Revista da Estatística UFOP, Vol III(3), 2014, ISSN 2237-8111
Edição especial: 59ª Reunião Anual da Regional Brasileira da Sociedade Internacional de Biometria - RBRAS
Trabalhos apresentados na versão poster.

Detecção de Clusters Espaciais Em Modelos de Regressão Poisson Duplo com Inflação de Zeros

José Cardoso Neto[a], Luiz H. Duczmal[b], Max S. de Lima[a] and Letícia P. Pinto[b]

[a] *Universidade Federal do Amazonas,* [b] *Universidade Federal de Minas Gerais*

1. Introdução

Recentemente a Estatística Scan (Kulldorff 1997) têm sido extendida para acomodar ajuste por covariáveis (Loh, 2007; Jung, 2009; Zhang and Lin, 2009) usando a regressão de Poisson. No entanto, na presença simultânea da sobredispersão e inflação de zeros, esta Estatística apresenta um excessivo erro tipo I (Lima et all, 2013). Sem ajuste por covariáveis, Cançado et all (2014) propôs uma Estatística Scan para o modelo Poisson com excesso de zeros **ZIP**, Zhang et all (2012) desenvolveram esta estatística em modelos de mistura Poisson-Gamma para dados com sobredispersão e uma Estatística Scan simultânea para modelos com excesso de zeros e sobredispersão foi proposta por Lima et al, (2013).

Geralmente o ajuste por covariável tem sido realizado somente na média do modelo. No entanto, se uma covariável está relacionada com a não ocorrência da doença e sua distribuição geográfica não é aleatória, pode-se detectar clusters espaciais de regiões com pequenas taxas que não são interessantes e sua interpretação pode ser errônea. Esses clusters podem ser explicados simplesmente pela distribuição espacial destas covariáveis. Por exemplo, o não registro de casos de câncer de mama em cidades onde não existe mamógrafos. Por isso, Neste artigo, uma modelagem via Regressão de Poisson Duplo em dados com execesso e sobredispersão é proposta para construir uma nova Estatística Scan onde todos os parâmetros são ajustados por covariáveis. A estimação é realizada via algortimo **EM**, a estatística de teste é logarítmo da Razão de Verossimilhança e a significância é avaliada usando o valor-p Bootstrap-**EM**. Uma ilustração do método é feita usando dados de casos de Haseníase no Estado do Amazonas-Brasil.

autor correspondente: J., Cardoso Neto(jcardoso@ufam.edu.br)

2. Estatística Scan para o Modelo de Regressão ZIDP(μ_l, ϕ_l, p_l)

2.1 O Modelo de Regress˜ao ZIDP(μ_l, ϕ_l, p_l)

Suponha que existam L localizações s_l e seja $\mathbf{Y} = (Y(s_1), ..., Y(s_L))'$, onde $Y_l \equiv Y(s_l)$ é a variável aleatória que representa a contagem de casos de uma determinada doença em s_l com população em risco n_l e valor observado y_l. O modelo proposto neste artigo é o de Regressão Poisson Duplo Inflacionado de Zeros, denotado por $\mathbf{ZIDP}(\mu_l, \phi_l, p_l), l = 1, 2, ..., L$, o qual assume

$$P(Y_l = y_l|p_l, \mu_l, \phi_l) = \begin{cases} p_l + (1-p_l)f_{DP}(0|\mu_l, \phi_l) & y_l = 0 \\ (1-p_l)f_{DP}(y_l|\mu_l, \phi_l) & y_l = 1, 2, ... \end{cases} \tag{2.1}$$

onde $f_{DP}(.|\mu_l, \phi_l)$ denota a função de probabilidade da distribuição Poisson Dupla aproximada por (veja Efron, 1986),

$$f_{DP}(y_l|\mu_l, \phi_l) = \phi_l^{1/2} \times \text{Poisson}(y_l) \times exp\left\{-\frac{\phi_l}{2} D_l(y_l, \mu_l)\right\} \tag{2.2}$$

