) + (26,3799 \times 0,1553) + (15,9508 \times 0,7400))) = 0,2793$, ou seja, uma probabilidade não tão pequena. Nesse casos, entre as variáveis inflacionadas, somente a aglomeração (*CROWDING*) foi considerada significativa.

Já na variável comportamento de saúde (*HEALTH_BEHAVIOR*), foi utilizada a observação (ID=57), região de Gangwon, província de Cheorwon. No modelo binomial negativo, percebe-se que o valor do comportamento de saúde aumentaria o número de casos de COVID-19 em $6,36\%$, pois $\exp(0,0617) = 1,0636$. Já no modelo binomial negativo inflacionado de zeros este valor aumentaria o número de casos de COVID-19 em $4,5\%$, pois $\exp(0,0438) = 1,04478$. Tendo como base a utilização da covariável $\boldsymbol{G}$ desta observação, tem-se:$\exp(G_i\boldsymbol{\gamma})/(1+\exp(G_i\boldsymbol{\gamma})) = \exp((1\times(-19,5610))+(9,8621\times 0,15435)+(13,2704\times 0,9604))/(1+\exp((1\times(-19,5610))+(9,8621\times 0,15435)+(13,2704\times 0,9604))) = 0,0050$, mostrando uma baixa influência da parte inflacionada de zeros. Nesse caso, entre as variáveis inflacionadas desta linha, somente o intercepto foi considerado significativo.

Por último, a parte das variáveis inflacionadas de zeros permite um conhecimento maior sobre a ocorrência ou não de casos de COVID-19, diferentemente da parte não-inflacionada que permitia estimar o aumento no número de casos de COVID-19. Sendo assim, ela fornece uma análise adicional em que é possível identificar quais variáveis estão mais relacionadas com a ocorrência ou não de casos de COVID-19, ou seja, ela permite identificar a significância desses casos. Para esta análise, foi considerada somente a variável *CROWDING* na parte inflacionada de zeros, pois foram feitos alguns testes com o modelo anterior, e boa parte das estimativas significativas da variável *HEALTHCARE_ACCESS* ficaram com valores muito acima do esperado, o que gerava probabilidades da parte inflacionada muito próximas de zero.

No entanto, ao executar o algoritmo do *Golden Section Search* novamente, o parâmetro de suavização do modelo binomial negativo inflacionado de zeros foi igual a $174$. Portanto para garantir uma comparação justa, no modelo binomial negativo também foi utilizado o parâmetro de suavização igual a $174$.

Na taxa de comorbidade (*MORBIDITY*), considerando a linha (ID=41), ao aumentar em uma unidade, o número esperado de casos de COVID-19 aumentaria por um fator de $\exp(0,0154) = 1,0155$, ou seja espera-se ver um aumento de $1,5\%$ no número de casos de COVID-19 no caso do modelo binomial negativo, diferente do modelo binomial negativo inflacionado de zeros, que essa esse aumento seria de $2,1\%$, pois $\exp(0,0209) = 1,0211$. Tendo como base a utilização da covariável $\boldsymbol{G}$ desta linha, tem-se: $\exp(G_i\boldsymbol{\gamma})/(1+\exp(G_i\boldsymbol{\gamma})) = \exp((1\times(-12,7738))+$

$(16,0111 \times 0,8315))/(1+\exp((1 \times (-12,7738))+(16,0111 \times 0,8315))) = 0,6317$, mostrando uma probabilidade não desprezível. Nesta observação, vale ressaltar que entre as variáveis inflacionadas, o intercepto e a variável *CROWDING* inflacionada foram consideradas significativas.

Na variável aglomeração (*CROWDING*), considerando a linha (ID=33), ao aumentar em uma unidade, o número esperado de casos de COVID-19 aumentaria por um fator de $\exp(0,0554) = 1,0569$, ou seja espera-se ver um aumento de $5,7\%$ no número de casos de COVID-19 no caso do modelo binomial negativo, diferente do modelo binomial negativo inflacionado de zeros, que essa esse aumento seria de $29,6\%$, pois $\exp(0,2596) = 1,2962$. Tendo como base a utilização da covariável $\boldsymbol{G}$ desta linha, tem-se: $\exp(G_i\boldsymbol{\gamma})/(1+\exp(G_i\boldsymbol{\gamma})) = \exp((1 \times (-14,5175)) + (13,2432 \times 0,9362))/(1+\exp((1 \times (-14,5175)) + (13,2432 \times 0,9362))) = 0,1073$. Neste caso, o intercepto inflacionado e a variável *CROWDING* inflacionada também foram consideradas significativas.

Nestes dois exemplos, as variáveis do modelo binomial negativo inflacionado de zeros foram consideradas significativas à 5% de significância, mas no modelo binomial negativo não. Já nas outras variáveis, não houveram casos (observações) em que o modelo binomial negativo inflacionado de zeros foi considerado significativo e que o modelo binomial negativo não foi considerado significativo.

Na Figura 7.23 é possível notar que existe uma chance mais do que duas vezes das localidades apresentaram casos de COVID-19 quando a variável aglomeração (*CROWDING*) aumenta em uma unidade na província de Gangwon, em algumas partes no Noroeste, sendo alguns trechos da região Gyeonggi, Chungcheongnam, Gangwon e na capital Seoul.

Uma outra alternativa de análise foi criada, dessa vez forçando nos dois modelos o parâmetro de suavização igual a $82$ (como anteriormente) e no caso do modelo binomial negativo inflacionado de zeros manteve-se somente a variável *CROWDING* na parte inflacionada de zeros.

Portanto, na taxa de comorbidade (*MORBIDITY*), considerando a linha (ID=200), ao aumentar em uma unidade, o número esperado de casos de COVID-19 aumentaria por um fator

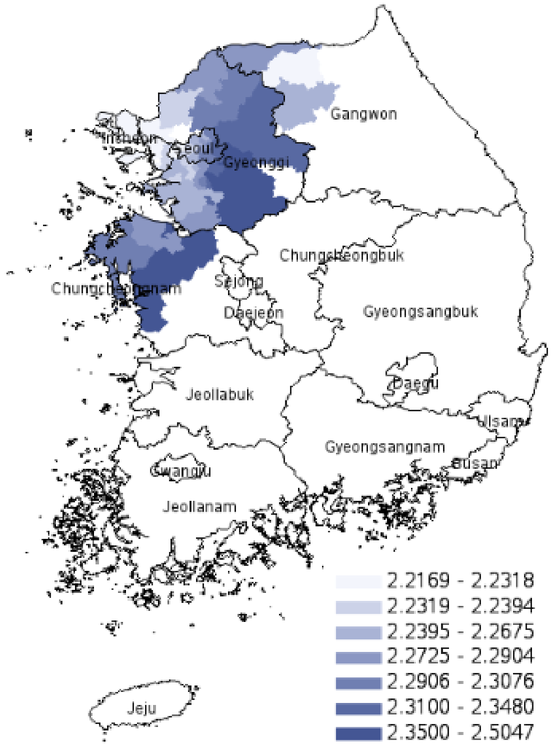



Figura 7.23: Modelo binomial negativo inflacionado de zeros - Variação espacial na razão de chances na variável inflacionada *CROWDING*, considerando parâmetro de suavização igual 174

de $\exp(0,0532) = 1,0546$, ou seja espera-se ver um aumento de $5,4\%$ no número de casos de COVID-19 no caso do modelo binomial negativo, diferente do modelo binomial negativo inflacionado de zeros, que essa esse aumento seria de $5,9\%$, pois $\exp(0,0576) = 1,0593$. Tendo como base a utilização da covariável $\boldsymbol{G}$ desta linha, tem-se: $\exp(G_i\boldsymbol{\gamma})/(1+\exp(G_i\boldsymbol{\gamma})) = \exp((1 \times (-15,3854)) + (18,1593 \times 0,9119))/(1+\exp((1 \times (-15,3854)) + (18,1593 \times 0,9119))) = 0,7639$, mostrando uma probabilidade não desprezível. Neste caso, somente o intercepto inflacionado foi considerado significativo.

Na variável dificuldade em distanciamento social (*DIFF_SD*), considerando a linha (ID = 176), ao aumentar em uma unidade, o número esperado de casos de COVID-19 aumentaria por um fator de $\exp(0,2226) = 1,2493$, ou seja espera-se ver um aumento de $24,9\%$ no número de casos de COVID-19 no caso do modelo binomial negativo, diferente do modelo binomial negativo inflacionado de zeros, que essa esse aumento seria de $22,49\%$, pois $\exp(0,2029) =$

$1,2249$. Utilizando a covariável $\boldsymbol{G}$ desta linha, tem-se: $\exp(G_i\boldsymbol{\gamma})/(1+\exp(G_i\boldsymbol{\gamma})) = \exp((1 \times (-15,4521)) + (16,0147 \times 0,9152))/(1+\exp((1 \times (-15,4521)) + (16,0147 \times 0,9152))) = 0,3111$, reforçando assim que todas as variáveis inflacionadas desta linha foram consideradas significativas.

Nestes dois exemplos, as variáveis do modelo binomial negativo inflacionado de zeros foram consideradas significativas à 5% de significância, mas no modelo binomial negativo não. Já nas outras variáveis, não houveram casos (observações) em que o modelo binomial negativo inflacionado de zeros foi considerado significativo e que o modelo binomial negativo não foi considerado significativo.

Na Figura 7.24, é possível notar também que existe uma chance mais do que duas vezes das localidades apresentaram casos de COVID-19 quando a variável aglomeração (*CROWDING*) na província de Daejeon, no sul da província de Jeollabuk e um pequeno trecho de Gangwon, ou seja, lugares diferentes do aquele estimado quando o parâmetro de suavização foi igual a 174. Ainda assim, aparecem alguns locais onde o risco relativo ficou muito elevado, mostrando um possível problema de convergência. Portanto, nesse caso, pode-se concluir que o modelo binomial negativo inflacionado de zeros se assemelha ao modelo binomial negativo, fazendo com que possivelmente o modelo binomial negativo seja o mais indicado nesta análise.

