Um Alisador de Máxima Verosimilhança Local (MVL) com o Modelo de Regressão de Poisson, para Análise de Regressão de Dados de Contagem

por

André Mateus de Carvalho Monteiro Faro Santana

Dissertação apresentada como requisito
parcial para obtenção do grau de

Mestre em
Estatística e Gestão de Informação

pelo

Instituto Superior de Estatística e Gestão de Informação
da
Universidade Nova de Lisboa

Resumo

O modelo de regressão de Poisson é a base da análise de regressão paramétrica de dados de contagem, mas as restrições impostas por este modelo são fortes e, frequentemente, não são respeitadas na prática, nomeadamente a hipótese de equidispersão, isto é, de que o valor médio e a variância condicionais são iguais. O modelo binomial negativo admite a possibilidade de sobredispersão, porém, quando esta é elevada, o modelo binomial negativo não se ajusta aos dados. Uma situação recorrente na prática é a presença de excesso de zeros, em que os modelos mais adequados são o modelo de Poisson inflacionado em zero (ZIP) e o modelo binomial negativo inflacionado em zero (ZINB).

Têm surgido diversos modelos não paramétricos e semi-paramétricos que não impõem as restrições dos modelos paramétricos, nem dependem da correcta especificação do modelo. Em Santos (2005), é desenvolvido um alisador de máxima verosimilhança local de Poisson e são deduzidos o seu viés, variância e distribuição assimptótica, apresentado boa aderência a dados reais e a dados de simulação. Este modelo, apesar de apresentar bom desempenho, revela-se computacionalmente pesado, devido à sua especificação exponencial.

Neste trabalho é apresentado um modelo de Poisson de máxima verosimilhança local, com base no alisador de núcleo e na regressão polinomial local, que deixa cair a especificação exponencial, dada em Santos (2005), passando o modelo a ser especificado localmente por um polinómio do primeiro grau.

Palavras-chave: Alisador de núcleo, dados de contagem, máxima verosimilhança local, regressão de Poisson, regressão polinomial local, regressão semi-paramétrica

Abstract

The Poisson regression model is the basis of parametric regression analysis of count data, but the restrictions imposed by such models are strong, and often they are not met in practice, namely the equal dispersion hypothesis, that is the equality between the conditional mean and the conditional variance. The negative binomial model allows for some overdispersion. However, when overdispersion is high, the negative binomial model fits poorly. A situation we often encounter in practice is excess zeros, in which the most adequate models are the zero-inflated Poisson (ZIP) and the zero-inflated negative binomial (ZINB).

Many nonparametric and semi-parametric models have been proposed that do not make such parametric model assumptions, nor do they depend on correct model specification. In Santos (2005), a local maximum likelihood estimator for Poisson regression is presented, as well as its bias, variance and assimptotic distribution are derived, showing a good fit to real data and to simulated data. This model, despite showing good performance, is quite heavy, due to its exponential specification.

In this work, an alternative Poisson local maximum likelihood model is presented, based on kernel smoothing and local polynomial regression, which drops the exponential specification, presented in Santos (2005) and a local polynomial specification is used instead.

Keywords: count data, kernel smoothing, local maximum likelihood, local polynomial regression, Poisson regression

Agradecimentos

A primeira palavra de agradecimento é endereçada ao meu orientador, Professor Doutor José António Rui Amaral Santos, cujo contributo para este trabalho foi inexcedível. Agradeço a sua disponibilidade e o seu interesse, e todas as suas sugestões, bem como as palavras de motivação. Por tudo isto, posso afirmar, com toda a justiça, que as suas contribuições extravasaram a simples orientação, foram antes um apoio com dedicação e amizade.

Aos meus pais, verdadeiros pilares de suporte na minha vida. Por todo o apoio que me dão, em tudo, e ao garantirem que nada me falta. Quaisquer palavras que escreva não chegarão para expressar a gratidão pelo seu esforço, e por aquilo de que prescindem, para que supere as minhas dificuldades, quaisquer que sejam, académicas ou não. O meu apreço não pode ser expresso em apenas algumas linhas.

Ao meu irmão Simão e à sua esposa Carla, minha cunhada. Também pelo contributo para a minha vida, e por toda a amizade e apoio. Ao meu sobrinho Diogo, que me dá força, com a sua presença e brincadeira de criança. Aos três não consigo agradecer o suficiente.

Ao meu primo Henrique, pela sua colaboração e ensinamentos técnicos, quanto à elaboração e pormenores, na redacção do documento. A ele, à sua irmã Helena e aos meus tios, pela sua amizade. Estendo o apreço a toda a minha família, tios e primos, apesar de distante, pela presença e amizade.

À Fátima, nossa amiga e funcionária (por esta ordem), por toda a sua pronta dedicação, empenho e amizade.

Muitos são, felizmente, os amigos e amigas que merecem o meu profundo agradecimento, apreço e consideração. A lista seria longa, pelo que não vou

particularizar, e deixo apenas estas singelas linhas, insuficientes para expressar a minha sentida gratidão pela amizade e companhia.

A toda a gente no ISEGI, que, sem excepção, desde que cheguei a esta casa, para a licenciatura, sempre demonstrou a maior disponibilidade e simpatia. A todos um agradecimento.

Lista de Abreviaturas e Tabelas e Figuras

Abreviaturas:

AMISE	Erro quadrático médio integrado assintótico
MAE	Erro absoluto médio
MAPE	Erro absoluto percentual médio
MISE	Erro quadrático médio integrado
MSE	Erro quadrático médio
MVL	Máxima verosimilhança local
RMSE	Raiz quadrada do erro quadrático médio
ZINB	Modelo binomial negativo inflacionado em zero
ZIP	Modelo de Poisson inflacionado em zero

Tabelas:

Figuras:

Índice

1. Introdução

Os dados de contagem podem ser vistos como resultantes de contagens de acontecimentos num período e/ou secção, pelo que a variável resposta apenas toma valores inteiros não negativos. Quando as contagens dependem de variáveis observáveis, tem interesse efectuar a análise de regressão de dados de contagem. A análise de dados de contagem apresenta interesse e aplicabilidade crescentes, em diversos campos, desde a sinistralidade, a tecnologia, as ciências sociais e económicas, bem como a epidemiologia e a saúde.

Porém, a natureza discreta e não negativa da variável resposta levanta desafios, nomeadamente, inviabilizando a aplicação do modelo clássico de regressão linear. O modelo clássico de regressão linear revela-se inadequado para aplicação a dados de contagem (cf. King (1988) citado em Santos (2005)).

A regressão paramétrica de dados de contagem assenta sobretudo em modelos cuja base é a distribuição de Poisson ou a distribuição binomial negativa. Estes modelos são "candidatos naturais" *a priori* da análise de dados de contagem. Um pré-requisito forte no modelo de Poisson é a hipótese de equidispersão, isto é, de que o valor médio e a variância condicionais da distribuição de probabilidade (das ocorrências) são iguais. Esta hipótese é frequentemente rejeitada na prática. Nestas situações, o modelo binomial negativo é uma alternativa melhor.

Uma situação frequentemente encontrada na prática é o excesso de contagens zero. Neste trabalho entende-se excesso, ou inflação, de contagens zero como dados em que as frequências de zero são superiores às esperadas pelo modelo de Poisson. Existem modelos aplicáveis nesta situação com base no modelo de Poisson (ZIP – Zero-Inflated Poisson – Lambert (1992)) e binomial negativo (ZINB – Zero-Inflated Negative Binomial – Hall (2000)).

No capítulo 2 são apresentados os principais modelos de regressão paramétrica de dados de contagem, nomeadamente o modelo de Poisson, o modelo Binomial Negativo, o modelo de Poisson inflacionado em zero (ZIP), e o modelo Binomial Negativo inflacionado em zero (ZINB).

Note-se que se a especificação do modelo paramétrico não for a correcta, a estimação através do modelo não é válida, e a inferência estatística poderá conduzir a conclusões erradas. É aqui que reside a vantagem dos modelos não paramétricos: não estabelece condições *a priori* sobre a distribuição da variável resposta, além de que podem fornecer indicações muito úteis sobre a distribuição da variável resposta, para posterior estimação de um modelo paramétrico. De entre os métodos de regressão não paramétrica, destacam-se os alisadores de núcleo, a regressão polinomial local, os alisadores de *spline*, séries ortogonais e os *wavelets*, a quasi-verosimilhança e a verosimilhança local. Os modelos de regressão do núcleo e de regressão polinomial local serão apresentados no capítulo 3.

Tem sido objecto de trabalho o desenvolvimento de modelos que combinam as abordagens paramétrica e não paramétrica, em formulações designadas semi-paramétricas, de grande interesse e aplicabilidade práticas. Veja-se Sun (2001), Podlich, Faddy e Smyth (2004) e, mais recentemente, Dean, Nathoo e Nielsen (2007), Nielsen e Dean (2008), Santos (2005) e Santos e Neves (2008), entre outros. A maior limitação destes modelos semi-paramétricos prende-se com o facto de serem computacionalmente pesados (Podlich, Faddy e Smyth (2004)).

Em Santos (2005) são propostas diversas formulações semi-paramétricas, com base no modelo de regressão de Poisson e no modelo de regressão binomial negativo, e na máxima verosimilhança local, proposta por Tibshirani e Hastie (1987). Em particular, é proposto o alisador de máxima verosimilhança local com o modelo de Poisson (ponto 5.2 pp 65-70), sendo deduzidos o seu viés, variância e distribuição assintótica. Este método, independentemente da

qualidade do ajustamento conseguido, revela-se computacionalmente pesado, uma vez que se trata de uma especificação exponencial. Neste modelo, o valor médio é especificado como a exponencial de uma combinação linear das covariáveis, uma função desconhecida a estimar por máxima verosimilhança local:

$$\lambda(x) = e^{m(x)}, \tag{1.1}$$

em que

$$m(x) \approx \beta_0 + \beta_1 (x_i - x), \tag{1.2}$$

no caso univariado.

No capítulo 4 é formulado um modelo de máxima verosimilhança local de Poisson, alternativo ao de Santos (2005), em que o valor médio não é dado pela função exponencial de uma combinação linear de covariáveis, mas antes pela combinação linear de covariáveis, ela própria uma função desconhecida São estimados o viés, a variância e a distribuição assintótica. Esta é, por conseguinte, uma formulação alternativa à proposta em Santos (2005):

$$\lambda(x) = m(x) \approx \beta_0 + \beta_1 (x_i - x). \tag{1.3}$$

Como o valor médio já não é estimado como uma exponencial, mas sim como um polinómio do primeiro grau, através da máxima verosimilhança local, são esperados ganhos computacionais significativos. É objectivo deste trabalho verificar se, por um lado, há ganhos computacionais assinaláveis e, por outro lado, se não há perda de aderência do modelo significativa.

No capítulo 5 é efectuado o estudo por simulação, e também um estudo de caso de regressão do número de infecções urinárias de indivíduos infectados com o vírus HIV em função do número de células CD4+. O modelo aqui proposto é aplicado aos dados de simulação e a um *data set* de dados reais gentilmente cedido por Santos(2005). Foi utilizado o *software* gratuito R (http://www.r-project.org/).

No capítulo 6 efectua-se a discussão dos resultados, e apresenta-se as principais conclusões. Pretende-se avaliar a robustez do desempenho do estimador face à especificação errada do modelo, bem como verificar se há ganhos computacionais assinaláveis, sem prejuízo da aderência aos dados. É, ainda, sugerido trabalho a desenvolver no futuro.

2. Apresentação dos Principais Modelos de Regressão Paramétrica de Dados de Contagem

Estes modelos, bem como as suas propriedades, são apresentados em referências bibliográficas como Cameron *et al* (1998), Winkelmann (2000) e Santos (2005). Os modelos paramétricos de regressão paramétrica de dados de contagem têm como base as distribuições discretas de Poisson e Binomial Negativa. Estes modelos apresentam condições de especificação fortes que, quando não verificadas, acarretam consequências nas inferências estatísticas, isto é, a incorrecta especificação do modelo implica conclusões erradas. Contudo, a sua utilização é frequente, nomeadamente se as condições impostas são satisfeitas, caso em que estes revelam-se mais poderosos, como por exemplo McCarthy *et al* (2008), num estudo de previsão da procura de serviços de urgência hospitalar, em que os histogramas observados são praticamente idênticos aos previstos pelo modelo de Poisson. Um pré-requisito forte no modelo de Poisson é a hipótese de equidispersão, isto é, de que o valor médio e a variância condicionais da distribuição de probabilidade do número de ocorrências são iguais. Esta hipótese é frequentemente rejeitada na prática. Muito frequentemente, na prática, os dados exibem sobredispersão (menos frequente é a ocorrência de subdispersão). Deng *et al* (2005) e Jung *et al* (2005) apresentam testes para a presença de sobredispersão. A distribuição binomial negativa, ao permitir uma especificação superior para a variância condicional (que na distribuição de Poisson é igual ao valor médio), e ao dar conta de alguma heterogeneidade não observada, faz com que, para uma sobredispersão moderada, o modelo de regressão binomial negativo produza melhores ajustamentos. Recentemente, Tsou (2006) apresenta uma metodologia paramétrica para análise de dados de contagem, com base no modelo de Poisson, com ajustamentos ao nível das funções de ligação e identidade, que torna a regressão robusta quanto às hipóteses do modelo de

Poisson, desde que as verdadeiras funções de distribuição subjacentes tenham momentos de segunda ordem finitos.

No entanto, é recorrente acontecer, na prática, um número de zeros, ou contagens zero, excessivo face ao esperado pela distribuição de Poisson. Quando ocorre esta situação, denominada inflação em zero, causa sobredispersão. Existem trabalhos recentes em que são propostos testes para a inflação em zero, em que se incluem Yang *et al* (2010a). Existem modelos paramétricos, com base nas distribuições acima referidas, truncadas em zero: o modelo de Poisson inflacionado em zero (ZIP), e o modelo binomial negativo inflacionado em zero (ZINB), respectivamente. Joe e Zhu (2005) apresentam uma comparação de desempenho de modelos com base nas distribuições de Poisson generalizada e binomial negativa, bem como as respectivas distribuições inflacionadas em zero, e de Poisson generalizados. Yang *et al* (2010b) estendem os testes de sobredispersão aos modelos inflacionados em zero. Existem ainda modelos de mistura de Poisson e de Poisson generalizados, que podem ser vistos, por exemplo, em Karlis e Xekalaki (2005), e Xie *et al* (2008).