em que $D_l(y_l, \mu_l)$ é a função desvio do modelo Poisson(μ_l). Os parâmetros $\boldsymbol{p} = (p_1, ..., p_L)'$, $\boldsymbol{\mu} = (\mu_1, ..., \mu_L)'$ e $\boldsymbol{\phi} = (\phi_1, ..., \phi_L)'$ são modelados através das funções de ligação $logit(p_l) = \mathbf{G}_l\boldsymbol{\gamma}$, $log(\mu_l) = \mathbf{B}_l\boldsymbol{\beta}$ e $logit(\boldsymbol{\phi}_l) = \mathbf{H}_l\boldsymbol{\delta}$. Onde, $(\boldsymbol{\beta}, \boldsymbol{\delta}, \boldsymbol{\gamma})$ são os parâmetros da regressão, ($\mathbf{B}$,$\mathbf{H}$,$\mathbf{G}$) são as correspondentes matrices do modelo e não, necessariamente, precisam ser distintas. Neste modelo, um zero ocorre com probabilidade p_l enquanto a outra parte envolve uma distribuição Poisson Duplo, $\mathbf{DP}(\mu_l, \phi_l)$, com probabilidade $1 - p_l$.

2.2 Estatística de Teste

Suponha que os dados $\mathbf{Y} = (Y(s_1), ..., Y(s_L))'$ são modelados por um modelo de regressão $\mathbf{ZIDP}(\mu_i, \phi_i, p_i)$. Assuma que neste modelo o $log(\mu_l) = \mathbf{B}_l\boldsymbol{\beta} + log\epsilon$ se $s_l \in Z$, e $log(\mu_l) = \mathbf{B}_l\boldsymbol{\beta}$ se $s_l \notin Z$. Onde Z é um conjunto de localizações espaciais candidato a cluster. Para detecção do cluster usamos o teste de hipóteses, $H_0 : \epsilon = 1$, para toda zona $Z \in \mathcal{Z}$ contra $H_1 : \epsilon > 1$ para alguma zona $Z \in \mathcal{Z}$. Em que $\mathcal{Z}$ é o conjunto de todos os possíveis candidatos a cluster. O parâmetro ϵ captura o risco relativo ajustado por covariável para indivíduos dentro da zona Z comparado com os indivíduos fora de Z. O interesse é a detecção de clusters com número de casos observados, significativamente, acima do esperado e procedemos da seguinte forma: Seja $\hat{\boldsymbol{\theta}}_0 = (\hat{\boldsymbol{\beta}}_0, \hat{\boldsymbol{\delta}}_0, \hat{\boldsymbol{\gamma}}_0)$ o Estimador de Máxima Verossimilhança para os parâmetros da regressão sob a hipótese nula e $\hat{\epsilon}$ o estimador de Máxima Verossimilhança de ϵ sob a hipótese alternativa