A diferença entre as Figuras 7.23 e 7.24 mostra que o parâmetro de suavização deve ser recalibrado a fim de considerar a influência de todas as variáveis no modelo. Isso porque como foi visto, a razão de chances estimada quando o parâmetro de suavização foi igual a 82, ou seja, igual ao modelo com as variáveis inflacionadas *HEALTH_BEHAVIOR* e *CROWDING*, se assemelhou muito ao modelo binominal negativo, além de ter valores muito grandes. Já o modelo com o parâmetro de suavização igual a 174, ou seja, quando o modelo foi recalibrado, mostrou resultados muito mais consistentes.

Além disso, a significância dessa variável *CROWDING* na parte inflacionada traz um ganho de informação não presente no modelo binomial negativo, que é a possibilidade de analisar quais variáveis estão mais relacionadas ao aparecimento ou não de casos de COVID-19, ou

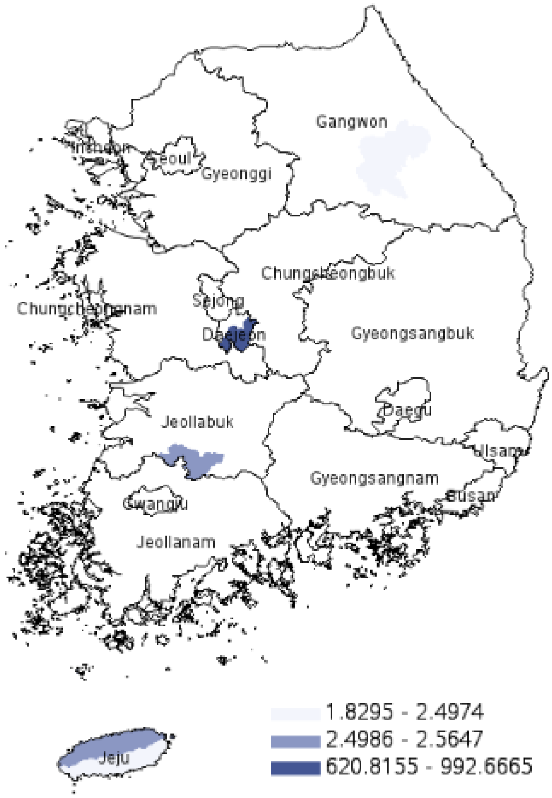



Figura 7.24: Modelo binomial negativo inflacionado de zeros - Variação espacial na razão de chances na variável inflacionada *CROWDING*, considerando parâmetro de suavização igual a 82

seja, é possível fazer uma análise da razão de chances. Apesar de ter uma pequena diferença entre as estimativas e medidas de ajuste (com vantagem para o modelo binomial negativo), a inclusão da variável *CROWDING* na parte inflacionada trouxe alteração na significância de algumas estimativas locais na parte não inflacionada do modelo.

# Capítulo 8

# Conclusões

O principal intuito deste trabalho foi trazer a abordagem da modelagem de dados de contagem em que existam uma quantidade considerável de zeros na distribuição. Sendo assim, partindo do modelo RBNGP, proposto por Da Silva e Rodrigues (2014), foi desenvolvido o modelo de regressão geograficamente ponderado utilizando a distribuição binomial negativa inflacionada de zeros, denominado regressão binomial negativa inflacionada de zeros geograficamente ponderada (RBNIZGP).

Neste trabalho foi desenvolvida uma estrutura geral, onde é possível entender que as distribuições Poisson, binomial negativa e Poisson inflacionada de zeros são casos especiais da distribuição binomial negativa inflacionada de zeros. As simulações permitiram verificar que o algoritmo desenvolvido para a RBNIZGP consegue acomodar todas essas distribuições, além de conseguir modelar dados sem dependência espacial.

No estudo de caso, inicialmente foi utilizado o modelo binomial negativo para analisar os casos de COVID-19 na Coréia do Sul, no entanto, tendo como base no histograma deste número de casos, foi possível verificar a quantidade de zeros que existia na distribuição, sendo assim, o modelo mais adequado para tal análise seria o binomial negativo inflacionado de zeros. Os resultados mostraram que em algumas localidades, os dados seriam modelados por uma distribuição Poisson e binomial negativa, mostrando a flexibilidade do algoritmo. Foi possível a comparação do resultados entre os modelos locais Poisson e binomial negativo, pois coincidentemente, o melhor parâmetro de suavização encontrado foi o mesmo nos dois modelos, como pode ser visto na Tabela 7.10. Além disso, os resultados mostraram que quando variáveis são retiradas e/ou adicionadas ao modelo, deve-se estimar novamente o parâmetro de suavização, a fim de melhor adequar a dependência espacial. Outra análise interessante foi a possibilidade de analisar a razão de chances, ou seja, a possibilidade de identificar quais variáveis estão mais relacionadas ao aparecimento ou não de casos de COVID-19. No caso, a variável aglomeração (*CROWDING*) se mostrou significativa, o que faz sentido e o que levou diversos países a decretarem o *lockdown*.

Portanto, pode-se concluir que o algoritmo RBNIZGP desenvolvido neste trabalho é mais geral do que os modelos RPGP, RBNGP e RPIZGP, facilitando dessa forma a análise por parte do usuário. Isso porque não há mais a necessidade de verificar qual a distribuição dos dados antes da análise, pois se houver muitos zeros ou se a quantidade de zeros não for muito elevada, o algoritmo RBNIZGP irá se ajustar de acordo com a distribuição dos dados. Reitera-se que o uso do modelo binomial negativo inflacionado de zeros só faz sentido se houver zeros na distribuição.

## 8.1 Limitações do trabalho

Uma das limitações deste trabalho seria em relação a ideia mostrada na Figura 6.5, que trata da relação entre os modelos na RBNIZGP em sua forma local. Na Figura 7.20a, que representa a base de dados com estimativas dos parâmetros de superdispersão $\alpha$ na distribuição binomial negativa inflacionada de zeros, percebe-se na observação 70, que o parâmetro de superdispersão $\alpha$ foi considerado significativo para $10\%$ de significância, no entanto, as estimativas desta linha nas respectivas bases de dados (Figuras 7.20b e 7.20d) não são iguais ou não são próximas devido às estimativas de $\gamma$, (ver Figura 7.20c), terem valores não nulos, mas não terem sido significativas, considerando por exemplo os mesmos $10\%$ de nível de significância.

Uma outra limitação identificada seria a questão do tempo de processamento do algoritmo. Na Tabela 8.1 estão os tempos de processamento dos algoritmos para a função Fixa e Adptável, tendo como base o modelo completo na binomial negativa inflacionada de zeros (Tabela 7.8). Note que os tempos de processamento do algoritmo RBNGP em comparação ao algoritmo RBNIZGP foram muito menores, sendo necessário apenas alguns segundos para sua conclusão.

Tabela 8.1: Tempo de processamento dos algoritmos

| **Algoritmo** | **Parâmetro de Suavização** | |
|---|---|---|
| | **Fixo** | **Adaptável** |
| *Golden Section Search (RBNIZGP)* | $06min : 44seg$ | $01h : 38min$ |
| *RBNIZGP* | $00min : 28seg$ | $11min : 58seg$ |
| *Golden Section Search (RBNGP)* | $00min : 15seg$ | $00min : 31seg$ |
| *RBNGP* | $00min : 1seg$ | $00min : 2seg$ |

## 8.2 Sugestões para trabalhos futuros

Com base nas conclusões apresentadas, seguem algumas sugestões para trabalhos futuros:

- Gerar uma estrutura local de significância, pois em alguns locais o modelo pode ser significativo para a distribuição binomial negativa, em outros para a distribuição binomial negativa inflacionada de zeros e assim sucessivamente para os demais modelos.
- Verificar uma forma de otimização no tempo de processamento do algoritmo.
- Analisar o motivo do pseudo $R^2$ ter tido uma grande variação nos parâmetros de suavização fixo e adaptável (considerando cada modelo), visto que os outros parâmetros se manteram estáveis.

# Apêndice A

# Algoritmos

**Algoritmo 8:** Regressão Poisson Inflacionada de Zeros

**Entrada:** $D_0{=}0, \quad DiffD{=}1, \quad offset{=}0, \quad itr{=}1$

$p_0{=}y[y{=}0], \quad p_1{=}y[y{>}0]$

$\boldsymbol{\mu}{=}(\boldsymbol{y}[p_1]{+}\overline{y}[p_1])/2$

$\boldsymbol{\beta}{=}[\boldsymbol{X}^T[p_1,]\boldsymbol{X}[p_1,]]^{-1}\boldsymbol{X}^T[p_1,]\log(\boldsymbol{\mu})$

$\boldsymbol{\eta}{=}\boldsymbol{X}\boldsymbol{\beta}{+}offset; \quad \boldsymbol{\mu}{=}\exp(\boldsymbol{\eta})$

$zk{=}1/(1{+}\exp(-\boldsymbol{G}\boldsymbol{\gamma}{-}\exp(\boldsymbol{X}\boldsymbol{\beta})))$

**enquanto** *(*$abs(DiffD){>}10^{-6}$*)* **faça**

  $D_0{=}0, \quad D{=}1$

  **enquanto** *(*$abs(dif \quad D){>}10^{-6}$*)* **faça**

    $\boldsymbol{A}{=}(1{-}zk)\boldsymbol{\mu}; \quad \boldsymbol{z}{=}\boldsymbol{\eta}{+}\frac{\boldsymbol{y}-\boldsymbol{\mu}}{\boldsymbol{\mu}}{-}offset$

    $\boldsymbol{\beta}{=}[\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{A}\boldsymbol{X}]^{-\boldsymbol{1}}\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{A}\boldsymbol{z}; \quad \boldsymbol{\eta}{=}\boldsymbol{X}\boldsymbol{\beta}{+}offset; \quad \boldsymbol{\mu}{=}\exp(\boldsymbol{\eta})$

    $oldD_0{=}D_0; \quad D_0{=}\sum_{i=1}^{n}[(1{-}zk)(\boldsymbol{y}\boldsymbol{\eta}{-}\boldsymbol{\mu})]; \quad D{=}D{-}oldD_0$

    $\boldsymbol{H}{=}[\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{A}]^{-1}\boldsymbol{X}$