2.1 Modelo de Regressão de Poisson

O modelo de Poisson, cuja formulação é apresentada, por exemplo, em Winkelmann (2000), é a base nuclear da regressão paramétrica de dados de contagem. É o modelo mais simples, razão pela qual é apelativo e é ainda largamente utilizado. Recentemente, Coxe *et al* (2009) discutem a sua aplicação e as suas alternativas. Mas a adequação do seu uso depende da correcta especificação do modelo, assim como as conclusões tiradas. A hipótese de equidispersão constitui um pré-requisito muito forte e rígido, que na prática é frequentemente contrariada. Como já foi referido, as situações de subdispersão (variância inferior ao valor médio condicionais) e sobredispersão (variância inferior superior ao valor médio condicionais) são frequentes, sobretudo esta última. A heterogeneidade não observada da população, o

contágio entre observações, e também os erros de medida das covariáveis contribuem para a rotura desta hipótese. Outro factor que influencia a validade da sua aplicação é a presença de deflação ou inflação em zero (defeito ou excesso de contagens zero), com prevalência de situações de inflação em zero, muitas vezes observadas, como resultantes naturais do fenómeno observado e a explicar. O modelo de regressão binomial negativo acomoda situações de quebra da equidispersão, enquanto os modelos inflacionados em zero são indicados para situações mais acentuadas de excesso de zeros (Ridout *et al* (1998)).

2.1.1 Especificação

O modelo de regressão de Poisson, (2000), parte de três hipóteses:

1) Igualdade entre valor médio e variância condicionais (equidispersão)

$$\lambda(\mathbf{x_i}) = E\left[Y_i \middle| \mathbf{x_i}\right] = Var\left[Y_i \middle| \mathbf{x_i}\right] \tag{2.1}$$

2) O valor médio, $\lambda(\mathbf{x_i})$, é a exponencial de uma combinação linear das covariáveis, $\mathbf{x_i}$.

$$\lambda(\mathbf{x_i}) = \lambda(\mathbf{x_i};\beta) = e^{\left(\mathbf{x_i}'\beta\right)} \Leftrightarrow \lambda(\mathbf{x_i}) = e^{\beta_0+\beta_1 x} \quad i = 1,\mathrm{K}\ ,n\,, \tag{2.2}$$

no caso univariado.

3) A distribuição condicional da variável resposta é:

$$Y_i \middle| \mathbf{x_i} \sim Poisson\left(\lambda(\mathbf{x_i})\right) \quad i = 1,\mathrm{K}\ ,n\,,$$

ou seja, a função massa de probabilidade é dada por:

$$P\left(Y_i = y_i \left| \mathbf{x_i} \right.\right) = \frac{e^{-\lambda(\mathbf{x_i})} \lambda^{y_i}}{y_i!} \quad i = 1, \mathrm{K}\ , n\ . \tag{2.3}$$

A primeira hipótese, da equidispersão, é frequentemente rejeitada na prática, quando há subdispersão, e sobretudo, sobredispersão (variância superior ao valor médio condicionais), o que pode ficar a dever-se, à existência de heterogeneidade na população, contágio entre acontecimentos, erros de medida de covariáveis, frequência elevada de zeros. A imposição desta hipótese pode causar uma subestimação de erros-padrão dos parâmetros, o que resultaria em estatísticas t sobrestimadas enviesando os testes de significância destes, empolando a significância dos regressores em geral. As conclusões retiradas, nomeadamente inferências sobre parâmetros e intervalos de confiança podem ser enganadoras. (cf Santos (2005)).

A hipótese 3) é frequentemente violada, por não haver independência entre as observações, na medida em que, frequentemente há autocorrelação (a contagem de acontecimentos depende de ocorrências prévias) em dados de natureza temporal, seccionais com autocorrelação espacial ou em rede.

2.1.2 Estimação

A função de log-verosimilhança é dada por:

$$\mathcal{L}(\beta|\mathbf{x}) = \sum_{i=1}^{n}\left(-\lambda\left(\mathbf{x_i}'\beta\right)\right) + y_i \ln\left(\lambda\left(\mathbf{x_i}'\beta\right)\right) - \ln(y_i!), \tag{2.4}$$

com $\lambda\left(\mathbf{x_i}'\beta\right) = e^{\left(\mathbf{x_i}'\beta\right)}$ (equação 2.2), donde resulta que as condições de primeira ordem sejam dadas por:

$$\sum_{i=1}^{n}\left[-e^{\lambda\left(\mathbf{x_i}'\beta\right)} + y_i\right]\mathbf{x_i} = \mathbf{0} \tag{2.5}$$

2.1.3 Propriedades

Demonstra-se que, sob a correcta especificação do modelo e certas condições de regularidade, o estimador é consistente, centrado e assintoticamente normal:

$$\hat{\beta} \overset{a}{\sim} N\left(\beta, \left(\sum_{i=1}^{n} e^{\left(\mathbf{x_i}'\beta\right)}\mathbf{x_i}\mathbf{x_i}'\right)^{-1}\right) \tag{2.6}$$

2.2 Modelo de Regressão Binomial Negativo

O modelo de regressão binomial negativo, embora tenha as mesmas limitações do modelo de regressão de Poisson, como qualquer modelo de regressão paramétrico (não é válido sob a incorrecta especificação do modelo), e, como neste, as propriedades assintóticas (tais como a eficiência, a consistência e a

normalidade assintótica) também dependam de algumas condições de regularidade (como a generalidade dos modelos de natureza paramétrica), possui a vantagem de dar conta de alguma heterogeneidade não observada da população, e, como se verá, permite modelar a sobredispersão devido a uma modelação mais flexível da variância condicional. É, por isso, um modelo mais adequado que o de Poisson.

2.2.1 Especificação

A especificação genérica é:

$$\lambda_i = \lambda(\mathbf{x_i}) = E\left[Y_i \middle| \mathbf{x_i}\right] = e^{\left(\mathbf{x_i}'\beta\right)}, \tag{2.7}$$

$$Var\left[Y_i \middle| \mathbf{x_i}\right] = \lambda_i + \alpha^p \lambda_i^2, \quad \alpha \geq 0, \quad p = 1,2,\mathrm{K}\ , \tag{2.8}$$

α é um parâmetro de dispersão a estimar. Como α é não negativo, verifica-se que a variância condicional é, ou pode ser, maior do que o valor médio. O parâmetro p indica a ordem do modelo: com p=1 é a especificação proposta em Cameron *et al* (1998) como NB1 e p=2 como NB2. Note-se que se α=0 tem-se o modelo de Poisson, pelo que se pode dizer que o modelo binomial negativo é uma generalização do de Poisson.

A função massa de probabilidade é dada por (NB2):

$$P\left(Y_i = y_i \middle| \mathbf{x_i}, \alpha\right) = \frac{\Gamma\left(y_i + \alpha^{-1}\right)}{\Gamma\left(y_i + 1\right)\Gamma\left(\alpha^{-1}\right)}\left(\frac{\alpha^{-1}}{\alpha^{-1} + \lambda_i}\right)^{\alpha^{-1}}\left(\frac{\lambda_i}{\alpha^{-1} + \lambda_i}\right)^{y_i} \quad y_i = 1,2,\mathrm{K}$$

2.2.2 Estimação

As condições de primeira ordem para a estimação de β e α são:

$$\sum_{i=1}^{n} \frac{y_i - e^{\left(\mathbf{x_i}'\beta\right)}}{1+\alpha e^{\left(\mathbf{x_i}'\beta\right)}} \mathbf{x_i} = \mathbf{0}$$

$$\sum_{i=1}^{n} \left[\frac{1}{\alpha^2} \left(\ln\left(1+\alpha e^{\left(\mathbf{x_i}'\beta\right)}\right) - \sum_{j=0}^{y_i-1} \frac{1}{j+\alpha^{-1}} \right) + \frac{y_i - e^{\left(\mathbf{x_i}'\beta\right)}}{\alpha\left(1+\alpha e^{\left(\mathbf{x_i}'\beta\right)}\right)} \right] = 0$$

2.2.3 Propriedades

Se a especificação do modelo for a correcta, tem-se assintoticamente:

$$\begin{bmatrix} \hat{\beta} \\ \hat{\alpha} \end{bmatrix} \overset{a}{\sim} N\left(\begin{bmatrix} \beta \\ \alpha \end{bmatrix}, \begin{bmatrix} Var\left[\hat{\beta}\right] & 0 \\ 0 & Var\left[\hat{\alpha}\right] \end{bmatrix} \right),$$

com

$$Var\left[\hat{\beta}\right] = \left(\sum_{i=1}^{n} \frac{e^{\left(\mathbf{x_i}'\beta\right)}}{1+\alpha e^{\left(\mathbf{x_i}'\beta\right)}} \mathbf{x_i}\mathbf{x_i}' \right)^{-1}$$

$$Var\left[\hat{\alpha}\right] = \left[\sum_{i=1}^{n} \left[\frac{1}{\alpha^4} \left(\ln\left(1+e^{\left(\mathbf{x_i}'\beta\right)}\right) \right) - \sum_{j=0}^{y_i-1} \frac{1}{j+\alpha^{-1}} \right] \right]^{-1}$$

$$\operatorname{cov}\left[\hat{\beta},\hat{\alpha}\right] = \mathbf{0}$$

2.3 Modelo de Poisson Inflacionado em Zero (ZIP)

O modelo de regressão de Poisson inflacionado em zero (ZIP) constitui um progresso face ao modelo de regressão de Poisson, na medida em que admite a heterogeneidade na população, assim como adere a dados em que o número de observações zero é superior ao esperado pelo modelo de Poisson, e portanto, dá conta de situações de sobredispersão acentuada. Uma forma de expor a especificificação é admitir que a heterogeneidade se prende com a divisão da população em estudo em duas subpopulações: a subpopulação A, onde apenas se verificam contagens zero, com probabilidade ψ, e a subpopulação B, onde as contagens são geradas segundo uma distribuição de Poisson. Hall e Shen (2010) apresentam modelos ZIP robustos quanto ao excesso de zeros. Existem diversos trabalhos em que são desenvolvidos testes para sub e sobredispersão em modelos ZIP, Van den Broeck (1995), Ridout *et al* (2001), Jansakul e Hinde (2002) *apud* Lee *et al* (2006), Deng e Paul (2005), Jung *et al* (2005) e Yang *et al* (2009). Lee *et al* (2011) apresentam testes de inflação em zero para modelos de Poisson inflacionados em zero bivariados. Mas, frequentemente, conforme salientam Lee *et al* (2006), devido ao desenho hierárquico do estudo ou ao método de recolha de dados, a inflação em zero ocorre simultaneamente com a falta de independência das observações, o que torna o modelo ZIP inadequado. apresentam uma classe de modelos ZIP multi-nível que dão conta de ambas as ocorrências em simultâneo. Os modelos são generalizados para dar conta de diferentes formas de correlação das ocorrências do fenómeno em estudo mais complexas.

Uma comparação entre o desempenho de modelos de Poisson e modelos ZIP pode ser vista em Naya *et al* (2008). Mir (2011) efectua a estimação com a distribuição de Poisson, em *software* R.

2.3.1 Especificação

O modelo ZIP de parâmetros λ e ψ, tem a seguinte especificação:

$$P\left(Y_i = y_i \left| \mathbf{x_i} \right.\right) = \begin{cases} \psi + (1-\psi) e^{-\lambda(\mathbf{x_i})}, & y_i = 0 \\ (1-\psi) \dfrac{e^{-\lambda(\mathbf{x_i})} \lambda(\mathbf{x_i})^{y_i}}{y_i!}, & y_i > 0 \end{cases}, \tag{2.9}$$

ou alternativamente:

$$P\left(Y_i = y_i \left| \mathbf{x_i} \right.\right) = d_i\psi + (1-d_i)(1-\psi) \frac{e^{-\lambda(\mathbf{x_i})} \left[\lambda(\mathbf{x_i})\right]^{y_i}}{y_i!}, \quad Y_i = 0,1,2,\mathrm{K} \tag{2.10}$$

em que $d_i = \begin{cases} 1, & i \in A \\ 0, & i \in B \end{cases}$ é uma variável indicatriz que define a pertença do indivíduo *i* da amostra à população A (subpopulação de contagens zero).

Pode demonstra-se que (Santos (2005)):

$$E\left[Y_i \left| \mathbf{x_i} \right.\right] = (1-\psi)\lambda(\mathbf{x_i})$$

$$Var\left[Y_i \left| \mathbf{x}_i \right.\right] = (1-\psi)\lambda(\mathbf{x_i})\left[1-\psi\lambda(\mathbf{x_i})\right]$$

.3.2 Estimação

As funções verosimilhança e logverosimilhança são dadas, respectivamente, por:

$$L\left(\beta, \psi \left| \mathbf{x}, \mathbf{y}, \mathbf{d} \right.\right) = \prod_{i=1}^{n} \left[d_i\psi + (1-d_i)(1-\psi) \frac{e^{-\lambda(\mathbf{x_i})} \left[\lambda(\mathbf{x_i})\right]^{y_i}}{y_i!} \right] \tag{2.11}$$

e

$$\mathcal{L}\left(\beta,\psi\left|\mathbf{x},\mathbf{y},\mathbf{d}\right.\right)=\sum_{i=1}^{n}\ln\left[d_i\psi+\left(1-d_i\right)\left(1-\psi\right)\frac{e^{-\lambda(\mathbf{x_i})}\left[\lambda\left(\mathbf{x_i}\right)\right]^{y_i}}{y_i!}\right]. \tag{2.12}$$

Assume-se d_i observada, isto é, se sabemos se a observação pertence a A ou B.