tal que sob H_0 teremos, $logit(\hat{\boldsymbol{p}}_0) = \mathbf{G}\hat{\boldsymbol{\gamma}}_0$, $log(\hat{\boldsymbol{\mu}}_0) = \mathbf{B}\hat{\boldsymbol{\beta}}_0$ e $logit(\hat{\boldsymbol{\phi}}_0) = \mathbf{H}\hat{\boldsymbol{\delta}}_0$. Sob H_1, o vetor de coeficientes $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\delta}, \boldsymbol{\gamma})$ é fixado usando as estimativas sob o modelo nulo. Ou seja, sob o modelo alternativo com mudança na média teremos, $\hat{\mu}_l = exp\left\{\mathbf{B}_l\hat{\boldsymbol{\beta}}_0 + log\hat{\epsilon}\right\} = \hat{\epsilon}e^{\mathbf{B}_l\hat{\boldsymbol{\beta}}_0} \quad \forall s_l \in Z$ e $\hat{\mu}_l = e^{\mathbf{B}_l\hat{\boldsymbol{\beta}}_0} \quad \forall s_l \notin Z$. Neste caso os coeficiente das covariáveis que são usadas para ajuste da média, inflação de zeros e sobredispersão, permanecem iguais mesmo para a varredura espacial de conjuntos distintos. A estatística de teste utilizada é $\mathcal{D} = \underset{Z\in\mathcal{Z}}{max}\mathcal{D}_Z$. Onde, $\mathcal{D}_Z = \left\{log\mathcal{L}_Z(\hat{\boldsymbol{\theta}}_0, \hat{\epsilon}; \mathbf{y}) - log\mathcal{L}_0(\hat{\boldsymbol{\theta}}_0, 1; \mathbf{y})\right\}$ com $\mathcal{L}_Z(\boldsymbol{\theta}, \epsilon; \mathbf{y})$ representando a verossimilhança sob H_1 para uma particular zona Z e $\mathcal{L}_0(\boldsymbol{\theta}, 1; \mathbf{y})$ é a verossimilhança sob H_0. Se $\epsilon = 1$ teremos $\mathcal{D}_Z = 0$, caso contrário

$$\begin{aligned}\mathcal{D}_Z &= \sum_{s_l\in Z} log\left[\frac{e^{\mathbf{G}_l\hat{\boldsymbol{\gamma}}_0} + (1+e^{-\mathbf{H}_l\hat{\boldsymbol{\delta}}_0})^{-1/2}exp\left\{-\hat{\epsilon}e^{\mathbf{B}_l\hat{\boldsymbol{\beta}}_0}/(1+e^{-\mathbf{H}_l\hat{\boldsymbol{\delta}}_0})\right\}}{e^{\mathbf{G}_l\hat{\boldsymbol{\gamma}}_0} + (1+e^{-\mathbf{H}_l\hat{\boldsymbol{\delta}}_0})^{-1/2}exp\left\{-e^{\mathbf{B}_l\hat{\boldsymbol{\beta}}_0}/(1+e^{-\mathbf{H}_l\hat{\boldsymbol{\delta}}_0})\right\}}\right] I(y_l = 0) \\ &+ \sum_{s_l\in Z}(1+e^{-\mathbf{H}_l\hat{\boldsymbol{\delta}}_0})^{-1}(y_l log\hat{\epsilon} - e^{\mathbf{B}_l\hat{\boldsymbol{\beta}}_0}(\hat{\epsilon}-1))I(y_l > 0).\end{aligned}$$

2.3 *Estimação dos parâmetros*

A estimação no modelo é realizada usando o algoritmo EM (Expectation-Maximization). Para tanto, seja o vetor de variáveis latentes $\mathbf{U} = (U_1, ..., U_L)$. Onde, $U_l = 1$ quando Y_l ocorre devido a um estado zero e $U_l = 0$ quando Y_l ocorre devido a um modelo Poisson Duplo. Então as verossimilhanças aumentadas são,

$$\mathcal{L}_Z^a(\boldsymbol{\theta}; \mathbf{y}, \boldsymbol{u}) = \prod_{s_l\in Z} p_l^{u_l}\left[(1-p_l)f_{DP}(y_l|\mu_l, \phi_l)\right]^{1-u_l} \times \prod_{s_l\notin Z} p_l^{u_l}\left[(1-p_l)f_{DP}(y_l|\mu_l, \phi_l)\right]^{1-u_l}.$$

$$\mathcal{L}_0^a(\boldsymbol{\theta}; \mathbf{y}, \boldsymbol{u}) = \prod_{s_l} p_l^{u_l}\left[(1-p_l)f_{DP}(y_l|\mu_l, \phi_l)\right]^{1-u_l}.$$