    $D_0{=}0, \quad D{=}1$

    $\boldsymbol{\eta}{=}\boldsymbol{G}\boldsymbol{\gamma}; \quad \pi{=}\frac{\exp(\boldsymbol{\eta})}{(1+\exp(\boldsymbol{\eta}))}$

  **fim**

  **enquanto** *(*$abs(dif \quad D){>}10^{-6}$*)* **faça**

    $\boldsymbol{A}{=}\pi(1{-}\pi); \quad \boldsymbol{z}{=}\boldsymbol{\eta}{+}(zk{-}\pi)\frac{1}{\pi(1-\pi)}$

    $\boldsymbol{\gamma}{=}[\boldsymbol{G}^{\boldsymbol{T}}\boldsymbol{A}\boldsymbol{G}]^{-1}\boldsymbol{G}^{\boldsymbol{T}}\boldsymbol{A}\boldsymbol{z}$

    $\boldsymbol{\eta}{=}\boldsymbol{G}\boldsymbol{\gamma}; \quad \pi{=}\frac{\exp(\boldsymbol{\eta})}{(1+\exp(\boldsymbol{\eta}))}$

    $oldD_0{=}D_0; \quad D_0{=}\sum_{i=1}^{n}(zk\boldsymbol{\eta}{-}\sum_{i=1}^{n}(\log(1{+}\exp(\boldsymbol{\eta})))); \quad D{=}D{-}oldD_0$

    $\boldsymbol{H}{=}[\boldsymbol{G}^{\boldsymbol{T}}\boldsymbol{A}]^{-1}\boldsymbol{G}; \quad zk{=}1/(1{+}\exp(-\boldsymbol{G}\boldsymbol{\gamma}{-}\exp(\boldsymbol{X}\boldsymbol{\beta})))$

  **fim**

  $OldD{=}D; \quad D{=}\sum_{i=1}^{n}(\log(\exp(\boldsymbol{G}[p_0,]\boldsymbol{\gamma}){+}\exp(-\exp(\boldsymbol{X}[p_0,]\boldsymbol{\beta})))){+}\sum_{i=1}^{n}(\boldsymbol{y}[p_1](\boldsymbol{X}[p_1,]\boldsymbol{\beta}){-}\exp(\boldsymbol{X}[p_1,]\boldsymbol{\beta})){-}$

  $\sum_{i=1}^{n}(\log(1{+}\exp(\boldsymbol{G}\boldsymbol{\gamma}))){-}\sum_{i=1}^{n}(\log(\boldsymbol{y}!\,[p_1]))$

  $DiffD{=}OldD{-}D; \quad itr{=}itr{+}1$

**fim**

**Algoritmo 9:** Regressão Binomial Negativa Inflacionada de Zeros

**Entrada:** $D_0{=}0, \quad DiffD{=}1, \quad offset{=}0, \quad itr{=}1, \quad k{=}1$

$p_0{=}y[y{=}0], \quad p_1{=}y[y{>}0]$

$\boldsymbol{\mu}{=}(\boldsymbol{y}[p_1]{+}\overline{y}[p_1])/2$

$\boldsymbol{\beta}{=}[\boldsymbol{X}^T[p_1,]\boldsymbol{X}[p_1,]]^{-1}\boldsymbol{X}^T[p_1,]\log(\boldsymbol{\mu})$

$\boldsymbol{\eta}{=}\boldsymbol{X}\boldsymbol{\beta}, \quad \boldsymbol{\mu}{=}\exp(\boldsymbol{\eta})$

$zk{=}1/(1{+}\exp(-\boldsymbol{G}\boldsymbol{\gamma}))(k/(k{+}\exp(\boldsymbol{X}\boldsymbol{\beta})))^k$

**enquanto** *(* $abs(DiffD){>}10^{-6} \quad \& \quad itr{<}100$*)* **faça**

> $dpar{=}1$
>
> **enquanto** *(* $abs(dif \quad dpar){>}10^{-6}$*)* **faça**
>
> > $aux1{=}1, \quad aux2{=}1, \quad ,Dk{=}1, \quad old_k{=}k, \quad itr{=}1$
>
> **fim**
>
> **enquanto** *(* $abs(dif \quad Dk){>}10^{-6}, \quad \& \quad aux2{<}500$*)* **faça**
>
> > $\boldsymbol{g}{=}\sum_{i=1}^{n}[\partial\Gamma(k{+}(1{-}zk)y){-}\partial\Gamma(k){+}\log(k){+}1{-}\log(k{+}(1{-}zk)\mu_i){-}(k{+}(1{-}zk)y_i)/(k{+}(1{-}zk)\mu_i)]$
> >
> > $\boldsymbol{H}{=}\sum_{i=1}^{n}[\partial^2\Gamma(k{+}(1{-}zk)y){-}\partial^2\Gamma(k){+}1/k{-}2/(k{+}(1{-}zk)\mu_i){+}((1{-}zk)y{+}k)/((k{+}(1{-}zk)\mu_i)^2)]$
> >
> > $k0{=}k, \quad k{=}k0{-}H^{-1}g, \quad \alpha{=}1/k, \quad D_0{=}0, \quad D{=}1$
>
> **fim**
>
> **enquanto** *(* $abs(D){>}10^{-6}, \quad \& \quad aux1{<}100$*)* **faça**
>
> > $\boldsymbol{A}{=}(1{-}zk)\boldsymbol{\mu}; \quad \boldsymbol{A0}{=}(1{-}zk)(\boldsymbol{\mu}/(1{+}\alpha\boldsymbol{\mu})){+}(\boldsymbol{y}{-}\boldsymbol{\mu})(\alpha\boldsymbol{\mu}/(1{+}2\alpha\boldsymbol{\mu}{+}\alpha^2\boldsymbol{\mu}^2))$
> >
> > $\boldsymbol{z}{=}\boldsymbol{\eta}{+}(\boldsymbol{y}{-}\boldsymbol{\mu})/(\boldsymbol{\mu})); \quad \boldsymbol{\beta}{=}[\boldsymbol{X^T A X}]^{-1}\boldsymbol{X^T A z}$
> >
> > $\boldsymbol{\eta}{=}\boldsymbol{X}\boldsymbol{\beta}{+}offset; \quad \boldsymbol{\mu}{=}\exp(\boldsymbol{\eta})$
> >
> > $oldD_0{=}D_0$
> >
> > $M{=}\boldsymbol{\mu}/(\boldsymbol{\mu}{+}k)^y(k/(\boldsymbol{\mu}{+}k))^k$
> >
> > $D_0{=}\sum_{i=1}^{n}[(1{-}zk)(\log(M)); \quad D{=}D_0{-}oldD_0$
> >
> > $aux1{=}aux1{+}1; \quad dpar{=}k{-}old_k$
> >
> > $\boldsymbol{Hb}{=}[\boldsymbol{X^T A0}]^{-1}\boldsymbol{X}$
> >
> > $D_0{=}0, \quad D{=}1$
> >
> > $\boldsymbol{\eta}{=}\boldsymbol{G}\boldsymbol{\gamma}$
> >
> > $\pi{=}\frac{\exp(\boldsymbol{\eta})}{(1{+}\exp(\boldsymbol{\eta}))}$
>
> **fim**
>
> **enquanto** *(*$abs(dif \quad D){>}10^{-6}$ *)* **faça**
>
> > $\boldsymbol{A}{=}\pi(1{-}\pi); \quad \boldsymbol{z}{=}\boldsymbol{\eta}{+}(zk{-}\pi)\frac{1}{\pi(1-\pi)}$
> >
> > $\boldsymbol{\gamma}{=}[\boldsymbol{G^T A G}]^{-1}\boldsymbol{G^T A z}$
> >
> > $\boldsymbol{\eta}{=}\boldsymbol{G}\boldsymbol{\gamma}$
> >
> > $\pi{=}\frac{\exp(\boldsymbol{\eta})}{(1{+}\exp(\boldsymbol{\eta}))}$
> >
> > $oldD_0{=}D_0$
> >
> > $D_0{=}\sum_{i=1}^{n}(zk\boldsymbol{\eta}{-}\sum_{i=1}^{n}(\log(1{+}\exp(\boldsymbol{\eta})); \quad D{=}D{-}oldD_0$
> >
> > $\boldsymbol{Hl}{=}[\boldsymbol{G^T A}]^{-1}\boldsymbol{G}$
> >
> > $zk{=}1/(1{+}\exp(-\boldsymbol{G}\boldsymbol{\gamma}(k/(k{+}\exp(\boldsymbol{X}\boldsymbol{\beta})))^k$
> >
> > $OldD{=}D, \quad D{=}\sum_{i=1}^{n}(-\log(1{+}\exp(\boldsymbol{G}[p_0,]\boldsymbol{\gamma}){+}\log(\exp\boldsymbol{X}[p_0,]\boldsymbol{\beta}){+}(k/(k{+}\exp(\boldsymbol{X}[p_0,]\boldsymbol{\beta})))^k{+}$
> >
> > $\sum_{i=1}^{n}(-\log(1{+}\exp(\boldsymbol{G}[p_1,]\boldsymbol{\gamma})){+}\boldsymbol{y}[p_1]\log(\exp(\boldsymbol{X}[p_1,]\boldsymbol{\beta})/(k{+}\exp(\boldsymbol{X}[p_1,]\boldsymbol{\beta}))){+}k\log(k/(k{+}\exp(\boldsymbol{X}[p_1,]\boldsymbol{\beta})))$
> >
> > $DiffD{=}OldD{-}D; \quad itr{=}itr{+}1$
>
> **fim**