As condições de primeira ordem são:

$$\frac{\partial\mathcal{L}}{\partial\beta}=\sum_{i=1}^{n}\frac{\left(1-d_i\right)\left(1-\psi\right)\left(-\lambda\left(\mathbf{x_i}\right)e^{-\lambda(\mathbf{x_i})}\left[\lambda\left(\mathbf{x_i}\right)\right]^{y_i}+y_i\left[\lambda\left(\mathbf{x_i}\right)\right]^{y_i}\right)}{y_i!d_i\psi+\left(1-d_i\right)\left(1-\psi\right)e^{-\lambda(\mathbf{x_i})}\left[\lambda\left(\mathbf{x_i}\right)\right]^{y_i}}\mathbf{x_i}=\mathbf{0}$$

$$\frac{\partial\mathcal{L}}{\partial\psi}=\sum_{i=1}^{n}\frac{y_i!d_i-\left(1-d_i\right)e^{-\lambda(\mathbf{x_i})}\left[\lambda\left(\mathbf{x_i}\right)\right]^{y_i}}{y_i!d_i\psi+\left(1-d_i\right)\left(1-\psi\right)e^{-\lambda(\mathbf{x_i})}\left[\lambda\left(\mathbf{x_i}\right)\right]^{y_i}}=0$$

2.3.3 Propriedades

Caso a especificação seja correcta, o estimador dos parâmetros do modelo segue assintoticamente uma distribuição normal:

$$\hat{\beta}\overset{a}{\sim}N\left(\beta;Var\left(\hat{\beta}\right)\right), \tag{2.13}$$

com

$$Var\left(\hat{\beta}\right)=\left(\sum\left(\psi-1\right)\left[\frac{e^{-\lambda(\mathbf{x_i})}\left[\lambda\left(\mathbf{x_i}\right)\right]^{y_i}\left[\left(y_i-\left(\lambda\left(\mathbf{x_i}\right)\right)^2-\lambda\left(\mathbf{x_i}\right)\right)\right]}{y_i!d_i\psi+\left(1-\psi\right)e^{\left[\lambda(\mathbf{x_i})\right]}\lambda\left(\mathbf{x_i}\right)^{y_i}}\right]^2\mathbf{x_i}\mathbf{x_i}'\right)$$

2.4 Modelo Binomial Negativo Inflacionado em Zero (ZINB) com probabilidade variável

Neste modelo, tal como no anterior, admite-se duas subpopulações A e B, em que a subpopulação A contém exclusivamente contagens zero, mas a subpopulação B contém contagens segundo uma distribuição NB2, com a respectiva variável indicatriz d_i definida e observada. Este modelo tem a vantagem de a errada especificação ser menos gravosa do que o modelo ZIP, no que toca à inconsistência dos estimadores. Acresce ainda que este último admite que as contagens não nulas têm distribuição de Poisson truncada em zero, pelo que não se ajusta em dados em que a sobredispersão é elevada. O modelo ZINB constitui então uma alternativa apropriada, adaptando-se melhor a dados de inflação ainda mais acentuada do que o modelo ZIP (Garay (2011)), apresentado no ponto anterior. Existem testes recentes de sobredispersão, Ridout *et al* (2001) e Hall *et al* (2002) cf Santos (2005) e também para a inflação em zero (Jansakul e Hinde (2009). Porém, tal como no modelo ZIP exposto, devido ao desenho hierárquico do estudo ou ao método de recolha de dados, a inflação em zero ocorre simultaneamente com a falta de independência das observações, o que torna o modelo ZIP inadequado. Moghimbeigi *et al* (2008) propõem modelos ZINB semi-paramétricos que, para estimação, combinam o algoritmo EM (*Expectation Maximization*), pseudoverosimilhança e máxima verosimilhança restrita, aplicáveis nestas condições.

2.4.1 Especificação

Seja $\mathbf{z_i}$ um vector de covariáveis que determinam $P(i \in A) = \psi_i$, de forma que:

$$\psi_i = \psi(\mathbf{z_i}) = P(d_i = 1 | \mathbf{z_i}) = F(\mathbf{z_i}\gamma), \tag{2.14}$$

em que $F(.)$ é uma função distribuição como a normal ou a logística:

$$P\left(Y_i = y_i \left| \mathbf{x_i}, \mathbf{z_i} \right.\right) = \begin{cases} \psi\left(\mathbf{z_i}\right) + \left(1-\psi\left(\mathbf{z_i}\right)\right)\left(\dfrac{\alpha^{-1}}{\alpha^{-1}+\lambda\left(\mathbf{x_i}\right)}\right)^{\alpha^{-1}}, & y_i = 0 \\ \left(1-\psi\left(\mathbf{z_i}\right)\right)\dfrac{\Gamma\left(y_i+\alpha^{-1}\right)}{\Gamma\left(y_i+1\right)\Gamma\left(\alpha^{-1}\right)}\left(\dfrac{\alpha^{-1}}{\alpha^{-1}+\lambda\left(\mathbf{x_i}\right)}\right)^{\alpha^{-1}}\left(\dfrac{\lambda\left(\mathbf{x_i}\right)}{\alpha^{-1}+\lambda\left(\mathbf{x_i}\right)}\right)^{y_i}, & y_i > 0 \end{cases} \tag{2.15}$$

ou simplificadamente:

$$P\left(Y_i = y_i \left| \mathbf{x_i}, \mathbf{z_i} \right.\right) = d_i\psi\left(\mathbf{z_i}\right) + \left(1-d_i\right)\left(1-\psi\left(\mathbf{z_i}\right)\right)\frac{\Gamma\left(y_i+\alpha^{-1}\right)}{\Gamma\left(y_i+1\right)\Gamma\left(\alpha^{-1}\right)}\left(\frac{\alpha^{-1}}{\alpha^{-1}+\lambda\left(\mathbf{x_i}\right)}\right)^{\alpha^{-1}}\left(\frac{\lambda\left(\mathbf{x_i}\right)}{\alpha^{-1}+\lambda\left(\mathbf{x_i}\right)}\right)$$

O valor médio e a variância condicionais são dados por:

$$E\left[Y_i \left| \mathbf{x_i}, \mathbf{z}_i \right.\right] = \left(1-\psi\left(\mathbf{z_i}\right)\right)\lambda\left(\mathbf{x_i}\right)$$

e

$$Var\left[Y_i \left| \mathbf{x_i}, \mathbf{z_i} \right.\right] = \left(1-\psi\left(\mathbf{z_i}\right)\right)\lambda\left(\mathbf{x_i}\right)\left[1+\left(\alpha+\psi\left(\mathbf{z_i}\right)\right)\lambda\left(\mathbf{x_i}\right)\right],$$

isto é, verifica-se a sobredispersão.

2.4.2 Estimação

A função verosimilhança é:

$$L\left(\beta,\gamma,\alpha\middle|\mathbf{X},\mathbf{Z},\mathbf{y},\mathbf{d}\right)=\prod_{i=1}^{n}\left[d_i\psi\left(\mathbf{z_i}\right)+\left(1-d_i\right)\left(1-\psi\left(\mathbf{z_i}\right)\right)\frac{\Gamma\left(y_i+\alpha^{-1}\right)}{\Gamma\left(y_i+1\right)\Gamma\left(\alpha^{-1}\right)}\right.$$
$$\left.\times\left(\frac{\alpha^{-1}}{\alpha^{-1}+\lambda\left(\mathbf{x_i}\right)}\right)^{\alpha^{-1}}\times\left(\frac{\lambda\left(\mathbf{x_i}\right)}{\alpha^{-1}+\lambda\left(\mathbf{x_i}\right)}\right)^{y_i}\right]$$

A função log-verosimilhança é dada por

$$\mathcal{L}\left(\beta,\gamma,\alpha\middle|\mathbf{X},\mathbf{Z},\mathbf{y},\mathbf{d}\right)=\sum_{i=1}^{n}\ln\left[d_i\psi\left(\mathbf{z_i}\right)+\left(1-d_i\right)\left(1-\psi\left(\mathbf{z_i}\right)\right)\frac{\Gamma\left(y_i+\alpha^{-1}\right)}{\Gamma\left(y_i+1\right)\Gamma\left(\alpha^{-1}\right)}\right.$$
$$\left.\times\left(\frac{\alpha^{-1}}{\alpha^{-1}+\lambda\left(\mathbf{x_i}\right)}\right)^{\alpha^{-1}}\times\left(\frac{\lambda\left(\mathbf{x_i}\right)}{\alpha^{-1}+\lambda\left(\mathbf{x_i}\right)}\right)^{y_i}\right]$$

As condições de primeira ordem para a estimação de β, γ e α são:

$$\sum_{i=1}^{n}\frac{\left(1-\psi\left(\mathbf{z_i}\right)\right)\prod_{j=0}^{y_i-1}\left(j+\alpha^{-1}\right)\left(\frac{\alpha^{-1}}{\alpha^{-1}+\lambda\left(\mathbf{x_i}\right)}\right)^{\alpha^{-1}+1}\left(\frac{\lambda\left(\mathbf{x_i}\right)}{\alpha^{-1}+\lambda\left(\mathbf{x_i}\right)}\right)^{y_i}\left(y_i-\lambda\left(\mathbf{x_i}\right)\right)}{y_i!d_i\psi\left(\mathbf{z_i}\right)+\left(1-\psi\left(\mathbf{z_i}\right)\right)\prod_{j=0}^{y_i-1}\left(j+\alpha^{-1}\right)\left(\frac{\alpha^{-1}}{\alpha^{-1}+\lambda\left(\mathbf{x_i}\right)}\right)^{\alpha^{-1}+1}\left(\frac{\lambda\left(\mathbf{x_i}\right)}{\alpha^{-1}+\lambda\left(\mathbf{x_i}\right)}\right)^{y_i}}\mathbf{x_i}=\mathbf{0}$$

$$\sum_{i=1}^{n}\frac{f\left(\mathbf{z_i}'\gamma\right)\left(y_i!d_i-\prod_{j=0}^{y_i-1}\left(j+\alpha^{-1}\right)\left(\frac{\alpha^{-1}}{\alpha^{-1}+\lambda\left(\mathbf{x_i}\right)}\right)^{\alpha^{-1}+1}\left(\frac{\lambda\left(\mathbf{x_i}\right)}{\alpha^{-1}+\lambda\left(\mathbf{x_i}\right)}\right)^{y_i}\right)}{y_i!d_i\psi\left(\mathbf{z_i}\right)+\left(1-\psi\left(\mathbf{z_i}\right)\right)\prod_{j=0}^{y_i-1}\left(j+\alpha^{-1}\right)\left(\frac{\alpha^{-1}}{\alpha^{-1}+\lambda\left(\mathbf{x_i}\right)}\right)^{\alpha^{-1}+1}\left(\frac{\lambda\left(\mathbf{x_i}\right)}{\alpha^{-1}+\lambda\left(\mathbf{x_i}\right)}\right)^{y_i}}\mathbf{z_i}=\mathbf{0}$$

$$\sum_{i=1}^{n}\frac{\left(1-\psi(\mathbf{z_i})\right)\prod_{j=0}^{y_i-1}\left(j+\alpha^{-1}\right)\left(\frac{\alpha^{-1}}{\alpha^{-1}+\lambda(\mathbf{x_i})}\right)^{\alpha^{-1}+1}\left(\frac{\lambda(\mathbf{x_i})}{\alpha^{-1}+\lambda(\mathbf{x_i})}\right)^{y_i}}{y_i!d_i\psi(\mathbf{z_i})+\left(1-\psi(\mathbf{z_i})\right)\prod_{j=0}^{y_i-1}\left(j+\alpha^{-1}\right)\left(\frac{\alpha^{-1}}{\alpha^{-1}+\lambda(\mathbf{x_i})}\right)^{\alpha^{-1}+1}\left(\frac{\lambda(\mathbf{x_i})}{\alpha^{-1}+\lambda(\mathbf{x_i})}\right)^{y_i}}$$

$$\times\alpha^{-2}\left(\frac{y_i-\lambda(\mathbf{x_i})}{\alpha^{-1}+\lambda(\mathbf{x_i})}-\sum_{j=0}^{y_i-1}\frac{1}{j+\alpha^{-1}}\right)=\mathbf{0}$$

em que $f\left(\mathbf{z_i}'\gamma\right)$ é dado em (2.14)

2.4.3 Propriedades

Se a especificação do modelo for a correcta e verificadas certas condições de regularidade, tem-se o seguinte resultado assintótico:

$$\begin{bmatrix}\hat{\beta}\\ \hat{\gamma}\\ \hat{\alpha}\end{bmatrix}\overset{a}{\sim}N\left(\begin{bmatrix}\beta\\ \gamma\\ \alpha\end{bmatrix},\begin{bmatrix}Var\left[\hat{\beta}\right] & Cov\left[\hat{\beta},\hat{\gamma}\right] & Cov\left[\hat{\beta},\hat{\alpha}\right]\\ & Var\left[\hat{\gamma}\right] & Cov\left[\hat{\gamma},\hat{\alpha}\right]\\ & & Var\left[\hat{\alpha}\right]\end{bmatrix}\right)$$

com

$$Var\left[\hat{\beta}\right]=\left(\sum_{i=1}^{n}\left[\left(\frac{\left(1-\psi(\mathbf{z_i})\right)\prod_{j=0}^{y_i-1}\left(j+\alpha^{-1}\right)\left(\frac{\alpha^{-1}}{\alpha^{-1}+\lambda(\mathbf{x_i})}\right)^{\alpha^{-1}+1}\left(\frac{\lambda(\mathbf{x_i})}{\alpha^{-1}+\lambda(\mathbf{x_i})}\right)^{y_i}\left(y_i-\lambda(\mathbf{x_i})\right)}{y_i!d_i\psi(\mathbf{z_i})+\left(1-\psi(\mathbf{z_i})\right)\prod_{j=0}^{y_i-1}\left(j+\alpha^{-1}\right)\left(\frac{\alpha^{-1}}{\alpha^{-1}+\lambda(\mathbf{x_i})}\right)^{\alpha^{-1}+1}\left(\frac{\lambda(\mathbf{x_i})}{\alpha^{-1}+\lambda(\mathbf{x_i})}\right)^{y_i}}\right)^2\right.\right.$$