Marginalmente $Y_l \sim \mathbf{ZIDP}(\mu_l, \phi_l, p_l)$. Para obter $\hat{\boldsymbol{\theta}}_0 = (\hat{\boldsymbol{\beta}}_0, \hat{\boldsymbol{\delta}}_0, \hat{\boldsymbol{\gamma}}_0)$ maximizamos a função $l_0^a(\boldsymbol{\theta}, 1; \mathbf{y}, \boldsymbol{u}) = \log\mathcal{L}_0^a(\boldsymbol{\theta}; \mathbf{y}, \boldsymbol{u})$ usando o algoritmo **EM**:

- No passo **E**: Na k-ésima iteração substituimos $u_l^{(k)}$ por $\mathbb{E}(u_l|y_l, \boldsymbol{\theta}^{(k)})$, onde $\boldsymbol{\theta}^{(k)} = (\boldsymbol{\beta}^{(k)}, \boldsymbol{\delta}^{(k)}, \boldsymbol{\gamma}^{(k)})$.
- No passo **M**: Na (k+1)-ésima iteração maximizamos $l_0^a(\boldsymbol{\theta}, 1; \mathbf{y}, \boldsymbol{u}^{(k)}) = \mathbb{E}\left\{l_0^a(\boldsymbol{\theta}, 1; \mathbf{y}, \boldsymbol{u})|\mathbf{y}, \boldsymbol{\theta}^{(k)}\right\}$ Esta maximização é realizada via o algoritmo Newton-Rapshon Score de Fisher (**NRSF**), tal que $\boldsymbol{\theta}^{(k+1)} = \boldsymbol{\theta}^{(k)} + \mathcal{I}(\boldsymbol{\theta}^{(k)})^{-1}\mathcal{S}(\boldsymbol{\theta}^{(k)})$. Onde, $\mathcal{S}(\boldsymbol{\theta})$ é a função escore e $\mathcal{I}(\boldsymbol{\theta})$ é a matirz de Informação de Fisher sob a hipótese nula. Na (k+1)-ésima

iteração maximizando a função $l_Z^a(\hat{\boldsymbol{\theta}}_0, \epsilon; \mathbf{y}, \boldsymbol{u}^{(k)})$ com respeito a ϵ obtemos,

$$\hat{\epsilon} = \epsilon^{(k+1)} = \frac{\sum_{s_l \in Z}(1 - u_l^{(k)})(1 + e^{-\mathbf{H}_l \hat{\boldsymbol{\delta}}_0})^{-1} y_l}{\sum_{s_l \in Z}(1 - u_l^{(k)})(1 + e^{-\mathbf{H}_l \hat{\boldsymbol{\delta}}_0})^{-1} e^{\mathbf{B}_l \hat{\boldsymbol{\beta}}_0}}$$

2.4 Bootstrap-EM for p-value of the Spatial Scan proposed

- Algoritmo Bootstrap-EM para $\mathcal{D}$.

  1. Baseado nos dado reais $\mathbf{y} = (y_1, ..., y_L)$ e matrizes de covariáveis $(\mathbf{B}, \mathbf{H}, \mathbf{G})$, use o algoritmo EM e compute $(\hat{\boldsymbol{\beta}}_0, \hat{\boldsymbol{\delta}}_0, \hat{\boldsymbol{\gamma}}_0)$ e ϵ. Derive o valor observado de $\mathcal{D}$ e denote por $\hat{d}_Q$.
  2. Gere amostras bootstrap $\mathbf{y}_q^* = (y_{1,q}^*, ..., y_{L,q}^*)$ de $\mathbf{ZIDP}(\mu_l(\hat{\boldsymbol{\beta}}_0), \phi_l(\hat{\boldsymbol{\delta}}_0), p_l(\hat{\boldsymbol{\gamma}}_0))$, $l = 1, 2, ..., L$.
  3. Com base nos dados gerados em 2, use o algoritmo EM e compute os pseudos estimadores $(\hat{\boldsymbol{\beta}}_0^*, \hat{\boldsymbol{\delta}}_0^*, \hat{\boldsymbol{\gamma}}_0^*)$. Derive o pseudo valor de $\mathcal{D}_q^*$ e denote por $\hat{d}_b^*$.
  4. Repitindo os passos 2 e 3 para $q = 1, ..., Q - 1$ compute o p-valor para $\mathcal{D}$ por $p_{valor} \stackrel{\circ}{=} p_{valor}^*(\mathcal{D}) = \frac{1}{Q}\sum_{q=1}^{Q} I(\hat{d} \geq \hat{d}_b^*)$