**fim**

# Apêndice B

# Demonstração parâmetros da regressão Poisson inflacionada de zeros

A demonstração dos parâmetros usam a função de log-verossimilhança para o vetor de parâmetros $\boldsymbol{\theta} = (\boldsymbol{\beta}^T, \boldsymbol{\gamma}^T)^T$ dada por:

$$
\begin{aligned}
l(\boldsymbol{\theta}) \;\; = \;\; & \sum_{i=1}^{n} \log\left\{\exp(\boldsymbol{z}_i^T\boldsymbol{\gamma}) + \exp(-\exp(\boldsymbol{x}_i^T\boldsymbol{\beta}))\right\} I_{(yi=0)} + \sum_{i=1}^{n}\log[1+\exp(\boldsymbol{z}_i^T\boldsymbol{\gamma})] \\
& + \sum_{i=1}^{n}\left\{y_i\boldsymbol{x}_i^T\boldsymbol{\beta} - \exp(\boldsymbol{x}_i^T\boldsymbol{\beta}\right\} I_{(yi>0)} \qquad \text{(B.1)}
\end{aligned}
$$

$$
\begin{aligned}
\frac{\partial l(\boldsymbol{\theta})}{\partial \gamma_j} \;\; &= \;\; \sum_{i;y=0} \frac{\exp(\boldsymbol{z}_i^T\boldsymbol{\gamma})\mathbf{z}_i}{\exp(\boldsymbol{z}_i^T\boldsymbol{\gamma}) + \exp(-\exp(\boldsymbol{x}_i^T\boldsymbol{\beta}))} - \sum_{i=1}^{n}\frac{\exp(\boldsymbol{z}_i^T\boldsymbol{\gamma})\mathbf{z}_i}{1+\exp(\boldsymbol{z}_i^T\boldsymbol{\gamma})} \\
\mathbf{L}_{\gamma_j\gamma_k} = \frac{\partial^2 l(\boldsymbol{\theta})}{\partial \gamma_j \partial \gamma_k} \;\; &= \;\; \sum_{i;y=0} \frac{-[\exp(\boldsymbol{z}_i^T\boldsymbol{\gamma})]^2 z_{ij}z_{ik}}{[\exp(\boldsymbol{z}_i^T\boldsymbol{\gamma}) + \exp(-\exp(\boldsymbol{x}_i^T\boldsymbol{\beta}))]^2} + \frac{\exp(\boldsymbol{z}_i^T\boldsymbol{\gamma})z_{ij}z_{ik}}{\exp(\boldsymbol{z}_i^T\boldsymbol{\gamma}) + \exp(-\exp(\boldsymbol{x}_i^T\boldsymbol{\beta}))} \\
& \quad + \sum_{i=1}^{n}\left\{\frac{[\exp(\boldsymbol{z}_i^T\boldsymbol{\gamma})]^2 z_{ij}z_{ik}}{[1+\exp(\boldsymbol{z}_i^T\boldsymbol{\gamma})]^2} - \frac{\exp(\boldsymbol{z}_i^T\boldsymbol{\gamma})z_{ij}z_{ik}}{1+\exp(\boldsymbol{z}_i^T\boldsymbol{\gamma})}\right\} \\
&= \;\; \sum_{i;y=0} \frac{\exp(\boldsymbol{z}_i^T\boldsymbol{\gamma})z_{ij}z_{ik}(-\exp(\boldsymbol{z}_i^T\boldsymbol{\gamma}) + \exp(\boldsymbol{z}_i^T\boldsymbol{\gamma}) + \exp(-\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})))}{[\exp(\boldsymbol{z}_i^T\boldsymbol{\gamma}) + \exp(-\exp(\boldsymbol{x}_i^T\boldsymbol{\beta}))]^2} \\
& \quad + \sum_{i=1}^{n}\frac{\exp(\boldsymbol{z}_i^T\boldsymbol{\gamma})z_{ij}z_{ik}(\exp(\boldsymbol{z}_i^T\boldsymbol{\gamma}) - 1 - \exp(\boldsymbol{z}_i^T\boldsymbol{\gamma}))}{[1+\exp(\boldsymbol{z}_i^T\boldsymbol{\gamma})]^2} \\
\mathbf{L}_{\gamma_j\gamma_k} = \frac{\partial^2 l(\boldsymbol{\theta})}{\partial \gamma_j \partial \gamma_k} \;\; &= \;\; \sum_{i;y=0} \frac{\exp(\boldsymbol{z}_i^T\boldsymbol{\gamma})z_{ij}z_{ik}\exp(-\exp(\boldsymbol{x}_i^T\boldsymbol{\beta}))}{[\exp(\boldsymbol{z}_i^T\boldsymbol{\gamma}) + \exp(-\exp(\boldsymbol{x}_i^T\boldsymbol{\beta}))]^2} - \sum_{i=1}^{n}\frac{\exp(\boldsymbol{z}_i^T\boldsymbol{\gamma})z_{ij}z_{ik}}{[1+\exp(\boldsymbol{z}_i^T\boldsymbol{\gamma})]^2}
\end{aligned}
$$

$$\mathbf{L}_{\gamma_j\beta_k} = \frac{\partial^2 l(\boldsymbol{\theta})}{\partial\gamma_j\partial\beta_k} \quad = \quad \sum_{i;y=0} \frac{\exp(\boldsymbol{z}_i^T\boldsymbol{\gamma})z_{ij}\exp(-\exp(\boldsymbol{x}_i^T\boldsymbol{\beta}))\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})x_{ik}}{[\exp(\boldsymbol{z}_i^T\boldsymbol{\gamma})+\exp(-\exp(\boldsymbol{x}_i^T\boldsymbol{\beta}))]^2}$$

$$\frac{\partial l(\boldsymbol{\theta})}{\partial\beta_j} \quad = \quad \sum_{i;y=0} \frac{-\exp(-\exp(\boldsymbol{x}_i^T\boldsymbol{\beta}))\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})\mathbf{x}_i}{\exp(\boldsymbol{z}_i^T\boldsymbol{\gamma})+\exp(-\exp(\boldsymbol{x}_i^T\boldsymbol{\beta}))} + \sum_{i;y>0}(y_i\mathbf{x}_i - \exp(\boldsymbol{x}_i^T\boldsymbol{\beta})\mathbf{x}_i)$$

$$\begin{aligned}\mathbf{L}_{\beta_j\beta_k} = \frac{\partial^2 l(\boldsymbol{\theta})}{\partial\beta_j\partial\beta_k} \quad &= \quad \sum_{i;y=0} \frac{-[-\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})\exp(-\exp(\boldsymbol{x}_i^T\boldsymbol{\beta}))]^2 x_{ij}x_{ik}}{[\exp(\boldsymbol{z}_i^T\boldsymbol{\gamma})+\exp(-\exp(\boldsymbol{x}_i^T\boldsymbol{\beta}))]^2} + \sum_{i;y=0} \frac{\exp(-\exp(\boldsymbol{x}_i^T\boldsymbol{\beta}))(\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})^2 x_{ij}x_{ik}}{\exp(\boldsymbol{z}_i^T\boldsymbol{\gamma})+\exp(-\exp(\boldsymbol{x}_i^T\boldsymbol{\beta}))} \\ &\quad - \sum_{i;y=0} \frac{\exp(-\exp(\boldsymbol{x}_i^T\boldsymbol{\beta}))\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})x_{ij}x_{ik}}{\exp(\boldsymbol{z}_i^T\boldsymbol{\gamma})+\exp(-\exp(\boldsymbol{x}_i^T\boldsymbol{\beta}))} - \sum_{i;y>0} x_{ij}x_{ik}\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})\end{aligned}$$

$$\begin{aligned}\mathbf{L}_{\beta_j\beta_k} = \frac{\partial^2 l(\boldsymbol{\theta})}{\partial\beta_j\partial\beta_k} \quad &= \quad \sum_{i;y=0} \frac{\frac{-[\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})]^2[\exp(-\exp(\boldsymbol{x}_i^T\boldsymbol{\beta}))]^2 x_{ij}x_{ik}}{\exp(\boldsymbol{z}_i^T\boldsymbol{\gamma})+\exp(-\exp(\boldsymbol{x}_i^T\boldsymbol{\beta}))} + \exp(-\exp(\boldsymbol{x}_i^T\boldsymbol{\beta}))[\exp(\boldsymbol{x}_i^T\boldsymbol{\beta}]^2 x_{ij}x_{ik}}{\exp(\boldsymbol{z}_i^T\boldsymbol{\gamma})+\exp(-\exp(\boldsymbol{x}_i^T\boldsymbol{\beta}))} \\ &\quad - \sum_{i;y=0} \frac{\exp(-\exp(\boldsymbol{x}_i^T\boldsymbol{\beta}))\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})x_{ij}x_{ik}}{\exp(\boldsymbol{z}_i^T\boldsymbol{\gamma})+\exp(-\exp(\boldsymbol{x}_i^T\boldsymbol{\beta}))} - \sum_{i;y>0} x_{ij}x_{ik}\exp(\boldsymbol{x}_i^T\boldsymbol{\beta}) \\ &= \quad \sum_{i;y=0} \frac{\exp(-\exp(\boldsymbol{x}_i^T\boldsymbol{\beta}))\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})x_{ij}x_{ik}\left(\exp(\boldsymbol{x}_i^T\boldsymbol{\beta}) - \frac{\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})\exp(-\exp(\boldsymbol{x}_i^T\boldsymbol{\beta}))}{\exp(\boldsymbol{z}_i^T\boldsymbol{\gamma})+\exp(-\exp(\boldsymbol{x}_i^T\boldsymbol{\beta}))} - 1\right)}{\exp(\boldsymbol{z}_i^T\boldsymbol{\gamma})+\exp(-\exp(\boldsymbol{x}_i^T\boldsymbol{\beta}))} \\ &\quad - \sum_{i;y>0} x_{ij}x_{ik}\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})\end{aligned}$$

# Apêndice C

# Demonstração parâmetros da regressão binomial negativa inflacionada de zeros

A demonstração dos parâmetros usam a função de log-verossimilhança para o vetor de parâmetros $\boldsymbol{\theta} = (\phi, \boldsymbol{\beta}^T, \boldsymbol{\gamma}^T)^T$ dada por:

$$\begin{aligned} l(\boldsymbol{\theta}) &= \sum_{i=1}^{n} \log \left\{ \exp(\boldsymbol{z}_i^T \boldsymbol{\gamma}) + \left[ \frac{\phi}{\phi + \exp(\boldsymbol{x}_i^T \boldsymbol{\beta})} \right]^{\phi} \right\} I_{(yi=0)} - \sum_{i=1}^{n} \log[1 + \exp(\boldsymbol{z}_i^T \boldsymbol{\gamma})] \\ &\quad + \sum_{i=1}^{n} \left\{ \log[\Gamma(\phi + y_i)] - \log[\Gamma(y_i + 1)] - \log[\Gamma(\phi)] + y_i \log \left[ \frac{\exp(\boldsymbol{x}_i^T \boldsymbol{\beta}}{\phi + \exp(\boldsymbol{x}_i^T \boldsymbol{\beta}} \right] \right. \\ &\quad \left. + \phi \log \left[ \frac{\phi}{\phi + \exp(\boldsymbol{x}_i^T \boldsymbol{\beta}} \right] \right\} I_{(yi>0)} \end{aligned} \tag{C.1}$$