$$-\frac{\left(1-\psi(\mathbf{z_i})\right)\prod_{j=0}^{y_i-1}\left(j+\alpha^{-1}\right)\left(\frac{\alpha^{-1}}{\alpha^{-1}+\lambda(\mathbf{x_i})}\right)^{\alpha^{-1}+1}\left(\frac{\lambda(\mathbf{x_i})}{\alpha^{-1}+\lambda(\mathbf{x_i})}\right)^{y_i+1}}{y_i!d_i\psi(\mathbf{z_i})+\left(1-\psi(\mathbf{z_i})\right)\prod_{j=0}^{y_i-1}\left(j+\alpha^{-1}\right)\left(\frac{\alpha^{-1}}{\alpha^{-1}+\lambda(\mathbf{x_i})}\right)^{\alpha^{-1}+1}\left(\frac{\lambda(\mathbf{x_i})}{\alpha^{-1}+\lambda(\mathbf{x_i})}\right)^{y_i}}$$

$$\left.\left.\times\left(\frac{\alpha^{-1}}{\lambda(\mathbf{x_i})}\left(y_i-\lambda(\mathbf{x_i})\right)^2-y_i-\alpha^{-1}\right)\right]\mathbf{x_i}\mathbf{x_i}'\right)^{-1}$$

$$Var[\hat{\gamma}] = \left(\sum_{i=1}^{n} \left[\left(\frac{f\left(\mathbf{z_i}'\gamma\right)\left(y_i!d_i - \prod_{j=0}^{y_i-1}\left(j+\alpha^{-1}\right)\left(\frac{\alpha^{-1}}{\alpha^{-1}+\lambda(\mathbf{x_i})}\right)^{\alpha^{-1}}\left(\frac{\lambda(\mathbf{x_i})}{\alpha^{-1}+\lambda(\mathbf{x_i})}\right)^{y_i}\right)}{y_i!d_i\psi(\mathbf{z_i})+(1-\psi(\mathbf{z_i}))\prod_{j=0}^{y_i-1}\left(j+\alpha^{-1}\right)\left(\frac{\alpha^{-1}}{\alpha^{-1}+\lambda(\mathbf{x_i})}\right)^{\alpha^{-1}}\left(\frac{\lambda(\mathbf{x_i})}{\alpha^{-1}+\lambda(\mathbf{x_i})}\right)^{y_i}} \right]^2 \right.$$

$$\left. - \frac{f'\left(\mathbf{z_i}'\gamma\right)\left(y_i!d_i - \prod_{j=0}^{y_i-1}\left(j+\alpha^{-1}\right)\left(\frac{\alpha^{-1}}{\alpha^{-1}+\lambda(\mathbf{x_i})}\right)^{\alpha^{-1}}\left(\frac{\lambda(\mathbf{x_i})}{\alpha^{-1}+\lambda(\mathbf{x_i})}\right)^{y_i}\right)}{y_i!d_i\psi(\mathbf{z_i})+(1-\psi(\mathbf{z_i}))\prod_{j=0}^{y_i-1}\left(j+\alpha^{-1}\right)\left(\frac{\alpha^{-1}}{\alpha^{-1}+\lambda(\mathbf{x_i})}\right)^{\alpha^{-1}}\left(\frac{\lambda(\mathbf{x_i})}{\alpha^{-1}+\lambda(\mathbf{x_i})}\right)^{y_i}} \mathbf{z_i}\mathbf{z_i}' \right)^{-1}$$

$$Var[\hat{\alpha}] = \left(\sum_{i=1}^{n} \left\{ \left(\frac{(1-\psi(\mathbf{z_i}))\prod_{j=0}^{y_i-1}\left(j+\alpha^{-1}\right)\left(\frac{\alpha^{-1}}{\alpha^{-1}+\lambda(\mathbf{x_i})}\right)^{\alpha^{-1}}\left(\frac{\lambda(\mathbf{x_i})}{\alpha^{-1}+\lambda(\mathbf{x_i})}\right)^{y_i}}{y_i d_i\psi(\mathbf{z_i})+(1-\psi(\mathbf{z_i}))\prod_{j=0}^{y_i-1}\left(j+\alpha^{-1}\right)\left(\frac{\alpha^{-1}}{\alpha^{-1}+\lambda(\mathbf{x_i})}\right)^{\alpha^{-1}}\left(\frac{\lambda(\mathbf{x_i})}{\alpha^{-1}+\lambda(\mathbf{x_i})}\right)^{y_i}} \right.\right.$$

$$\times \left. \alpha^{-2}\left(\frac{y_i-\lambda(\mathbf{x_i})}{\alpha^{-1}+\lambda(\mathbf{x_i})} - \sum_{j=0}^{y_i-1}\frac{1}{j+\alpha^{-1}} \right) \right)^2$$

$$- \frac{(1-\psi(\mathbf{z_i}))\prod_{j=0}^{y_i-1}\left(j+\alpha^{-1}\right)\left(\frac{\alpha^{-1}}{\alpha^{-1}+\lambda(\mathbf{x_i})}\right)^{\alpha^{-1}}\left(\frac{\lambda(\mathbf{x_i})}{\alpha^{-1}+\lambda(\mathbf{x_i})}\right)^{y_i}}{y_i d_i\psi(\mathbf{z_i})+(1-\psi(\mathbf{z_i}))\prod_{j=0}^{y_i-1}\left(j+\alpha^{-1}\right)\left(\frac{\alpha^{-1}}{\alpha^{-1}+\lambda(\mathbf{x_i})}\right)^{\alpha^{-1}}\left(\frac{\lambda(\mathbf{x_i})}{\alpha^{-1}+\lambda(\mathbf{x_i})}\right)^{y_i}}$$

$$\times \left[\alpha^{-4}\left(\frac{y_i-\lambda(\mathbf{x_i})}{\alpha^{-1}+\lambda(\mathbf{x_i})} - \sum_{j=0}^{y_i-1}\frac{1}{j+\alpha^{-1}} \right)^2 \right.$$

$$\left.\left. - \frac{y_i-\lambda(\mathbf{x_i})}{\left(\alpha^{-1}+\lambda(\mathbf{x_i})\right)^2}\alpha^{-3}\left(2\lambda(\mathbf{x_i})+\alpha^{-1}\right) + \sum_{j=0}^{y_i-1}\frac{\alpha^{-3}}{\left(j+\alpha^{-1}\right)^2}\left(2j+\alpha^{-1}\right) \right] \right\} \right)^{-1}$$

e as covariâncias:

$$Cov\left[\hat{\beta},\hat{\gamma}\right]=$$

$$\left(\sum_{i=1}^{n} f\left(\mathbf{z_i}'\gamma\right)\left(\frac{\left(1-\psi(\mathbf{z_i})\right)\prod_{j=0}^{y_i-1}\left(j+\alpha^{-1}\right)\left(\frac{\alpha^{-1}}{\alpha^{-1}+\lambda(\mathbf{x_i})}\right)^{\alpha^{-1}+1}\left(\frac{\lambda(\mathbf{x_i})}{\alpha^{-1}+\lambda(\mathbf{x_i})}\right)^{y_i}\left(y_i-\lambda(\mathbf{x_i})\right)}{\left((y_i)!d_i\psi(\mathbf{z_i})+\left(1-\psi(\mathbf{z_i})\right)\prod_{j=0}^{y_i-1}\left(j+\alpha^{-1}\right)\left(\frac{\alpha^{-1}}{\alpha^{-1}+\lambda(\mathbf{x_i})}\right)^{\alpha^{-1}+1}\left(\frac{\lambda(\mathbf{x_i})}{\alpha^{-1}+\lambda(\mathbf{x_i})}\right)^{y_i}\right)^2}\right.$$

$$\times\left(y_i!d_i-\prod_{j=0}^{y_i-1}\left(j+\alpha^{-1}\right)\left(\frac{\alpha^{-1}}{\alpha^{-1}+\lambda(\mathbf{x_i})}\right)^{\alpha^{-1}+1}\left(\frac{\lambda(\mathbf{x_i})}{\alpha^{-1}+\lambda(\mathbf{x_i})}\right)^{y_i}\right)$$

$$\left.+\frac{\prod_{j=0}^{y_i-1}\left(j+\alpha^{-1}\right)\left(\frac{\alpha^{-1}}{\alpha^{-1}+\lambda(\mathbf{x_i})}\right)^{\alpha^{-1}+1}\left(\frac{\lambda(\mathbf{x_i})}{\alpha^{-1}+\lambda(\mathbf{x_i})}\right)^{y_i}}{(y_i)!d_i\psi(\mathbf{z_i})+\left(1-\psi(\mathbf{z_i})\right)\prod_{j=0}^{y_i-1}\left(j+\alpha^{-1}\right)\left(\frac{\alpha^{-1}}{\alpha^{-1}+\lambda(\mathbf{x_i})}\right)^{\alpha^{-1}+1}\left(\frac{\lambda(\mathbf{x_i})}{\alpha^{-1}+\lambda(\mathbf{x_i})}\right)^{y_i}}\mathbf{x_i}\mathbf{z_i}'\right)^{-1}$$

$$Cov\left[\hat{\beta},\hat{\alpha}\right]=$$

$$\left(\sum_{i=1}^{n}\left(\frac{\left(1-\psi(\mathbf{z_i})\right)\prod_{j=0}^{y_i-1}\left(j+\alpha^{-1}\right)\left(\frac{\alpha^{-1}}{\alpha^{-1}+\lambda(\mathbf{x_i})}\right)^{\alpha^{-1}+1}\left(\frac{\lambda(\mathbf{x_i})}{\alpha^{-1}+\lambda(\mathbf{x_i})}\right)^{y_i}\left(y_i-\lambda(\mathbf{x_i})\right)}{y_i!d_i\psi(\mathbf{z_i})+\left(1-\psi(\mathbf{z_i})\right)\prod_{j=0}^{y_i-1}\left(j+\alpha^{-1}\right)\left(\frac{\alpha^{-1}}{\alpha^{-1}+\lambda(\mathbf{x_i})}\right)^{\alpha^{-1}+1}\left(\frac{\lambda(\mathbf{x_i})}{\alpha^{-1}+\lambda(\mathbf{x_i})}\right)^{y_i}}\right.\right.$$

$$\times\left(\frac{\alpha^{-2}\left(y_i-\lambda(\mathbf{x_i})\right)-\lambda(\mathbf{x_i})\alpha^{-1}}{\alpha^{-1}+\lambda(\mathbf{x_i})}-\sum_{j=0}^{y_i-1}\frac{\alpha^2}{j+\alpha^{-1}}\right)$$

$$-\frac{\left(1-\psi(\mathbf{z_i})\right)\prod_{j=0}^{y_i-1}\left(j+\alpha^{-1}\right)\left(\frac{\alpha^{-1}}{\alpha^{-1}+\lambda(\mathbf{x_i})}\right)^{\alpha^{-1}+1}\left(\frac{\lambda(\mathbf{x_i})}{\alpha^{-1}+\lambda(\mathbf{x_i})}\right)^{y_i}\left(y_i-\lambda(\mathbf{x_i})\right)}{\left(y_i!d_i\psi(\mathbf{z_i})+\left(1-\psi(\mathbf{z_i})\right)\prod_{j=0}^{y_i-1}\left(j+\alpha^{-1}\right)\left(\frac{\alpha^{-1}}{\alpha^{-1}+\lambda(\mathbf{x_i})}\right)^{\alpha^{-1}+1}\left(\frac{\lambda(\mathbf{x_i})}{\alpha^{-1}+\lambda(\mathbf{x_i})}\right)^{y_i}\right)^2}$$

$$\times\left(1-\psi(\mathbf{z_i})\right)\prod_{j=0}^{y_i-1}\left(j+\alpha^{-1}\right)\left(\frac{\alpha^{-1}}{\alpha^{-1}+\lambda(\mathbf{x_i})}\right)^{\alpha^{-1}+1}\left(\frac{\lambda(\mathbf{x_i})}{\alpha^{-1}+\lambda(\mathbf{x_i})}\right)^{y_i}$$

$$\left.\left.\times\alpha^{-2}\left(\frac{y_i-\lambda(\mathbf{x_i})}{\alpha^{-1}+\lambda(\mathbf{x_i})}-\sum_{j=0}^{y_i-1}\frac{1}{j+\alpha^{-1}}\right)\right)\mathbf{x_i}\right)^{-1}$$

$$Cov[\hat{\gamma},\hat{\alpha}] =$$

$$\left(\sum_{i=1}^{n} \left(\frac{f\left(\mathbf{z_i}'\gamma\right)\prod_{j=0}^{y_i-1}\left(j+\alpha^{-1}\right)\left(\frac{\alpha^{-1}}{\alpha^{-1}+\lambda(\mathbf{x_i})}\right)^{\alpha^{-1}}\left(\frac{\lambda(\mathbf{x_i})}{\alpha^{-1}+\lambda(\mathbf{x_i})}\right)^{y_i}}{y_i!d_i\psi(\mathbf{z_i})+(1-\psi(\mathbf{z_i}))\prod_{j=0}^{y_i-1}\left(j+\alpha^{-1}\right)\left(\frac{\alpha^{-1}}{\alpha^{-1}+\lambda(\mathbf{x_i})}\right)^{\alpha^{-1}}\left(\frac{\lambda(\mathbf{x_i})}{\alpha^{-1}+\lambda(\mathbf{x_i})}\right)^{y_i}} \right.\right.$$

$$- \frac{f\left(\mathbf{z_i}'\gamma\right)\left(y_i!d_i-\prod_{j=0}^{y_i-1}\left(j+\alpha^{-1}\right)\left(\frac{\alpha^{-1}}{\alpha^{-1}+\lambda(\mathbf{x_i})}\right)^{\alpha^{-1}}\left(\frac{\lambda(\mathbf{x_i})}{\alpha^{-1}+\lambda(\mathbf{x_i})}\right)^{y_i}\right)}{\left(y_i!d_i\psi(\mathbf{z_i})+(1-\psi(\mathbf{z_i}))\prod_{j=0}^{y_i-1}\left(j+\alpha^{-1}\right)\left(\frac{\alpha^{-1}}{\alpha^{-1}+\lambda(\mathbf{x_i})}\right)^{\alpha^{-1}}\left(\frac{\lambda(\mathbf{x_i})}{\alpha^{-1}+\lambda(\mathbf{x_i})}\right)^{y_i}\right)^2}$$

$$\left.\times \left(1-\psi(\mathbf{z_i})\right)\prod_{j=0}^{y_i-1}\left(j+\alpha^{-1}\right)\left(\frac{\alpha^{-1}}{\alpha^{-1}+\lambda(\mathbf{x_i})}\right)^{\alpha^{-1}}\left(\frac{\lambda(\mathbf{x_i})}{\alpha^{-1}+\lambda(\mathbf{x_i})}\right)^{y_i}\right)$$

$$\times \alpha^{-2}\left(\sum_{j=0}^{y_i-1}\frac{1}{j+\alpha^{-1}}-\frac{y_i-\lambda(\mathbf{x_i})}{\alpha^{-1}+\lambda(\mathbf{x_i})}\right)\mathbf{z_i}\Bigg)^{-1}$$