3. Aplicação e discussões

Nesta aplicação utilizamos dados de novos casos de Haseníase em menores de 15 que ocorreram nos 62 municípios do Estado do Amazonas, Brasil-2010. Foram observados 190 novos casos com uma taxa de 1.525 por 10.000 habitantes. A média e variancia forma respectivamente 2.706 and 7.421. Em 30 municípios (48%) não foram registrados novos casos da doença. Na maioria dos municípios o teste para detecção da doença não é realizado e por isso usamos a distância padronizada entre o i-ésimo município e a capital do estado Manaus (área desenvolvida) como covariável de ajuste para inflação de zeros. Como a Haseníase é uma doença associada a condições de vida da população, usamos para ajuste na média o Índice de Vunerabilidade Social do município(IVS). Nesta aplicação, a sobredispersão é estimada sem o ajuste por covariável. Após a aplicação do método descrito na seção 2, obtivemos as estimativas $\widehat{\boldsymbol{\beta}} = (-8.09578, -0.70807)$, $\widehat{\boldsymbol{\gamma}} = (-1.33778, -0.022434)$, $\widehat{\phi} = 0.439$ e $\widehat{\epsilon} = 3.53534$. O valor-p observado foi 0.00826 e o cluster detectado é apresentado na Figura 5.1.

4. Conclusões

Neste artigo foi proposta uma Nova Estatística Scan ajustada por covariáveis para detecção de cluster espaciais em dados com sobredispersão e inflação de zeros. Um

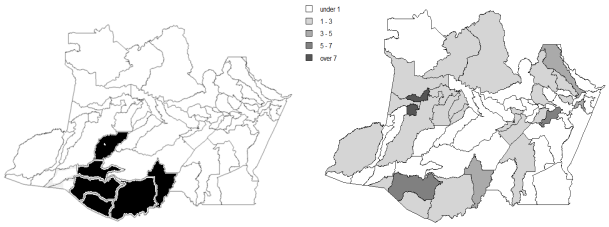

Figura 5.1: Cluster detectado à esquerda; mapa de taxa de novos casos à direita

algoritmo EM foi desenvolvido para estimação dos parâmetros e a significância do cluster foi avaliada via valor-p Bootstrap. O resultado de nossa aplicação sugere que a metodologia proposta é bem efetiva na detecção do cluster e pode ser facilmente modificada para utilização em problemas de vigilância epidemiológica no espaço-tempo.

References

Cançado A., da Silva C. and da Silva M.(2013) A zero-inflated Poisson-based spatial scan statistic. Submitted.

Efron, B. (1986) Double Exponential Families and Their Use in Generalized Linear Regression, *JASA.*, **81**, 709-721.

Kulldorff, M. (1997) A spatial scan statistic. *Communs Statist. Theory Meth.*, **26**, 1481-1496.

Lima, M. S. ; Duczmal, L.H ; Cardoso Neto, J. ; Pinto, L. P.(2013). Spatial Scan Statistics for Models with Overdispersion and Excess Zeros. Submitted.

Meng, B.J., Zhu Z.(2007) Accounting for spatial correlation in the Scan Statitics. *The Annals of Applied Statistics*, **2**, 560-584.

Zhang, T., Lin, G. (2009) Spatial scan statistics in loglinear models *Computational Statistics and Data Analysis*, **53**, 2851-2858.

Zhang, T., Zhang Z.; Lin G.,(2012) Spatial scan statistics with overdispersion. *Statatistics Medicine*, **2** (8):762-774.