Para as derivadas com respeito a $\phi$, a função $\left( \frac{\phi}{\phi + \exp(\boldsymbol{x}_i^T \boldsymbol{\beta})} \right)^{\phi}$ é do tipo $f(x)^x$, e a derivada desse tipo de função é $\frac{\partial f(x)^x}{\partial x} = f(x)^x \left[ \log(f(x)) + \frac{f'(x)x}{f(x)} \right]$.

$$\begin{aligned} \frac{\partial l(\boldsymbol{\theta})}{\partial \phi} &= \sum_{i;y=0} \frac{\left[ \frac{\phi}{\phi + \exp(\boldsymbol{x}_i^T \boldsymbol{\beta})} \right]^{\phi} \left[ \log \left( \frac{\phi}{\phi + \exp(\boldsymbol{x}_i^T \boldsymbol{\beta})} \right) + \phi \left( \frac{\phi}{\phi + \exp(\boldsymbol{x}_i^T \boldsymbol{\beta})} \right)^{-1} \left( \frac{\exp(\boldsymbol{x}_i^T \boldsymbol{\beta}) + \phi - \phi}{[\exp(\boldsymbol{x}_i^T \boldsymbol{\beta}) + \phi]^2} \right) \right]}{\exp(\boldsymbol{z}_i^T \boldsymbol{\gamma}) + \left[ \frac{\phi}{\phi + \exp(\boldsymbol{x}_i^T \boldsymbol{\beta})} \right]^{\phi}} \\ &\quad + \sum_{i;y>0} \left\{ \psi(\phi + y_i) - \psi(\phi) - \frac{(y_i + \phi)}{\exp(\boldsymbol{x}_i^T \boldsymbol{\beta}) + \phi} + \log \left( \frac{\phi}{\exp(\boldsymbol{x}_i^T \boldsymbol{\beta}) + \phi} \right) + \frac{\phi}{\phi} \right\} \end{aligned}$$

$$
\begin{aligned}
\frac{\partial l(\boldsymbol{\theta})}{\partial \phi} &= \sum_{i;y=0} \frac{\left[\frac{\phi}{\phi+\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})}\right]^{\phi}\left[\log\left(\frac{\phi}{\phi+\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})}\right)+\phi\left(\frac{\phi+\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})}{\phi}\right)\left(\frac{\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})}{[\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})+\phi]^2}\right)\right]}{\exp(\boldsymbol{z}_i^T\boldsymbol{\gamma})+\left[\frac{\phi}{\phi+\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})}\right]^{\phi}} \\
&+ \sum_{i;y>0}\left\{\psi(\phi+y_i)-\psi(\phi)-\frac{(y_i+\phi)}{\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})+\phi}+\log\left(\frac{\phi}{\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})+\phi}\right)\right\} \\
&= \sum_{i;y=0} \frac{\left[\frac{\phi}{\phi+\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})}\right]^{\phi}\left[\log\left(\frac{\phi}{\phi+\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})}\right)+\left(\frac{\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})}{\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})+\phi}\right)\right]}{\exp(\boldsymbol{z}_i^T\boldsymbol{\gamma})+\left[\frac{\phi}{\phi+\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})}\right]^{\phi}} \\
&+ \sum_{i;y>0}\left\{\psi(\phi+y_i)-\psi(\phi)-\frac{(y_i+\phi)}{\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})+\phi}+\log\left(\frac{\phi}{\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})+\phi}\right)\right\}
\end{aligned}
$$

$$
\begin{aligned}
\mathbf{L}_{\phi\phi} = \frac{\partial^2 l(\boldsymbol{\theta})}{\partial \phi^2} &= \sum_{i;y=0} - \frac{\left(\left[\frac{\phi}{\phi+\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})}\right]^{\phi}\left[\log\left(\frac{\phi}{\phi+\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})}\right)+\left(\frac{\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})}{\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})+\phi}\right)\right]\right)^2}{\left(\exp(\boldsymbol{z}_i^T\boldsymbol{\gamma})+\left[\frac{\phi}{\phi+\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})}\right]^{\phi}\right)^2} \\
&+ \sum_{i;y=0} \frac{\left(\left[\frac{\phi}{\phi+\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})}\right]^{\phi}\left[\log\left(\frac{\phi}{\phi+\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})}\right)+\left(\frac{\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})}{\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})+\phi}\right)\right]\right)^2}{\exp(\boldsymbol{z}_i^T\boldsymbol{\gamma})+\left[\frac{\phi}{\phi+\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})}\right]^{\phi}} \\
&+ \sum_{i;y=0} \frac{\left[\frac{\phi}{\phi+\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})}\right]^{\phi}\left[\frac{\phi+\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})}{\phi}\,\frac{\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})}{[\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})+\phi]^2}-\frac{\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})}{[\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})+\phi]^2}\right]}{\exp(\boldsymbol{z}_i^T\boldsymbol{\gamma})+\left[\frac{\phi}{\phi+\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})}\right]^{\phi}} \\
&+ \sum_{i;y>0}\left\{\psi'(\phi+y_i)-\psi'(\phi)+\frac{(y_i+\phi)}{[\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})+\phi]^2}+\frac{1}{\phi}-\frac{2}{\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})+\phi}\right\} \\
&= \sum_{i;y=0} \frac{\left(\left[\frac{\phi}{\phi+\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})}\right]^{\phi}\left[\log\left(\frac{\phi}{\phi+\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})}\right)+\left(\frac{\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})}{\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})+\phi}\right)\right]\right)^2}{\exp(\boldsymbol{z}_i^T\boldsymbol{\gamma})+\left[\frac{\phi}{\phi+\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})}\right]^{\phi}}\left(1-\frac{1}{\exp(\boldsymbol{z}_i^T\boldsymbol{\gamma})+\left[\frac{\phi}{\phi+\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})}\right]^{\phi}}\right) \\
&+ \sum_{i;y=0} \frac{\left[\frac{\phi}{\phi+\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})}\right]^{\phi}\left[\frac{\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})}{\phi+\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})}\left(\frac{1}{\phi}-\frac{1}{\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})+\phi}\right)\right]}{\exp(\boldsymbol{z}_i^T\boldsymbol{\gamma})+\left[\frac{\phi}{\phi+\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})}\right]^{\phi}} \\
&+ \sum_{i;y>0}\left\{\psi'(\phi+y_i)-\psi'(\phi)+\phi^{-1}+\frac{(y_i+\phi)}{[\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})+\phi]^2}-\frac{2}{\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})+\phi}\right\}
\end{aligned}
$$

$$\mathbf{L}_{\phi\phi} = \frac{\partial^2 l(\boldsymbol{\theta})}{\partial \phi^2} = \sum_{i;y=0} \frac{\left(\left[\frac{\phi}{\phi+\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})}\right]^{\phi}\left[\log\left(\frac{\phi}{\phi+\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})}\right)+\left(\frac{\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})}{\exp(\boldsymbol{z}_i^T\boldsymbol{\gamma})+\phi}\right)\right]\right)^2}{\exp(\boldsymbol{z}_i^T\boldsymbol{\gamma})+\left[\frac{\phi}{\phi+\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})}\right]^{\phi}} \times$$
$$\left(1-\frac{1}{\exp(\boldsymbol{z}_i^T\boldsymbol{\gamma})+\left[\frac{\phi}{\phi+\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})}\right]^{\phi}}\right) + \sum_{i;y=0}\frac{\left[\frac{\phi}{\phi+\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})}\right]^{\phi}\left[\phi^{-1}\left(\frac{\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})}{\phi+\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})}\right)^2\right]}{\exp(\boldsymbol{z}_i^T\boldsymbol{\gamma})+\left[\frac{\phi}{\phi+\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})}\right]^{\phi}}$$
$$+\sum_{i;y>0}\left\{\psi'(\phi+y_i)-\psi'(\phi)+\phi^{-1}+\frac{(y_i+\phi)}{[\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})+\phi]^2}-\frac{2}{\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})+\phi}\right\}$$

Usando a mesma nomenclatura de Garay et al. (2011), $\psi(k) = \frac{\partial \log[\Gamma(k)]}{\partial k}$, $\psi'(k) = \frac{\partial \psi(k)}{\partial k}$, $g_1(\mathbf{x}_i) = \frac{\phi}{\phi+\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})}$, $g_2(\mathbf{x}_i) = \frac{\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})}{\phi+\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})}$, $h(\mathbf{z}_i,\mathbf{x}_i) = \exp(\boldsymbol{z}_i^T\boldsymbol{\gamma}) + \left[\frac{\phi}{\phi+\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})}\right]^{\phi}$, $\mathbf{L}_{\phi\phi}$ pode ser reescrito como:

$$\mathbf{L}_{\phi\phi} = \frac{\partial^2 l(\boldsymbol{\theta})}{\partial \phi^2} = \sum_{i;y=0}\left\{\frac{\left([g_1(\mathbf{x}_i)]^{\phi}[\log(g_1(\mathbf{x}_i))+g_2(\mathbf{x}_i)]\right)^2}{h(\mathbf{z}_i,\mathbf{x}_i)}\left(1-\frac{1}{h(\mathbf{z}_i,\mathbf{x}_i)}\right)+\frac{[g_1(\mathbf{x}_i)]^{\phi}\left[\phi^{-1}[g_2(\mathbf{x}_i)]^2\right]}{h(\mathbf{z}_i,\mathbf{x}_i)}\right\}$$
$$+\sum_{i;y>0}\left\{\psi'(\phi+y_i)-\psi'(\phi)+\phi^{-1}+\frac{(y_i+\phi)}{[\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})+\phi]^2}-\frac{2}{\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})+\phi}\right\}$$