3. Principais Abordagens Semi- ou Não Paramétricas

A análise de regressão não paramétrica permite uma primeira abordagem exploratória e flexível do ponto de vista das hipóteses postuladas sobre a distribuição da contagem, sendo adequada quando o conhecimento sobre esta é escasso ou nulo. É vasta a obra neste campo, e existem diversos métodos de análise de regressão não paramétrica. A bibliografia sobre técnicas de alisamento não paramétricas é também muito vasta, (cf. Hall, 2000) de que é apresentada uma amostra em Santos (2005). De entre os métodos de regressão não paramétrica, destacam-se os alisadores de núcleo, a regressão polinomial local, os alisadores de *spline*, alisadores de séries ortogonais, alisadores de *wavelets*, a quasiverosimilhança e a verosimilhança local. Neste capítulo são apresentadas as três primeiras metodologias. Tem sido objecto de trabalho o desenvolvimento de modelos que combinam as abordagens paramétrica e não paramétrica, em formulações designadas semi-paramétricas, de grande interesse e aplicabilidade práticos, em que se desenvolve uma dada especificação de um modelo, com incorporação de metodologias cuja abordagem não estabelece condições *a priori*. Esta combinação conseguida nas formulações semi-paramétricas, pode ser efectuada de diversas formas, nomeadamente através do método dos momentos, da verosimilhança penalizada, a quasi-verosimilhança (Wedderburn (1974)), quasi-verosimilhança estendida (Nelder e Pregibon (1987)), a pseudoverosimilhança (Besag (1975)): Sun (2001) Podlich, Faddy e Smyth (2004) e mais recentemente, Zhao e Sun (2006), Dean, Nathoo e Nielsen (2007), Nielsen e Dean (2008), Santos (2005) e Santos e Neves (2008). Saha (2008) apresentam metodologias de estimação semi-paramétricas para análise de dados de contagem, na presença de sub e sobredispersão.

Se, pelo modelo de regressão não paramétrico mais simples, pretendermos modelar a relação, no caso univariado, entre uma variável resposta Y, e um regressor X,

$$Y_i = m(x_i) + \varepsilon_i \tag{3.1}$$

em que $m(x_i)$ é a média condicional ao regressor,

$$m(x_i) = E\left[Y \middle| X = x_i\right], \tag{3.2}$$

e ε_i é o erro aleatório, que respeita as seguintes condições:

$$E\left[\varepsilon_i \middle| X = x_i\right] = 0 \text{ e} \tag{3.3}$$

$$Var\left[\varepsilon_i \middle| X = x_i\right] = \sigma^2(x_i), \tag{3.4}$$

em que X e ε são assumidos como independentes, sem prejuízo de generalidade, e ε, como se pode ver, não é necessariamente constante. A estimação dos parâmetros dos modelos é feita sem assumir quaisquer pressupostos. No caso de modelos semi-paramétricos, faz-se apenas alguns pressupostos, combinando algum modelo paramétrico de base com a flexibilidade da abordagem não paramétrica. Kohler e Krzyzak (2007) desenvolvem um alisador de núcleo por regressão polinomial de Poisson, e desenvolvem estimativas assintoticamente convergentes para os intervalos de confiança.

Na vertente não paramétrica, Mulkhophadyay e Marsh (2006) introduzem um novo alisador de núcleo de Poisson e um alisador de núcleo binomial. Uma revisão de metodologias pode ser vista em Zhao e Sun (2006).

Para sublinhar a crescente importância de trabalhos de natureza semi-paramétrica, desenvolvidos por métodos de verosimilhança penalizada, quasi-

verosimilhança, quasi-verosimilhança estendida e a pseudoverosimilhança já referidos anteriormente, em trabalhos recentes, Antoniadis *et al* (2011) apresentam modelos lineares generalizados estimados por regressão de verosimilhança penalizada, com aplicação na actividade seguradora, Silva e Tenreyro (2011) mostram, no âmbito da regressão de Poisson, que o estimador de pseudoverosimilhança de Poisson tem bom desempenho, mesmo na presença elevada de zeros.

3.1 Regressão de Núcleo

A regressão com alisadores de núcleo é a mais simples e mais usada. Baseia-se no princípio de que as observações mais próximas do ponto de interesse onde a função deve ser estimada têm também médias mais próximas de cada ponto xi, e como tal, devem ter maior ponderação na estimação da variável resposta. Da equação (3.2) tínhamos:

$$\begin{aligned} m(x_i) &= E\left[Y \middle| X = x_i\right] \\ &= \int_{-\infty}^{+\infty} y f_{Y|X=x}(y)\, dy \\ &= \int_{-\infty}^{+\infty} y \frac{f_{(X,Y)}(x,y)}{f_X(x)} dy, \end{aligned}$$

correcta especificação do modelo em que uma uma estimativa para a função marginal de X é

$$\hat{f}_X(x) = \frac{1}{nh_x} \sum_{i=1}^{n} K_x\left(\frac{x - x_i}{h_x}\right) \tag{3.5}$$

e para a função densidade conjunta de X e Y, por produto de núcleos é

$$\hat{f}_{(X,Y)}(x,y) = \frac{1}{nh_x h_y} \sum_{i=1}^{n} K_x\left(\frac{x - x_i}{h_x}\right) K_y\left(\frac{y - y_i}{h_y}\right), \quad (3.6)$$

em que h é a largura de banda, que deverá ser convenientemente escolhida. O alisador atribui pesos às observações x_i=1,…,n, consoante a sua proximidade ao ponto de interesse x, reduzindo o seu peso à medida que aumenta a distância a este.

- Valores de h_x pequenos resultam em alisadores com elevada variância local (maior rugosidade) e reduzido viés.
- Valores elevados de h_x resultam numa diminuição menos rápida dos pesos diminuem, resultando numa menor variância local (maior alisamento), e elevado viés.

A escolha da largura de banda é então um compromisso entre viés e variância, por um lado, e alisamento e rugosidade, por outro. A função núcleo é real, contínua, limitada e simétrica, e verifica:

$$\int K(z)\,dz = 1$$

Alguns exemplos de funções núcleo são apresentados na Tabela 2). Todos têm suporte [-1,1] à excepção do núcleo Gaussiano, com suporte]-∞,+∞[. Para todas as funções núcleo à excepção desta, se a largura de banda for fixa, o número de observações consideradas a cada estimativa é variável, consoante a densidade de observações na vizinhança de cada ponto de interesse (se a distribuição não for uniforme). Podemos considerar, em alternativa a largura de banda variável $h(x)$ em que o número de observações em torno de cada ponto de interesse x, $h(x): \#\{x_j \in [x - h(x), x + h(x)]\} = m$, seja fixo.

Núcleo	Forma
Epanechnikov	$\frac{3}{4}\left(1-z^2\right)$
Biweight	$\frac{15}{16}\left(1-z^2\right)^2$
Triweight	$\frac{35}{32}\left(1-z^2\right)^3$
Gauss	$\frac{1}{\sqrt{2\pi}}e^{-\frac{z^2}{2}}$
Uniforme	$\frac{1}{2}$

Tabela 1) Núcleos mais usados e sua forma

Os métodos automáticos de selecção da largura de banda h, parâmetro que controla o alisamento do estimador, dividem-se essencialmente em dois grupos: a abordagem clássica e a abordagem *plug-in*:

- A abordagem clássica centra-se na minimização da média do erro quadrático médio, ou de um estimador aproximadamente centrado desta:

$$\frac{1}{n}E\left[\left(\hat{\mathbf{m}}_{\mathbf{h}}-\mathbf{m}\right)'\left(\hat{\mathbf{m}}_{\mathbf{h}}-\mathbf{m}\right)\right],$$

com $\mathbf{m}=\left(m(x_1),\mathrm{K}\,,m(x_n)\right)'$, ou na minimização de:

$\log\left(\hat{\sigma}^2\right)+\psi\left(H\right)$

com $\hat{\sigma}^2=\frac{1}{n}\sum_{i=1}^{n}\left[y_i-\hat{m}_h\left(x_i\right)\right]^2$ e ψ é uma função penalizadora de rugosidade que decresce com o alisamento de $\hat{\mathbf{m}}_{\mathbf{h}}$.

- A abordagem *plug-in* assenta na minimização da média condicional ponderada do erro quadrático médio integrado (MISE):

$$E\left[\int\left\{\hat{m}(u)-m(u)\right\}^2 f_X(u)\,du\,\middle|\,x_1,\mathrm{K}\,,x_n\right]$$

em que $f_X(x)$ é a função densidade do delineamento experimental. A solução assintótica para o alisador linear local é dada por:

$$h=\left[\frac{R(K)\sigma_0^2}{N\mu_2(K)^2\int m''(u)^2\,f_X(u)\,du}\right]^{\frac{1}{5}} \qquad (3.7)$$

em que $R(K)=\int K(u)^2\,du$ e $\mu_2(K)=\int u^2K(u)\,du$. A largura de banda *plug-in* é dada por (3.7) com σ_0^2 e $\int m''(u)^2\,f_X(u)\,du$ dados pelas suas estimativas.

O estimador de núcleo de Nadarayda-Watson obtém-se substituindo (3.5) e (3.6) em (3.2), que corresponde a uma combinação linear das variáveis resposta:

$$\hat{m}(x)=\frac{\sum_{i=1}^{n}y_iK\left(\frac{x-x_i}{h}\right)}{\sum_{i=1}^{n}K\left(\frac{x-x_i}{h}\right)}=\sum_{i=1}^{n}y_iw_i\,, \qquad (3.8)$$

com os pesos dados por $w_i = \frac{K\left(\frac{x - x_i}{h}\right)}{\sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right)}$.

3.2 Regressão Polinomial Local

A regressão polinomial local surge inicialmente pelos trabalhos de Stone (1977) e Cleveland (1979). Wand *et al* (1995) apresentam trabalho mais aprofundado neste tema.

A regressão de núcleo do ponto anterior pode ser expandida ao ajustamento polinomial local num ponto de interesse x, com os ponderadores dados pela função núcleo. Estes pontos de interesse x poderão ser quaisquer valores do suporte do regressor, e não necessariamente os pontos da amostra, embora na prática, por conveniência, se façam coincidir os pontos de interesse x com os pontos amostrais.

A regressão polinomial de ordem p a estimar é:

$$y_i = \beta_0 + \beta_1 (x_i - x) + \mathrm{K} + \beta_p (x_i - x)^p + \varepsilon_i, \tag{3.9}$$

através dos mínimos quadrados ponderados pela função núcleo, que minimizam o erros aleatórios:

$$h^{-1} \sum_{i=1}^{n} \varepsilon_i^2 K\left\{\frac{x_i - x}{h}\right\}, \tag{3.10}$$

ou seja,

$$h^{-1} \sum_{i=1}^{n} \left[y_i - \beta_0 - \beta_1 (x_i - x) - \mathrm{K} - \beta_p (x_i - x)^p \right]^2 K\left\{\frac{x - x_i}{h}\right\}.$$

A matriz dos regressores é:

$$\mathbf{X}_x = \begin{bmatrix} 1 & x_1 - x & \text{L} & (x_1 - x)^p \\ \text{M} & \text{M} & \text{O} & \text{M} \\ 1 & x_n - x & \text{L} & (x_n - x)^p \end{bmatrix}$$

e a matriz dos ponderadores

$$\text{´} \qquad \mathbf{W}_x = h^{-1} \begin{bmatrix} K\left\{\frac{x - x_1}{h}\right\} & & \mathbf{0} \\ & \text{O} & \\ \mathbf{0} & & K\left\{\frac{x - x_p}{h}\right\} \end{bmatrix}.$$

Podemos reescrever (3.9) em forma matricial:

$$h^{-1} \sum_{i=1}^{n} \varepsilon_i^2 K\left\{\frac{x_i - x}{h}\right\} = \varepsilon' \mathbf{W}_x \varepsilon = \left(\mathbf{y} - \mathbf{X}_x \beta\right)' \mathbf{W}_x \left(\mathbf{y} - \mathbf{X}_x \beta\right).$$

Se a matriz $\mathbf{X}_x{}' \mathbf{W}_x \mathbf{X}_x$ for invertível, a solução $\hat{\beta}_{x;p,h}$ que minimiza os mínimos quadrados de (3.9) é dada por:

$$\hat{\beta}_{x;p,h} = \left(\mathbf{X}_x{}' \mathbf{W}_x \mathbf{X}_x\right)^{-1} \mathbf{X}_x{}' \mathbf{W}_x \mathbf{y}. \tag{3.11}$$

O alisador de regressão polinomial local de ordem p é dado pelo estimador da constante β_0 da equação (3.8) dado por:

$$\hat{m}_p(x;h) = \mathbf{e_1}' \left(\mathbf{X}_x{}' \mathbf{W}_x \mathbf{X}_x\right)^{-1} \mathbf{X}_x{}' \mathbf{W}_x \mathbf{y} \tag{3.12}$$

em que $\mathbf{e_1}$ é um vector de ordem p+1 com 1 na primeira entrada e 0 nas restantes.

Na regressão polinomial local, o analista selecciona o grau do polinómio ajustado localmente, a função de núcleo $\boldsymbol{K}(z)$, a largura de banda h e, no caso robusto, o número de iterações. A escolha de um maior grau de ajustamento tem o mesmo efeito no alisador que a de uma maior largura de banda, isto é, menor viés e maior variância. O alisador de núcleo, exposto no ponto anterior, é um caso particular deste, fazendo p=0.