$$\mathbf{L}_{\phi\gamma_j} = \frac{\partial^2 l(\boldsymbol{\theta})}{\partial\phi\partial\gamma_j} = \sum_{i;y=0} -\frac{\left[\frac{\phi}{\phi+\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})}\right]^{\phi}\left[\log\left(\frac{\phi}{\phi+\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})}\right)+\left(\frac{\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})}{\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})+\phi}\right)\right]}{\left(\exp(\boldsymbol{z}_i^T\boldsymbol{\gamma})+\left[\frac{\phi}{\phi+\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})}\right]^{\phi}\right)^2}\exp(\boldsymbol{z}_i^T\boldsymbol{\gamma})z_{ij}$$
$$\mathbf{L}_{\phi\gamma_j} = \frac{\partial^2 l(\boldsymbol{\theta})}{\partial\phi\partial\gamma_j} = \sum_{i;y=0} -\frac{[g_1(\mathbf{x}_i)]^{\phi}\left[\log\left(g_1(\mathbf{x}_i)\right)+g_2(\mathbf{x}_i)\right]}{h(\mathbf{z}_i,\mathbf{x}_i)^2}\exp(\boldsymbol{z}_i^T\boldsymbol{\gamma})z_{ij}$$

$$
\begin{aligned}
\mathbf{L}_{\phi\beta} = \frac{\partial^2 l(\boldsymbol{\theta})}{\partial\phi\partial\beta_j} \quad = \quad & \sum_{i;y=0} - \frac{\left[\frac{\phi}{\phi+\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})}\right]^{\phi}\left[\log\left(\frac{\phi}{\phi+\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})}\right)+\left(\frac{\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})}{\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})+\phi}\right)\right]\phi\left(\frac{\phi}{\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})+\phi}\right)^{\phi-1}\frac{(-\phi)}{[\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})+\phi]^2}\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})x_{ij}}{\left(\exp(\boldsymbol{z}_i^T\boldsymbol{\gamma})+\left[\frac{\phi}{\phi+\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})}\right]^{\phi}\right)^2} \\
& - \sum_{i;y=0}\frac{\frac{\phi}{[\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})+\phi]^2}\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})x_{ij}\phi\left(\frac{\phi}{\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})+\phi}\right)^{\phi-1}\left[\log\left(\frac{\phi}{\phi+\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})}\right)+\left(\frac{\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})}{\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})+\phi}\right)\right]}{\exp(\boldsymbol{z}_i^T\boldsymbol{\gamma})+\left[\frac{\phi}{\phi+\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})}\right]^{\phi}} \\
& + \sum_{i;y=0}\frac{\left(\frac{\phi}{\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})+\phi}\right)^{\phi}\left[\left(\frac{\phi+\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})}{\phi}\right)\frac{-\phi}{[\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})+\phi]^2}\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})x_{ij}+\frac{\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})x_{ij}(\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})+\phi)-[\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})]^2x_{ij}}{[\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})+\phi]^2}\right]}{\exp(\boldsymbol{z}_i^T\boldsymbol{\gamma})+\left[\frac{\phi}{\phi+\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})}\right]^{\phi}} \\
& + \sum_{i;y>0}\left\{\frac{(y_i+\phi)}{[\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})+\phi]^2}\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})x_{ij}-\frac{(\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})+\phi)}{\phi}\frac{\phi\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})x_{ij}}{[\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})+\phi]^2}\right\} \\
= \quad & \sum_{i;y=0}\frac{\left[\frac{\phi}{\phi+\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})}\right]^{2\phi}\frac{\phi^2}{[\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})+\phi]^2}\frac{(\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})+\phi)}{\phi}\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})x_{ij}\left[\log\left(\frac{\phi}{\phi+\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})}\right)+\left(\frac{\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})}{\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})+\phi}\right)\right]}{\left(\exp(\boldsymbol{z}_i^T\boldsymbol{\gamma})+\left[\frac{\phi}{\phi+\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})}\right]^{\phi}\right)^2} \\
& - \sum_{i;y=0}\frac{\frac{\phi^2}{[\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})+\phi]^2}\frac{(\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})+\phi)}{\phi}\left(\frac{\phi}{\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})+\phi}\right)^{\phi}\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})x_{ij}\left[\log\left(\frac{\phi}{\phi+\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})}\right)+\left(\frac{\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})}{\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})+\phi}\right)\right]}{\exp(\boldsymbol{z}_i^T\boldsymbol{\gamma})+\left[\frac{\phi}{\phi+\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})}\right]^{\phi}} \\
& + \sum_{i;y=0}\frac{\left(\frac{\phi}{\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})+\phi}\right)^{\phi}\left[-\left(\frac{\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})}{\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})+\phi}\right)x_{ij}+\left(\frac{\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})}{\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})+\phi}\right)x_{ij}-\left(\frac{\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})}{\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})+\phi}\right)^2x_{ij}\right]}{\exp(\boldsymbol{z}_i^T\boldsymbol{\gamma})+\left[\frac{\phi}{\phi+\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})}\right]^{\phi}} \\
& + \sum_{i;y>0}\left\{\frac{\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})x_{ij}}{(\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})+\phi)}\left(\frac{(y_i+\phi)}{(\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})+\phi)}-1\right)\right\} \\
= \quad & \sum_{i;y=0}\frac{\left[\frac{\phi}{\phi+\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})}\right]^{2\phi+1}\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})x_{ij}\left[\log\left(\frac{\phi}{\phi+\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})}\right)+\left(\frac{\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})}{\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})+\phi}\right)\right]}{\left(\exp(\boldsymbol{z}_i^T\boldsymbol{\gamma})+\left[\frac{\phi}{\phi+\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})}\right]^{\phi}\right)^2} \\
& - \sum_{i;y=0}\frac{\left(\frac{\phi}{\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})+\phi}\right)^{\phi+1}\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})x_{ij}\left[\log\left(\frac{\phi}{\phi+\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})}\right)+\left(\frac{\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})}{\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})+\phi}\right)\right]-\left(\frac{\phi}{\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})+\phi}\right)^{\phi}\left(\frac{\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})}{\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})+\phi}\right)^2x_{ij}}{\exp(\boldsymbol{z}_i^T\boldsymbol{\gamma})+\left[\frac{\phi}{\phi+\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})}\right]^{\phi}} \\
& + \sum_{i;y>0}\left\{\frac{\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})x_{ij}}{(\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})+\phi)}\left(\frac{y_i-\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})}{\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})+\phi}\right)\right\}
\end{aligned}
$$

$$\mathbf{L}_{\phi\beta_j} = \frac{\partial^2 l(\boldsymbol{\theta})}{\partial\phi\partial\beta_j} = \sum_{i;y=0} \frac{\left[\frac{\phi}{\phi+\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})}\right]^{2\phi+1}\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})x_{ij}\left[\log\left(\frac{\phi}{\phi+\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})}\right)+\left(\frac{\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})}{\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})+\phi}\right)\right]}{\left(\exp(\boldsymbol{z}_i^T\boldsymbol{\gamma})+\left[\frac{\phi}{\phi+\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})}\right]^{\phi}\right)^2}$$

$$-\sum_{i;y=0} \frac{\left(\frac{\phi}{\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})+\phi}\right)^{\phi}x_{ij}\left(\frac{\phi(\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})}{\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})+\phi}\left[\log\left(\frac{\phi}{\phi+\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})}\right)+\left(\frac{\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})}{\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})+\phi}\right)\right]-\left(\frac{\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})}{\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})+\phi}\right)^2\right)}{\exp(\boldsymbol{z}_i^T\boldsymbol{\gamma})+\left[\frac{\phi}{\phi+\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})}\right]^{\phi}}$$

$$+\sum_{i;y>0}\left\{\frac{\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})x_{ij}}{(\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})+\phi)}\left(\frac{y_i-\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})}{\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})+\phi}\right)\right\}$$

$$\mathbf{L}_{\phi\beta_j} = \frac{\partial^2 l(\boldsymbol{\theta})}{\partial\phi\partial\beta_j} = \sum_{i;y=0}\frac{[g_1(\mathbf{x}_i)]^{2\phi+1}\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})x_{ij}\left[\log\left(g_1(\mathbf{x}_i)\right)+g_2(\mathbf{x}_i)\right]}{h(\mathbf{z}_i,\mathbf{x}_i)^2}$$

$$-\sum_{i;y=0}\frac{[g_1(\mathbf{x}_i)]^{\phi}x_{ij}\left(\phi g_2(\mathbf{x}_i)\left[\log\left(g_1(\mathbf{x}_i)\right)+g_2(\mathbf{x}_i)\right]-[g_2(\mathbf{x}_i)]^2\right)}{h(\mathbf{z}_i,\mathbf{x}_i)}$$

$$+\sum_{i;y>0}\left\{g_2(\mathbf{x}_i)x_{ij}\left(\frac{y_i-\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})}{\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})+\phi}\right)\right\}$$

$$\frac{\partial l(\boldsymbol{\theta})}{\partial\gamma_j} = \sum_{i;y=0}\frac{\exp(\boldsymbol{z}_i^T\boldsymbol{\gamma})z_{ij}}{\left(\exp(\boldsymbol{z}_i^T\boldsymbol{\gamma})+\left[\frac{\phi}{\phi+\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})}\right]^{\phi}\right)}-\sum_{i=1}^{n}\frac{\exp(\boldsymbol{z}_i^T\boldsymbol{\gamma})z_{ij}}{1+\exp(\boldsymbol{z}_i^T\boldsymbol{\gamma})}$$

$$\mathbf{L}_{\gamma_j\gamma_k} = \frac{\partial^2 l(\boldsymbol{\theta})}{\partial\gamma_j\partial\gamma_k} = \sum_{i;y=0}-\frac{[\exp(\boldsymbol{z}_i^T\boldsymbol{\gamma})]^2 z_{ij}z_{ik}}{\left(\exp(\boldsymbol{z}_i^T\boldsymbol{\gamma})+\left[\frac{\phi}{\phi+\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})}\right]^{\phi}\right)^2}+\frac{\exp(\boldsymbol{z}_i^T\boldsymbol{\gamma})z_{ij}z_{ik}}{\exp(\boldsymbol{z}_i^T\boldsymbol{\gamma})+\left[\frac{\phi}{\phi+\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})}\right]^{\phi}}$$

$$+\sum_{i=1}^{n}\frac{[\exp(\boldsymbol{z}_i^T\boldsymbol{\gamma})]^2 z_{ij}z_{ik}}{[1+\exp(\boldsymbol{z}_i^T\boldsymbol{\gamma})]^2}-\frac{\exp(\boldsymbol{z}_i^T\boldsymbol{\gamma})z_{ij}z_{ik}}{1+\exp(\boldsymbol{z}_i^T\boldsymbol{\gamma})}$$

$$= \sum_{i;y=0}\frac{-[\exp(\boldsymbol{z}_i^T\boldsymbol{\gamma})]^2 z_{ij}z_{ik}+[\exp(\boldsymbol{z}_i^T\boldsymbol{\gamma})]^2 z_{ij}z_{ik}+z_{ij}z_{ik}\exp(\boldsymbol{z}_i^T\boldsymbol{\gamma})\left[\frac{\phi}{\phi+\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})}\right]^{\phi}}{\left(\exp(\boldsymbol{z}_i^T\boldsymbol{\gamma})+\left[\frac{\phi}{\phi+\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})}\right]^{\phi}\right)^2}$$

$$+\sum_{i=1}^{n}\frac{[\exp(\boldsymbol{z}_i^T\boldsymbol{\gamma})]^2 z_{ij}z_{ik}-\exp(\boldsymbol{z}_i^T\boldsymbol{\gamma})z_{ij}z_{ik}-[\exp(\boldsymbol{z}_i^T\boldsymbol{\gamma})]^2 z_{ij}z_{ik}}{[1+\exp(\boldsymbol{z}_i^T\boldsymbol{\gamma})]^2}$$

$$= \sum_{i;y=0}\frac{z_{ij}z_{ik}\exp(\boldsymbol{z}_i^T\boldsymbol{\gamma})\left[\frac{\phi}{\phi+\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})}\right]^{\phi}}{\left(\exp(\boldsymbol{z}_i^T\boldsymbol{\gamma})+\left[\frac{\phi}{\phi+\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})}\right]^{\phi}\right)^2}-\sum_{i=1}^{n}\frac{\exp(\boldsymbol{z}_i^T\boldsymbol{\gamma})z_{ij}z_{ik}}{[1+\exp(\boldsymbol{z}_i^T\boldsymbol{\gamma})]^2}$$

$$\mathbf{L}_{\gamma_j\gamma_k} = \frac{\partial^2 l(\boldsymbol{\theta})}{\partial\gamma_j\partial\gamma_k} = \sum_{i;y=0}\frac{z_{ij}z_{ik}\exp(\boldsymbol{z}_i^T\boldsymbol{\gamma})\,[g_1(\mathbf{x}_i)]^{\phi}}{h(\mathbf{z}_i\mathbf{x}_i)^2}-\sum_{i=1}^{n}\frac{\exp(\boldsymbol{z}_i^T\boldsymbol{\gamma})z_{ij}z_{ik}}{[1+\exp(\boldsymbol{z}_i^T\boldsymbol{\gamma})]^2}$$

$$\mathbf{L}_{\gamma_j\beta_k} = \frac{\partial^2 l(\boldsymbol{\theta})}{\partial\gamma_j\partial\beta_k} = \sum_{i;y=0} \frac{\exp(\boldsymbol{z}_i^T\boldsymbol{\gamma})z_{ij}\phi\left[\frac{\phi}{\phi+\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})}\right]^{\phi}\frac{\phi+\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})}{\phi}\frac{\phi}{[\phi+\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})]^2}\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})x_{ik}}{\left(\exp(\boldsymbol{z}_i^T\boldsymbol{\gamma})+\left[\frac{\phi}{\phi+\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})}\right]^{\phi}\right)^2}$$

$$= \sum_{i;y=0} \frac{\phi z_{ij}x_{ik}\exp(\boldsymbol{z}_i^T\boldsymbol{\gamma})\left[\frac{\phi}{\phi+\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})}\right]^{\phi}\frac{\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})}{\phi+\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})}}{\left(\exp(\boldsymbol{z}_i^T\boldsymbol{\gamma})+\left[\frac{\phi}{\phi+\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})}\right]^{\phi}\right)^2}$$

$$\mathbf{L}_{\gamma_j\beta_k} = \frac{\partial^2 l(\boldsymbol{\theta})}{\partial\gamma_j\partial\beta_k} = \sum_{i;y=0} \frac{\phi z_{ij}x_{ik}\exp(\boldsymbol{z}_i^T\boldsymbol{\gamma})\,[g_1(\mathbf{x}_i)]^{\phi}\,g_2(\mathbf{x}_i)}{h(\mathbf{z}_i\mathbf{x}_i)^2}$$

$$\frac{\partial l(\boldsymbol{\theta})}{\partial\beta_j} = \sum_{i;y=0}\frac{-\phi\left[\frac{\phi}{\phi+\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})}\right]^{\phi-1}\frac{\phi}{[\phi+\exp(\boldsymbol{x}_i^T\boldsymbol{\beta}))]^2}\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})x_{ij}}{\exp(\boldsymbol{z}_i^T\boldsymbol{\gamma})+\left[\frac{\phi}{\phi+\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})}\right]^{\phi}} + \sum_{i;y>0}\phi\frac{(\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})+\phi)}{\phi}\frac{(-\phi)\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})x_{ij}}{[\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})+\phi]^2}$$

$$+\sum_{i;y>0}y_i\frac{\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})+\phi}{\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})}\left(\frac{\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})x_{ij}}{\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})+\phi}-\frac{\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})x_{ij}}{[\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})+\phi]^2}\right)$$

$$= \sum_{i;y=0}\frac{-\left[\frac{\phi}{\phi+\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})}\right]^{\phi}\frac{\phi^2}{[\phi+\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})]^2}\frac{\phi+\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})}{\phi}\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})x_{ij}}{\exp(\boldsymbol{z}_i^T\boldsymbol{\gamma})+\left[\frac{\phi}{\phi+\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})}\right]^{\phi}} - \sum_{i;y>0}\frac{\phi\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})x_{ij}}{\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})+\phi}$$

$$+\sum_{i;y>0}y_i\left(x_{ij}-\frac{\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})x_{ij}}{\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})+\phi}\right)$$

$$= \sum_{i;y=0}\frac{-\phi\left[\frac{\phi}{\phi+\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})}\right]^{\phi}\frac{\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})}{\phi+\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})}x_{ij}}{\exp(\boldsymbol{z}_i^T\boldsymbol{\gamma})+\left[\frac{\phi}{\phi+\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})}\right]^{\phi}} + \sum_{i;y>0}\frac{\phi x_{ij}(y_i-\exp(\boldsymbol{x}_i^T\boldsymbol{\beta}))}{\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})+\phi}$$

$$
\begin{aligned}
\mathbf{L}_{\beta_j\beta_k} = \frac{\partial^2 l(\boldsymbol{\theta})}{\partial\beta_j\partial\beta_k} &= \sum_{i;y=0} -\frac{\phi^2\left(\left[\frac{\phi}{\phi+\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})}\right]^{\phi}\left[\frac{\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})}{\phi+\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})}\right]\right)^2 x_{ij}x_{ik}}{\left(\exp(\boldsymbol{z}_i^T\boldsymbol{\gamma})+\left[\frac{\phi}{\phi+\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})}\right]^{\phi}\right)^2} \\
&+ \sum_{i;y=0} \frac{\phi^2\left[\frac{\phi}{\phi+\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})}\right]^{\phi}\frac{\phi+\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})}{\phi}\frac{[\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})]^2}{\phi+\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})}x_{ij}x_{ik}\frac{\phi}{[\phi+\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})]^2}}{\exp(\boldsymbol{z}_i^T\boldsymbol{\gamma})+\left[\frac{\phi}{\phi+\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})}\right]^{\phi}} \\
&- \sum_{i;y=0} \frac{\phi\left[\frac{\phi}{\phi+\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})}\right]^{\phi}\left(\frac{\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})x_{ij}x_{ik}}{\phi+\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})}-\frac{[\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})]^2x_{ij}x_{ik}}{[\phi+\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})]^2}\right)}{\exp(\boldsymbol{z}_i^T\boldsymbol{\gamma})+\left[\frac{\phi}{\phi+\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})}\right]^{\phi}} \\
&+ \sum_{i;y>0}\left\{\frac{-\phi x_{ij}(y_i-\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})x_{ik}}{[\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})+\phi]^2}-\frac{\phi x_{ij}x_{ik}\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})}{\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})+\phi}\right\} \\
&= \sum_{i;y=0} -\frac{\phi^2\left(\left[\frac{\phi}{\phi+\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})}\right]^{\phi}\left[\frac{\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})}{\phi+\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})}\right]\right)^2 x_{ij}x_{ik}}{\left(\exp(\boldsymbol{z}_i^T\boldsymbol{\gamma})+\left[\frac{\phi}{\phi+\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})}\right]^{\phi}\right)^2} \\
&+ \sum_{i;y=0} \frac{\phi^2\left[\frac{\phi}{\phi+\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})}\right]^{\phi}\frac{[\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})]^2}{[\phi+\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})]^2}x_{ij}x_{ik}\left(1-\frac{1}{\phi}\frac{[\phi+\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})]}{\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})}+\frac{1}{\phi}\right)}{\exp(\boldsymbol{z}_i^T\boldsymbol{\gamma})+\left[\frac{\phi}{\phi+\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})}\right]^{\phi}} \\
&- \sum_{i;y>0}\left\{\frac{\phi x_{ij}x_{ik}\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})}{\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})+\phi}\left[\frac{y_i-\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})}{\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})+\phi}+1\right]\right\} \\
&= \sum_{i;y=0} -\frac{\phi^2\left(\left[\frac{\phi}{\phi+\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})}\right]^{\phi}\left[\frac{\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})}{\phi+\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})}\right]\right)^2 x_{ij}x_{ik}}{\left(\exp(\boldsymbol{z}_i^T\boldsymbol{\gamma})+\left[\frac{\phi}{\phi+\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})}\right]^{\phi}\right)^2} \\
&+ \sum_{i;y=0} \frac{\phi^2\left[\frac{\phi}{\phi+\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})}\right]^{\phi}\frac{[\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})]^2}{[\phi+\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})]^2}x_{ij}x_{ik}\left(1-\frac{1}{\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})}\right)}{\exp(\boldsymbol{z}_i^T\boldsymbol{\gamma})+\left[\frac{\phi}{\phi+\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})}\right]^{\phi}} \\
&- \sum_{i;y>0}\left\{\frac{\phi x_{ij}x_{ik}\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})}{\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})+\phi}\left[\frac{y_i-\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})}{\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})+\phi}+1\right]\right\}
\end{aligned}
$$