3.3 Alisadores de *Spline*

Os alisadores de *Spline* são apresentados em vários trabalhos na literatura. Uma exposição sucinta pode ser vista em Aydin (2007) [a], a qual será adaptada para esta exposição. Uma exposição detalhada pode ser vista em Wahba (1990), Green e Silverman (1994), Eubank (1999), Santos (2005) ou Aydin (2007) [b].

A metodologia consiste na solução do problema de minimização da função:

$$S(f) = \sum_{i=1}^{n} \{y_i - f(x_i)\}^2 + \alpha \int_a^b \{f''(x)\}^2 \, dx \qquad (3.13)$$

em que f na equação acima corresponde ao termo $m(x)$ da equação (3.1) na notação deste trabalho. O estimador da função, bem como a função, devem respeitar a condição $\hat{f} \in C^2[a;b]$, isto é, devem ser diferenciáveis duas vezes no intervalo, e $f''(x)^2$ integráveis neste intervalo.

- O primeiro termo da equação (3.13) corresponde à soma dos quadrados dos resíduos, e penaliza a "má qualidade do ajustamento" ("*lack of fit*", expressão original).

- O segundo termo desta equação, ponderado pelo factor α (denominado parâmetro de alisamento, ou "*smoothing parameter*", expressão original) penaliza a rugosidade de f.

O parâmetro α varia entre 0 e +∞. À medida que α cresce, a solução varia da interpolação para um modelo linear:

- Se $\alpha \to 0$, a penalização da rugosidade desaparece e o alisador comporta-se como interpolador linear dos dados (segmentos de recta).

- Se $\alpha \to \infty$, a penalização da rugosidade predomina sobre o resto e o alisador será ($\alpha \int_a^b \{ f''(x) \}^2 dx$). A curvatura exibida demonstrará completo alisamento (com prejuízo da aderência aos dados).

A escolha de α é portanto um compromisso entre a qualidade do ajustamento, (medida pela soma dos quadrados dos resíduos, $\sum_{i=1}^{n} \{ y_i - f(x_i) \}^2$) e o alisamento da estimativa (medido por $\int_a^b \{ f''(x) \}^2 dx$).

Sejam, para um dado α escolhido, $\mathbf{f} = \left(f(x_1), \mathrm{K}, f(x_n) \right)'$ o vector de valores de f (f nas codições especificadas acima) nos nós $x_1,\ldots,x_n$, e $\mathbf{y} = \left(y_1, \mathrm{K}, y_n \right)'$ o vector das respostas e ainda $\mathbf{S}_\alpha$ uma matriz semi-definida positiva (simétrica) conhecida (especificada em detalhe na literatura referida acima), que depende dos nós $x_1,\ldots,x_n$, e de α, mas não de $\mathbf{y}$. Então, a estimativa $\hat{\mathbf{f}}_\alpha$ é dada por:

$$\underset{(n\times 1)}{\hat{\mathbf{f}}_\alpha} = \underset{(n\times 1)}{\begin{bmatrix} \hat{f}_\alpha(x_1) \\ \hat{f}_\alpha(x_2) \\ \mathrm{M} \\ \hat{f}_\alpha(x_n) \end{bmatrix}} = \underset{(n\times n)}{\mathbf{S}_\alpha} . \underset{(n\times 1)}{\begin{bmatrix} y_1 \\ y_2 \\ \mathrm{M} \\ y_n \end{bmatrix}} \quad \text{ou} \quad \hat{\mathbf{f}}_\alpha = \mathbf{S}_\alpha . \mathbf{y}$$

A solução da minimização de (3.13), $\hat{\mathbf{f}}_{\alpha}$, existe e é única (cf. Santos (2005) pp 52, a equação acima, com a devida adaptação de notação, é equivalente à equação (4.32) daquele trabalho), denominada ***spline cúbico natural***, com nós em $x_1,\ldots,x_n$.

4. Formulação de um Modelo de Máxima Verosimilhança Local (MVL) de Poisson

O conceito de máxima verosimilhança local (MVL), metodologia que permite associar a análise de regressão não paramétrica, nomeadamente a regressão polinomial local, a modelos assentes no princípio da máxima verosimilhança, estendida ao conceito local (Eguchi *et al* (2003)), foi apresentado por Tibshirani e Hastie (1987). Em Fan *et al* (1998), é apresentado um método que permite estimar o seu viés, e a sua variância, bem como um método para a selecção da largura de banda. Em Santos (2005) são propostas diversas formulações semi-paramétricas, com base no modelo de regressão de Poisson, e no modelo de regressão binomial negativo, e até no modelo logístico. Em particular, é proposto o alisador de máxima verosimilhança local, com o modelo de Poisson (ponto 5.2 pp 65-70), sendo deduzidos o seu viés, variância e distribuição assintótica. Este método, independentemente do ajustamento conseguido, revela-se computacionalmente muito pesado, uma vez que se trata de uma especificação exponencial, dificultando, nomeadamente, a determinação da solução computacional das condições de primeira ordem de maximização da função log-verosimilhança local (equações (5.14) pp 66), de que resulta o estimador do viés e da variância.

Neste trabalho propõe-se uma especificação alternativa, tendo como base a verosimilhança local, o alisador de núcleo e a regressão polinomial local, que pode ser entendida como uma simplificação da anterior, proposta por Santos (2005). A ideia subjacente à modelação alternativa que se vai desenvolver, reside no facto de, na especificação do valor médio, $\lambda(\mathrm{x})$, a exponencial de uma função desconhecida ser, ela própria, uma função desconhecida. Faz, então, sentido especificar $\lambda(\mathrm{x})$ como uma função polinomial, mais simples de tratar e computacionalmente menos exigente.

O conceito de verosimilhança local, bem como o do alisador de máxima verosimilhança local, nos quais assenta toda esta formulação, são explicados no ponto 5.1 de Santos (2005) (pp 58-64). A especificação, dedução do viés, da variância e distribuição assintótica seguem o exposto nesse subcapítulo.

4.1 Especificação do modelo

O valor médio segue uma especificação por regressão polinomial:

$$\lambda(\mathbf{x_i}) \approx \beta_0 + \beta_1 (x_i - x) \tag{4.1}$$

e a função massa de probabilidade é dada por:

$$P(Y_i = y_i | \mathbf{x_i}) = \frac{e^{-\lambda(\mathbf{x_i})} \left[\lambda(\mathbf{x_i})\right]^{y_i}}{y_i!}, \quad y = 0,1,2,\mathrm{K}\ . \tag{4.2}$$

Como se pode ver, o modelo possui uma base paramétrica de Poisson, porém o valor médio é estimado através de regressão polinomial local. O conceito de verosimilhança local, bem como o do alisador de máxima verosimilhança local, nos quais assenta toda esta formulação, são introduzidos e explicados no capítulo 5.1 de Santos (2005).

A função de logverosimilhança local, no ponto de interesse x, é dada por:

$$\mathcal{L}(\beta_0, \beta_1 | \mathbf{x}, \mathbf{y}, x, h) = h^{-1} \sum_{i=1}^{n} \left\{ (-\lambda(x_i) + y_i \ln \lambda(x_i) - \ln(y_i)) K\left(\frac{x - x_i}{h}\right) \right\}, \tag{4.3}$$

substituindo $\lambda(x_i)$ por $\beta_0+\beta_1(x_i-x)$ dá

$$\mathcal{L}\left(\beta_0,\beta_1\left|\mathbf{x},\mathbf{y},x,h\right.\right)=h^{-1}\sum_{i=1}^{n}\left\{\left(-\left(\beta_0+\beta_1\left(x_i-x\right)\right)+y_i\ln\left(\beta_0+\beta_1\left(x_i-x\right)\right)-\ln\left(y_i\right)\right)K\left(\frac{x-x_i}{h}\right)\right\}$$

logo, as condições de primeira ordem são:

$$\frac{\partial\mathcal{L}_x^{(1)}\left(\cdot\right)}{\partial\beta_0}=h^{-1}\sum_{i=1}^{n}\left\{\left(-1+\frac{y_i}{\beta_0+\beta_1\left(x_i-x\right)}\right)K\left(\cdot\right)\right\}=0$$

$$\frac{\partial\mathcal{L}_x^{(1)}\left(\cdot\right)}{\partial\beta_1}=h^{-1}\sum_{i=1}^{n}\left\{\left(-\left(x-x_i\right)+\frac{y_i\left(x-x_i\right)}{\beta_0+\beta_1\left(x_i-x\right)}\right)K\left(\cdot\right)\right\}=0$$

na forma matricial

$$\frac{\partial\mathcal{L}_x^{(1)}\left(\cdot\right)}{\partial\beta}=h^{-1}\sum_{i=1}^{n}\left\{\left(-1+\frac{y_i}{\beta_0+\beta_1\left(x_i-x\right)}\right)K\left(\cdot\right)\right\}\begin{bmatrix}1\\x_i-x\end{bmatrix}=\begin{bmatrix}0\\0\end{bmatrix}. \qquad (4.4)$$

Para efeitos de programação, as equações foram separadas da seguinte forma:

$$\sum_{i=1}^{n}\frac{y_i}{\beta_0+\beta_1\left(x_i-x\right)}K\left(\frac{x-x_i}{h}\right)-\sum_{i=1}^{n}K\left(\frac{x-x_i}{h}\right)=0$$

$$\sum_{i=1}^{n}\frac{y_i\left(x_i-x\right)}{\beta_0+\beta_1\left(x_i-x\right)}K\left(\frac{x-x_i}{h}\right)-\sum_{i=1}^{n}\left(x_i-x\right)K\left(\frac{x-x_i}{h}\right)=0$$

4.2) Viés

Usando um ajustamento polinomial de Taylor de grau 3, tal como foi feito em Santos (2005), o logaritmo da verosimilhança local ponderada pela função núcleo é

$$\mathcal{L}_1^*\left(\beta_0,\beta_1\middle|\mathbf{x},\mathbf{y},x,h\right)=h^{-1}\sum_{i=1}^{n}\left[\left(-\lambda\left(x_i\right)\right)+\ln\lambda\left(x_i\right)-\ln\left(y_i!\right)K\left(\frac{x-x_i}{h}\right)\right], \quad (4.5)$$

com

$$\lambda\left(x_i\right)=\beta_0+\beta_1\left(x_i-x\right)+\beta_2\left(x_i-x\right)^2+\beta_3\left(x_i-x\right)^3, \quad i=1,\mathrm{K}\,,n, \quad (4.6)$$

As condições de primeira ordem são:

$$\frac{\partial\mathcal{L}_x^{(3)}(\cdot)}{\partial\beta_0}=h_*^{-1}\sum_{i=1}^{n}\left\{\left(-1+\frac{y_i}{\beta_0+\beta_1\left(x_i-x\right)+\beta_2\left(x_i-x\right)^2+\beta_3\left(x_i-x\right)^3}\right)K(\cdot)\right\}=0$$

$$\frac{\partial\mathcal{L}_x^{(3)}(\cdot)}{\partial\beta_1}=h_*^{-1}\sum_{i=1}^{n}\left\{\left(-1+\frac{y_i\left(x_i-x\right)}{\beta_0+\beta_1\left(x_i-x\right)+\beta_2\left(x_i-x\right)^2+\beta_3\left(x_i-x\right)^3}\right)K(\cdot)\right\}=0$$

$$\frac{\partial\mathcal{L}_x^{(3)}(\cdot)}{\partial\beta_2}=h_*^{-1}\sum_{i=1}^{n}\left\{\left(-1+\frac{y_i\left(x_i-x\right)^2}{\beta_0+\beta_1\left(x_i-x\right)+\beta_2\left(x_i-x\right)^2+\beta_3\left(x_i-x\right)^3}\right)K(\cdot)\right\}=0$$

$$\frac{\partial\mathcal{L}_x^{(3)}(\cdot)}{\partial\beta_3}=h_*^{-1}\sum_{i=1}^{n}\left\{\left(-1+\frac{y_i\left(x_i-x\right)^3}{\beta_0+\beta_1\left(x_i-x\right)+\beta_2\left(x_i-x\right)^2+\beta_3\left(x_i-x\right)^3}\right)K(\cdot)\right\}=0$$

na forma matricial

$$\frac{\partial \mathcal{L}_x^{(3)}(\cdot)}{\partial \boldsymbol{\beta}} = h_*^{-1}\sum_{i=1}^{n}\left\{\left(-1+\frac{y_i}{\beta_0+\beta_1(x_i-x)+\beta_2(x_i-x)^2+\beta_3(x_i-x)^3}\right)K(\cdot)\right\}\begin{bmatrix}1\\(x_i-x)\\(x_i-x)^2\\(x_i-x)^3\end{bmatrix}=\begin{bmatrix}0\\0\\0\\0\end{bmatrix}$$

$$\boldsymbol{\beta}=(\beta_0,\beta_1,\beta_2,\beta_3)' \qquad (4.7)$$

Do sistema obtêm-se as soluções para $\boldsymbol{\beta}$. Considera-se as soluções/estimativas para β_2 e β_3 do sistema para estimar o viés do alisador de verosimilhança local, dado por

$$\hat{r}=\hat{\beta}_2(x_i-x)^2+\hat{\beta}_3(x_i-x)^3, \quad i=1,\mathrm{K}\,,n\,. \qquad (4.8)$$

A verosimilhança local com inclusão da estimativa do resto

$$\mathcal{L}_1^*(\cdot)=h^{-1}\sum_{i=1}^{n}\left[\left(-\left(\beta_0+\beta_1(x_i-x)+r_i\right)+y_i\ln\left(\beta_0+\beta_1(x_i-x)+r_i\right)-\ln(y_i!)\right)K(\cdot)\right] \qquad (4.9)$$

cujo gradiente é

$$\mathcal{L}_1^{*'}(\beta_0,\beta_1|\mathbf{x},\mathbf{y},x,h)=\begin{bmatrix}h^{-1}\sum_{i=1}^{n}\left(-1+\frac{y_i}{\beta_0+\beta_1(x_i-x)+r_i}\right)K(\cdot)\\ h^{-1}\sum_{i=1}^{n}\left(-(x_i-x)+\frac{y_i(x_i-x)}{\beta_0+\beta_1(x_i-x)+r_i}\right)K(\cdot)\end{bmatrix}$$

$$=h^{-1}\sum_{i=1}^{n}\left[\left(\frac{y_i}{\beta_0+\beta_1(x_i-x)+r_i}-1\right)K(\cdot)\right]\begin{bmatrix}1\\x_i-x\end{bmatrix}\Leftrightarrow$$

$$\underbrace{1\ 2\ 3}_{x_i}$$

$$\mathcal{L}_1^{*'}\left(\beta_0,\beta_1\middle|\mathbf{x},\mathbf{y},x,h\right)=h^{-1}\sum_{i=1}^{n}\left[\left(\frac{y_i}{\beta_0+\beta_1\left(x_i-x\right)+r_i}-1\right)K\left(\frac{x-x_i}{h}\right)\mathbf{x_i}\right] \tag{4.10}$$

na forma matricial, com $\mathbf{x_i}=\begin{bmatrix}1\\ x_i-x\end{bmatrix}$.