$$
\begin{aligned}
\mathbf{L}_{\beta_j\beta_k} = \frac{\partial^2 l(\boldsymbol{\theta})}{\partial\beta_j\partial\beta_k} &= \sum_{i;y=0} -\frac{\phi^2\left([g_1(\mathbf{x}_i)]^{\phi}\,g_2(\mathbf{x}_i)\right)^2 x_{ij}x_{ik}}{h(\mathbf{z}_i\mathbf{x}_i)^2}+\frac{\phi^2\,[g_1(\mathbf{x}_i)]^{\phi}\,[g_2(\mathbf{x}_i)]^2x_{ij}x_{ik}\left(1-\frac{1}{\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})}\right)}{h(\mathbf{z}_i\mathbf{x}_i)} \\
&- \sum_{i;y>0}\phi x_{ij}x_{ik}g_2(\mathbf{x}_i)\left[\frac{y_i-\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})}{\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})+\phi}+1\right]
\end{aligned}
$$

# Referências Bibliográficas


Agresti, A. (2003). *Categorical Data Analysis*, (2nd ed.). Wiley.

Atkinson, P. M., German, S. E., Sear, D. A., e Clark, M. J. (2003). Exploring the relations between riverbank erosion and geomorphological controls using geographically weighted logistic regression. *Geographical Analysis*, 35(1):58–82.

Brunsdon, C., Fotheringham, A. S., e Charlton, M. E. (1996). Geographically weighted regression: a method for exploring spatial nonstationarity. *Geographical Analysis*, 28(4):281–298.

Brunsdon, C., Fotheringham, A. S., e Charlton, M. E. (1998). Geographically weighted regression - modelling spatial non-stationarity. *The Statistician*, 47(3):431–443.

Brunsdon, C., Fotheringham, A. S., e Charlton, M. E. (2000). Quantitative geography – perspectives on spatial data analysis.

Cameron, A. e Windmeijer, F. (1997). An R-squared measure of goodness of fit for some common nonlinear regression models. *Econometrics*, 77:329–342.

Casella, G. e Berger, R. L. (2014). *Statistical Inference*, (2nd ed.). Cengage Learning.

Cleveland, W. S. (1979). Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association*, AC(74):829–836.

Cohen, A. C. (1963). *Estimation in Mixtures of Discrete Distributions. International Symposium on Discrete Distributions*. Montreal: International Symposium on Discrete Distributions.

Conte, S. D. (1965). *Elementary Numerical Analysis*. MacGraw-Hill.

Da Silva, A. R. e Fotheringham, A. S. (2016). The multiple testing issue in geographically weighted regression. *Geographical Analysis*, (48):233–247.

Da Silva, A. R. e Mendes, F. F. (2018). On comparing some algorithms for finding the optimal bandwidth in geographically weighted regression. *Applied Soft Computing Journal*, 73:943–957.

Da Silva, A. R. e Rodrigues, T. C. V. (2014). Geographically weighted negative binomial regression - incorporating overdispersion. *Statistics and Computing*, 24(5):769–783.

Dempster, A. P., Laird, N. M., e Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society*, 39(1):1–38.

Dempster, A. P., Laird, N. M., e Rubin, D. B. (2009). Geographically weighted regression. *White paper. National Centre for Geocomputation. National University of Ireland Maynooth*.

Dobson, A. J. e Barnett, A. G. (2008). *An introduction to generalized linear models*. 3rd ed. Chapman and Hall/CRC.

Druck, S., Carvalho, M., Câmara, G., e Monteiro, A. (2004). *Análise espacial de dados geográficos*. EMBRAPA.

Erdman, D., Jackson, L., e Sinko, A. (2008). Zero-inflated poisson and zero-inflated negative binomial models using the countreg procedure. Technical report, SAS Global Forum 2008. SAS Global Forum 2008.

Fotheringham, A. S., Charlton, M., e Brunsdon, C. (2002). *Geographically Weighted Regression: The Analysis of Spatially Varying Relationships*. Wiley.

Fumes, G. (2009). Uso de modelos inflacionados de zeros na análise de questionários de frequência alimentar. Master's thesis, Universidade Estadual Paulista Júlio de Mesquita Filho.

Garay, A. M., Hashimoto, E. M., Ortega, E. M. M., e Lachos, V. H. (2011). On estimation and influence diagnostics for zero-inflated negative binomial regression models. *Computational Statistics & Data Analysis*, 55(3):1304–1318.

Hall, D. B. (2000). Zero-inflated poisson and binomial regression with random effects: A case study. *Biometrics*, 56:1030–1039.

Hilbe, J. M. (2011). *Negative Binomial Regression*. Cambridge University Press.

Hinde, J. e Demétrio, C. G. B. (1998). Overdispersion: Models and estimation. *Computational Statistics and Data Analysis*, 27:151–170.

Hurvich, C. M. e Tsai, C. L. (1989). Regression and time series model selection in small samples. *Biometrika*, 76:297–307.

Johnson, N. e Kotz, S. (1969). Distributions in statistics: Discrete distributions. *Boston: Houghton Mifflin, Wiley/Houghton-Mifflin*, page 328.

Kalagirou, S. (2016). Destination choice of athenians: An application of geographically weighted versions of standard and zero inflated poisson spatial interaction models. *Geographical Analysis*, 48:191–230.

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

Lawless, J. F. (1987). Negative binomial and mixed poisson regression. *The Canadian Journal of Statistics*, 15(3):209–225.

Leung, Y., Chang-Lin, M., e Wen-Xiu, Z. (2000). Statistical tests for spatial nonstationarity based on the geographically weighted regression model. *Environment and Planning A*, 32(1):9–32.

Martin, J. e Hall, D. B. (2016). $R^2$ measures for zero-inflated regression models for count data with excess zeros. *Journal of Statistical Computation and Simulation*, 86(18):3777–3790.

Mittlböck, M. e Waldhör, T. (2000). Adjustments for $R^2$-measures for poisson regression models. *Comput Statist Data Anal*, 34:461–472.

Nakaya, T., Fotheringham, A., Brunsdon, C., e Charlton, M. (2005). Geographically weighted poisson regression for disease association mapping. *Statistics in Medicine*, 24(17):2695–2717.

Naya, H., Urioste, J. I., Chang, Y. M., Rodrigues-Motta, M., Kremer, R., e Gionola, D. (2008). A comparison between poisson and zero-inflated poisson regression models with an application to number of black spots in corriedale sheep. *Genet Sel Evol.*, 40(2):379–394.

Nelder, J. A. e Wedderburn, R. W. M. (1972). Generalized linear models. *Journal of the Royal Statistical Society. Series A (General)*, 135(3):370–384.

Neter, J., Wasserman, W., e Kutner, M. H. (1983). *Applied Linear Regression Models*. Richard D. Irwin, Inc. Homewood, Illinois.

Paula, G. A. (2013). *Modelos de regressão com apoio computacional*. IME-USP,São Paulo.

Purhadi, Dewi, N. S., Qurotul, A., e Irhamah (2021). Geographically weighted bivariate zero inflated generalized poisson regression model and its application. *Heliyon*, 7(7):e07491.

Purhadi, Yuliani, S. D., e Luthfatul, A. (2015). Zero inflated poisson and geographically weighted zeroinflated poisson regression model: Application to elephantiasis (filariasis) counts data. *Journal of Mathematics and Statistics*, 11(2):52–60.

Ridout, M., Demétrio, C. G. B., e Hinde, J. (1998). *Models for count data with many zeros*. International Biometric Conference.

SAS (2011). Cary, NC: SAS Institute Inc. `v9doc.sas.com`. [SAS On Line Doc Version 9.3, SURVEYREG Procedure].

SINGH, S. A. (1963). A note on zero inflated poisson distribution. *Journal of the Indian Statistical Association*, 1(1):140–144.

Staniswalis, J. G. (1989). The kernel estimate of a regression function in likellihood based models. *Journal of the American Statistical Association*, 84:276–283.

Wang, Z., Shuangge, M., Gao, W., e Wang, C. Y. (2015). Variable selection for zero-inflated and overdispersed data with application to health care demand in germany. *Biom J.*, 57(5):867–884.

Washington, D. (2015). Map of South Korea. `https://www.loc.gov/item/2015587021/`. Acesso em 3 Jun, 2022.

Weinstein, B., da Silva, A. R., Kouzoukas, D. E., Bose, T., Kim, G. J., Correa, P. A., Pondugula, S., Lee, Y., Kim, J., e Carpenter, D. O. (2021). Precision mapping of covid-19 vulnerable locales by epidemiological and socioeconomic risk factors, developed using south korean data. *International Journal of Environmental Research and Public Health*, 604(18):1–14.

Wikipédia (2020). COVID-19 pandemic in mainland China — Wikipédia, a enciclopédia livre. https://en.wikipedia.org/wiki/COVID-19_pandemic_in_mainland_China. Acesso em 3 Jun, 2022.

Wikipédia (2021). COVID-19 pandemic in South Korea — Wikipédia, a enciclopédia livre. https://en.wikipedia.org/wiki/COVID-19_pandemic_in_South_Korea#:~:text=The%20first%20case%20in%20South%20Korea%20was%20announced%20on%2020%20January%202020. Acesso em 3 Jun, 2022.

Yau, K. K. W., Wang, K., e Lee, A. H. (2003). Zero-inflated negative binomial mixed regression modeling of over-dispersed count data with extra zeros. *Biometrical Journal*, 45(4):437–452.