A matriz hessiana é dada por:

$$\mathcal{L}_1^{*''}\left(\beta_0,\beta_1\middle|\mathbf{x},\mathbf{y},x,h\right)=\begin{bmatrix}h^{-1}\sum_{i=1}^{n}\frac{-y_i}{\beta_0+\beta_1\left(x_i-x\right)}K(\cdot) & h^{-1}\sum_{i=1}^{n}\frac{-y_i\left(x_i-x\right)}{\beta_0+\beta_1\left(x_i-x\right)}K(\cdot)\\ h^{-1}\sum_{i=1}^{n}\frac{-y_i\left(x_i-x\right)}{\beta_0+\beta_1\left(x_i-x\right)}K(\cdot) & h^{-1}\sum_{i=1}^{n}\frac{-y_i\left(x_i-x\right)^2}{\beta_0+\beta_1\left(x_i-x\right)}K(\cdot)\end{bmatrix}=$$

$$h^{-1}\sum_{i=1}^{n}\left[h^{-1}\sum_{i=1}^{n}\frac{-y_i}{\beta_0+\beta_1\left(x_i-x\right)}K(\cdot)\right]\begin{bmatrix}1 & \left(x_i-x\right)\\ \left(x_i-x\right) & \left(x_i-x\right)^2\end{bmatrix}\Leftrightarrow$$

$$\mathcal{L}_1^{*''}\left(\beta_0,\beta_1\middle|\mathbf{x},\mathbf{y},x,h\right)=-h^{-1}\sum_{i=1}^{n}\left[h^{-1}\sum_{i=1}^{n}\frac{y_i}{\beta_0+\beta_1\left(x_i-x\right)}K\left(\frac{x-x_i}{h}\right)\right]\mathbf{x_i}\mathbf{x_i}' \tag{4.11}$$

O viés, definido como $\mathbf{v}=E\left[\hat{\beta}\middle|\mathbf{x}\right]-\beta$, pode ser estimado, devidamente adaptado, através do estimador dado pela equação (5.19) de Santos (2005),

$$\hat{\mathbf{v}}\left(x;p\right)=\mathcal{L}_1^{*''}\left(\hat{\beta}_0,\hat{\beta}_1\middle|\mathbf{x},\mathbf{y},x,h\right)^{-1}\mathcal{L}_1^{*'}\left(\hat{\beta}_0,\hat{\beta}_1\middle|\mathbf{x},\mathbf{y},x,h\right) \tag{4.12}$$

$$=\left(-h^{-1}\sum_{i=1}^{n}\left[h^{-1}\sum_{i=1}^{n}\frac{y_i}{\beta_0+\beta_1\left(x_i-x\right)}K(\cdot)\right]\mathbf{x_i}\mathbf{x_i}'\right)^{-1}h^{-1}\sum_{i=1}^{n}\left[\left(\frac{y_i}{\beta_0+\beta_1\left(x_i-x\right)+r_i}-1\right)K\left(\frac{x-x_i}{h}\right)\mathbf{x_i}\right]$$

$$=\left(\sum_{i=1}^{n}\left[h^{-1}\sum_{i=1}^{n}\frac{y_i}{\beta_0+\beta_1\left(x_i-x\right)}K(\cdot)\right]\mathbf{x_i}\mathbf{x_i}'\right)^{-1}\left(\sum_{i=1}^{n}\left[\left(\frac{y_i}{\beta_0+\beta_1\left(x_i-x\right)+r_i}-1\right)K\left(\frac{x-x_i}{h}\right)\mathbf{x_i}\right]\right)$$

4.3 Variância

A contribuição individual de cada observação (X_i,Y_i) para a log-verosimilhança condicional local é:

$$l\{m(X_i),Y_i\} = -\lambda(X_i) + Y_i \ln\lambda(X_i) - \ln(Y_i)$$
$$= -(\beta_0 + \beta_1(X_i - x)) + Y_i \ln(\beta_0 + \beta_1(X_i - x)) - \ln(Y_i)$$

que, no ponto de interesse x, tem a expressão

$$l\{m(x_i),Y\} = -\beta_0 + Y \ln\beta_0 - \ln(Y!).$$

A sua derivada é derivada:

$$l'\{m(x),Y\} = \frac{\partial}{\partial\beta_0} l\{m(x),Y\} = -1 + Y\frac{1}{\beta_0}.$$

A variância vem

$$Var\left[l'\{m(x)\},Y\right] = Var\left[-1 + Y\frac{1}{\beta_0}\right] = \frac{1}{\beta_0^2} Var[Y],$$

do modelo de Poisson, temos $Var\left[Y \middle| X = x\right] = \lambda(x) = \beta_0$,

logo, $Var\left[l'\{m(x)\}, Y \middle| X = x \right] = \frac{1}{\beta_0^2} \times \beta_0 = \frac{1}{\beta_0}$.

De acordo com Fan *et al* (cf. Santos (2005) pp. 69 - a equação (5.22)), o estimador da variância é:

$$Var\left[\hat{\beta} \middle| \mathbf{x} \right] = Var\left[l'\{m(x), Y\} \middle| X = x \right] \mathcal{L}_1''\left(\hat{\beta}_0, \hat{\beta}_1 \middle| x, h \right)^{-1} \overline{S}_n \mathcal{L}_1''\left(\hat{\beta}_0, \hat{\beta}_1 \middle| x, h \right)^{-1} \qquad (4.13)$$

com $\overline{S}_n = h^{-2} \sum_{i=1}^{n} \mathbf{x_i} \mathbf{x_i}' K^2 \left\{ \frac{x - x_i}{h} \right\}$, que, no modelo aqui apresentado, resulta em:

$$\begin{aligned} Var\left[\hat{\beta} \middle| \mathbf{x} \right] = & \frac{1}{\beta_0} \left(-h^{-1} \sum_{i=1}^{n} \left[h^{-1} \sum_{i=1}^{n} \frac{y_i}{\beta_0 + \beta_1 (x_i - x)} K(\cdot) \right] \mathbf{x_i} \mathbf{x_i}' \right)^{-1} \\ & \times h^{-2} \sum_{i=1}^{n} \mathbf{x_i} \mathbf{x_i}' K^2 \left\{ \frac{x - x_i}{h} \right\} \times \left(-h^{-1} \sum_{i=1}^{n} \left[h^{-1} \sum_{i=1}^{n} \frac{y_i}{\beta_0 + \beta_1 (x_i - x)} K(\cdot) \right] \mathbf{x_i} \mathbf{x_i}' \right)^{-1} \end{aligned} \qquad (4.14)$$

4.4 Distribuição Assintótica

Segundo Fan *et al* (1996), o estimador de máxima variância local, $\hat{\beta}_j, j = 1, \mathrm{K}, p$, tem como distribuição assintótica relativamente a um ponto de interesse x do regressor é normal:

$$\left(Var\left[\hat{\beta}_j \middle| \mathbf{x} \right] \right)^{-\frac{1}{2}} \left[\hat{\beta}_j - \beta_j - v_x\left(\hat{\beta}_j \right) \right] \overset{a}{\sim} N(0,1), \qquad (4.15)$$

em que $v_x\left(\hat{\beta}_j \right)$ é o *j*-ésimo elemento do vector do viés e $Var\left[\hat{\beta}_j \middle| \mathbf{x} \right]$ é o *j*-ésimo elemento da diagonal da matriz de variâncias.

Podemos construir intervalos de confiança a $(1\text{-}\alpha)\%$, utilizando o viés e variância estimados, que dão origem a bandas de confiança, quando considerados todos os pontos da amostra em geral:

$$\hat{\beta}_j - \hat{v}\left(\hat{\beta}_j\right) \pm z_{1-\frac{\alpha}{2}} \left(Var\left[\hat{\beta}_j \middle| \mathbf{x}\right]\right)^{-\frac{1}{2}}. \tag{4.16}$$

5.Simulação e Estudo de um Caso

O estimador apresentado no capítulo anterior irá agora ser agora aplicado a dados de simulação, bem como a dados reais, de forma semelhante à que foi efectuada em Santos (2005). O procedimento será idêntico ao descrito no capítulo 6 de Santos (2005). Foi utilizado o *software* gratuito *R*. Os dados de simulação, em quatro amostras de dimensões diferentes, serão gerados segundo uma distribuição uniforme, no intervalo de [-2,2], com as respostas geradas segundo o modelo de Poisson e segundo o modelo de Poisson inflacionado em zero (ZIP). Os dados reais referem-se ao número de infecções do tracto urinário em indivíduos do sexo masculino, infectados com o HIV, em função do número de células CD4+ presentes em cada indivíduo. Os dados foram gentilmente cedidos por van den Broek.

5.1 Simulação

Os dados de simulação são gerados aleatoriamente, segundo uma distribuição uniforme, com valores entre [-2,2]. São geradas quatro amostras com dimensões n=50, 100, 500 e 1000. A selecção do valor óptimo de largura de banda, h, fez-se por partição da amostra em duas subamostras: de treino ou aprendizagem, e de teste ou validação. A amostra de teste é obtida recolhendo aleatoriamente m=20% da amostra aleatória. A amostra de treino é constituída pelos restantes 80%, que são usados para a estimação dos parâmetros do modelo.

O alisador de máxima verosimilhança local é baseado no núcleo Gaussiano $K(z)=\frac{1}{\sqrt{2\pi}}e^{-\frac{z^2}{2}}{}_{\{z\in[-\infty;+\infty]\}}$. A largura de banda óptima foi seleccionada na amostra de teste, segundo o critério de minimização $\hat{h}_{opt}=\arg\min_h \sum_{i=1}^{m}(y_i-\hat{y}_i)^2$, com

$\hat{y}_i = \hat{\lambda}_i = \hat{\beta}_0$. Os valores da variável resposta são gerados aleatoriamente segundo o modelo de regressão de Poisson univariado, com $\lambda_i = \beta_0 + \beta_1 x_i = -2 + 2x_i$. A figura 1) ilustra, para as várias dimensões amostras, as respostas geradas:

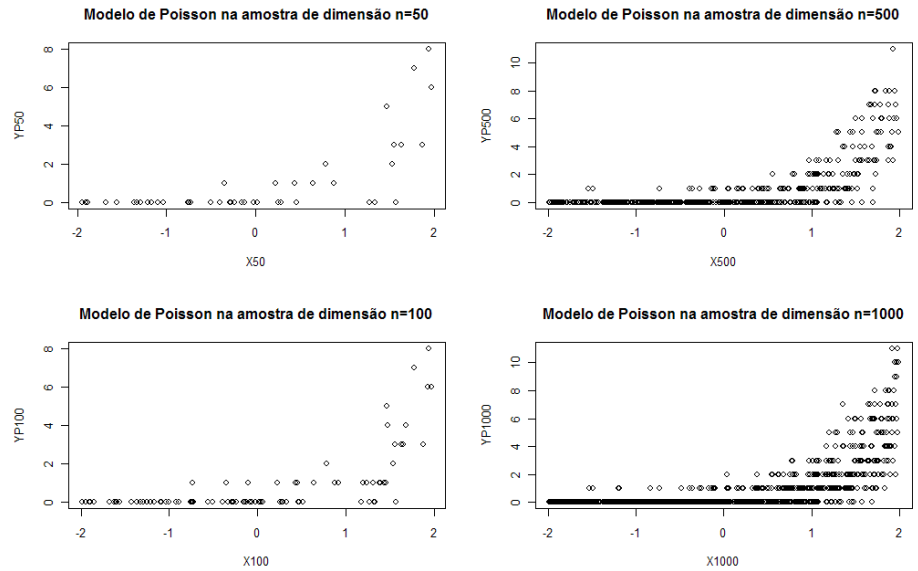

Figura 1) Respostas geradas segundo o modelo de Poisson

e segundo o modelo de Poisson inflacionado em zero (ZIP), com ψ=0,2, ou seja, com especificação

$$P\left(Y_i = y_i \left| \mathbf{x_i} \right.\right) = \begin{cases} 0,2 + 0,8e^{(-\lambda_i)}, & y_i = 0 \\ 0,8\dfrac{e^{(-\lambda_i)}\lambda_i^{y_i}}{y_i!}, & y_i > 0 \end{cases}.$$

como se pode ver na figura 2):

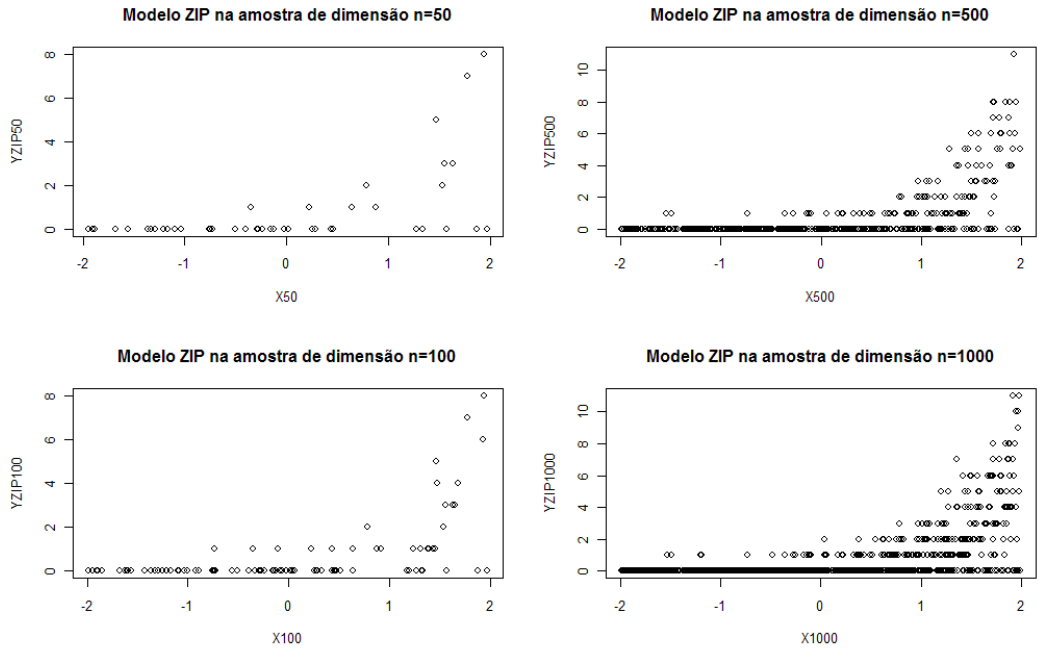

Figura 2) Respostas geradas segundo o modelo ZIP

5.2 Estudo de Caso

Os dados de regressão no estudo de caso consistem no número de infecções no tracto urinário, em noventa e oito indivíduos do sexo masculino infectados com HIV, em função do número de células CD4+. A variável resposta é o número de infecções urinárias contraídas por cada indivíduo, durante o período de observação do estudo, e o regressor é a contagem de células CD4+ desse indivíduo. A frequência de contagens zero é elevada, como se pode ver na seguinte tabela de frequências relativas:

Episódios	0	1	2	3
Frequências (%)	82,7	9,2	7,1	1

Tabela 2) Frequências relativas (cf. Tabela 6.2 Santos (2005))

Foi utilizado o método de Newton-Raphson para a resolução do sistema de equações de primeira ordem (equação 4.4). O método de Newton para

resolução de sistemas não lineares pode ser visto, por exemplo, em Scheid (1988). Uma apresentação sucinta é aqui reproduzida:

Um sistema de equações não lineares, com a seguinte forma genérica,

$$\mathbf{F}(\mathbf{x}) = \mathbf{0},$$

onde **F**, **x** e **0** são vectores n-dimensionais. O método iterativo pode ser expresso na forma matricial com a série de Taylor

$$\mathbf{F}\left(\mathbf{x}^{(n-1)} + \mathbf{h}\right) = \mathbf{F}\left(\mathbf{x}^{(n-1)}\right) + \mathbf{J}\left(\mathbf{x}^{(n-1)}\right)\mathbf{h} + \mathrm{K}$$

ignorando os termos de ordem superior e em que **J** é a matriz jacobiana de **F**, ou seja,

$$J_{ij} = \frac{\partial f_i}{\partial x_j} \Leftrightarrow \mathbf{J} = \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \mathrm{L} & \frac{\partial f_1}{\partial x_n} \\ \mathrm{M} & \mathrm{O} & \mathrm{M} \\ \frac{\partial f_n}{\partial x_1} & \mathrm{L} & \frac{\partial f_n}{\partial x_n} \end{bmatrix}.$$

Se fizermos o primeiro termo igual a zero e resolvendo em ordem a **h**, obtemos

$$\mathbf{J}\left(\mathbf{x}^{(n-1)}\right)\mathbf{h} = -\mathbf{F}\left(\mathbf{x}^{(n-1)}\right)$$
$$\Leftrightarrow$$
$$\mathbf{h} = -\left[\mathbf{J}\left(\mathbf{x}^{(n-1)}\right)\right]^{-1}\mathbf{F}\left(\mathbf{x}^{(n-1)}\right)$$

ou numa formulação equivalente:

$$\left(\mathbf{x}^{(n+1)} - \mathbf{x}^{(n)}\right) = -\left[\mathbf{J}\left(\mathbf{x}^{(n)}\right)\right]^{-1} \mathbf{F}\left(\mathbf{x}^{(n)}\right)$$

$$\Leftrightarrow$$

$$\mathbf{x}^{(n+1)} = -\left[\mathbf{J}\left(\mathbf{x}^{(n)}\right)\right]^{-1} \mathbf{F}\left(\mathbf{x}^{(n)}\right) + \mathbf{x}^{(n)}$$

$$\Leftrightarrow$$

$$\mathbf{x}^{(n+1)} = \mathbf{x}^{(n)} + \mathbf{h}$$

que é um sistema linear. Em geral, não existe uma maneira simples de garantir a convergência para a solução pretendida ou simplesmente de localizar a solução. O sucesso deste método depende muito do ponto de inicialização escolhido do algoritmo. A convergência depende assim, entre outras condições, da estimativa de inicialização do algoritmo. Estas condições são as mesmas que para o caso univariado:

$$x_n = x_{n-1} - \frac{f\left(x_{n-1}\right)}{f'\left(x_{n-1}\right)}$$

e devem ser respeitadas por **<u>todas</u>** as equações do sistema:

1. x_0 deve estar "suficientemente próxima" de uma raiz da equação
2. A função f deve ser diferenciável e $f'(x) \neq 0, \forall x \in [a;b]$
3. $f'(x)$ não deve ser muito próxima de zero (em valor absoluto)
4. $f''(x)$ não deve tomar, a cada iteração, valores "excessivamente" grandes

Em rigor, as condições suficientes para a convergência do algoritmo são:

Teorema: Sejam $f(x)$ diferenciável duas vezes (C^2) e **[*a;b*]** que contém uma única raiz x_0 para $f(x)=0$:

1. $f(a).f(b) \leq 0$
2. $f'(x) \neq 0, \forall x \in [a;b]$
3. $f''(x)$ é de sinal constante, isto é, $f''(x) \geq 0$ ou $f''(x) \leq 0$, para $x \in]a;b[$

4.a) $\left|\frac{f(a)}{f'(a)}\right| < |a-b|$ **e** $\left|\frac{f(b)}{f'(b)}\right| < |a-b|$ (pontos extremos a e b)

ou

4.b) $f(x_0).f''(x) \geq 0, \forall x \in [a;b]$

então a equação $f(x)=0$ tem uma única solução em **[*a;b*]** e o algoritmo tende para essa solução:

- qualquer que seja o x_0 pertencente a **[*a;b*]** se se verificar 4.a)
- para os x_0 em **[*a;b*]** que verificarem 4.b)

6. Discussões e Conclusões. Trabalho Futuro

Em face dos resultados obtidos, serão efectuados comentários de análise de desempenho do alisador de máxima verosimilhança local (MVL) de Poisson alternativo, aqui proposto, e é efectuada a comparação com o modelo apresentado em Santos (2005).

Os valores iniciais, ou de arranque, foram obtidos através da regressão pelo modelo de Poisson, efectuando a estimação nos elementos da amostra de teste em cada amostra. Dada a importância dos pontos de inicialização foram testados vários pontos, de acordo com vários processos com origem nos resultados da regressão de Poisson. Para cada ponto, utilizou-se a regressão global, depois dentro de sucessivas janelas de largura variável em torno do ponto de interesse, para melhorar a possibilidade de sucesso do algoritmo.

No entanto, infelizmente, não houve qualquer convergência em nenhum ponto de interesse, bem como em nenhum dos pontos de inicialização.

6.1 Discussões e Conclusões

Pode-se discutir, em face dos resultados, os seguintes aspectos:

1) Numa primeira análise, o modelo proposto em Santos (2005), ao estabelecer para o valor médio $\lambda(\mathrm{x})$ uma especificação exponencial de uma função desconhecida a estimar, de que resulta a seguinte função de log-verosimilhança,

$$L_1\left(\beta_0,\beta_1\middle|\mathbf{x},\mathbf{y},x,h\right)=h^{-1}\sum_{i=1}^{n}\left\{\left(-e^{(\beta_0+\beta_1(x_i-x))}+y_i\ln\left(e^{(\beta_0+\beta_1(x_i-x))}\right)-\ln\left(y_i!\right)\right)K\left\{\frac{\left(x_i-x\right)}{h}\right\}\right\},$$

que gera o seguinte sistema de condições de primeira ordem (eq. 5.14, pp 66),

$$\sum_{i=1}^{n}\left[\left(y_i - e^{(\beta_0+\beta_1(x_i-x))}\right)K\left\{\frac{(x_i-x)}{h}\right\}\right]\begin{bmatrix}1\\(x_i-x)\end{bmatrix}=0,$$

o qual acarreta um peso computacional elevado. Comparando com a formulação apresentada neste trabalho, em que o valor médio $\lambda(x)$ é especificado como um polinómio, resultando na seguinte função de log-verosimilhança:

$$\mathcal{L}_1\left(\beta_0,\beta_1\middle|\mathbf{x},\mathbf{y},x,h\right)=h^{-1}\sum_{i=1}^{n}\left[\left(-\left(\beta_0+\beta_1\left(x_i-x\right)\right)+y_i\ln\left(\beta_0+\beta_1\left(x_i-x\right)\right)-\ln\left(y_i!\right)\right)K\left\{\frac{(x_i-x)}{h}\right\}\right]$$

e nas seguintes condições de primeira ordem (eq. 4.4):

$$\sum_{i=1}^{n}\left[\left(\frac{y_i}{\beta_0+\beta_1(x_i-x)}-1\right)K\left\{\frac{(x_i-x)}{h}\right\}\right]\begin{bmatrix}1\\(x_i-x)\end{bmatrix}=0,$$

que, como se pode ver, não apresenta nenhum termo com a função exponencial, pelo que é de esperar um ganho computacional significativo, no modelo aqui proposto.

Fazendo a mesma comparação para os estimadores do viés, em Santos (2005) o estimador para o viés é:

$$\hat{\mathbf{v}}(x;p)=-\left(\sum_{i=1}^{n}e^{\hat{\beta}_0-\hat{\beta}_1(x_i-x)+\hat{r}_i}.K\left\{\frac{x_i-x}{h}\right\}\mathbf{x_i}\mathbf{x_i'}\right)^{-1}\times\left(\sum_{i=1}^{n}\left(y_i-e^{\hat{\beta}_0-\hat{\beta}_1(x_i-x)+\hat{r}_i}\right).K\left\{\frac{x_i-x}{h}\right\}\mathbf{x_i}\right)$$

enquanto que neste trabalho é:

$$\hat{\mathbf{v}}(x;p)=\left(\sum_{i=1}^{n}\left(\frac{y_i}{\left(\beta_0+\beta_1(x_i-x)+r_i\right)^2}K\left\{\frac{x_i-x}{h}\right\}\right)\mathbf{x}_i\mathbf{x}_i'\right)^{-1}\left(h^{-1}\sum_{i=1}^{n}\left[\left(\frac{y_i}{\beta_0+\beta_1(x_i-x)+r_i}-1\right)K\left\{\frac{x_i-x}{h}\right\}\mathbf{x}_i\right]\right)$$

que, mais uma vez, é uma expressão matematicamente menos complexa, sem termos exponenciais.

Comparando os estimadores da variância:

$$\widehat{Var}\left[\hat{\beta}|\mathbf{x}\right]=e^{\hat{\beta}_0}.\left(\sum_{i=1}^{n}\left[e^{\hat{\beta}_0-\hat{\beta}_1(x_i-x)}\right]K\left\{\frac{x-x_i}{h}\right\}\right)^{-1}\times\left(\sum_{i=1}^{n}\mathbf{x_i}\mathbf{x_i'}K^2\left\{\frac{x-x_i}{h}\right\}\right)\times\left(\sum_{i=1}^{n}\left[e^{\hat{\beta}_0-\hat{\beta}_1(x_i-x)}\right]K\left\{\frac{x-x_i}{h}\right\}\right)^{-1}$$

em Santos (2005), e neste trabalho:

$$\widehat{Var}\left[\hat{\beta}|\mathbf{x}\right]=\frac{1}{\beta_0}\left(\sum_{i=1}^{n}\left(\frac{y_i}{\left(\hat{\beta}_0+\hat{\beta}_1(x_i-x)\right)^2}K\left\{\frac{x-x_i}{h}\right\}\right)\mathbf{x}_i\mathbf{x}_i'\right)^{-1}\times\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i'K^2\left\{\frac{x-x_i}{h}\right\}\times\left(\sum_{i=1}^{n}\left(\frac{y_i}{\left(\hat{\beta}_0+\hat{\beta}_1(x_i-x)\right)^2}K\left\{\frac{x-x_i}{h}\right\}\right)\mathbf{x}_i\mathbf{x}_i'\right)^{-1}$$

novamente, a exponencial é substituída por um quociente.

2) Outra questão a referir *a priori* é a da função núcleo utilizada no alisador: O critério mais frequente para avaliar a qualidade do ajustamento de um alisador é o erro quadrático médio integrado (MISE), definido pela seguinte expressão:

$MISE\left\{\hat{f}(.;h)\right\}=E\left[ISE\left\{\hat{f}(.;h)\right\}\right]=E\int\left\{\hat{f}(.;h)-f(x)\right\}^2dx$ (Wand e Jones (1995) pp. 15),

com aproximação assintótica para grandes amostras pelo erro quadrático médio integrado assintótico (AMISE). No modelo apresentado em Santos (2005), o alisador foi ajustado com a função de núcleo de Epanechnikov, $K(z)=\frac{3}{4}\left(1-z^2\right)\mathbf{1}_{z\in[-1,1]}$. O núcleo de Epanechnikov minimiza o erro quadrático médio integrado assintótico (AMISE), sendo portanto a solução óptima. Porém, esta escolha traz dificuldades computacionais bastante acrescidas, com a agravante de que, com este núcleo, termos de ter um número mínimo de observações em cada janela (largura de banda) para poder estimar o modelo localmente (graus de liberdade: nº de parâmetros+1), o que em muitos casos não acontece. No núcleo Gaussiano, este problema não existe, já que o seu suporte é a recta real, ¡ . Em Silverman (1986) e Simonoff (1996), por exemplo, pode ver-se a eficiência relativa entre funções núcleo relativamente à de Epanechnikov:

Núcleo	Eficiência Relativa
Epanechnikov	1,000
Biweight	0,994
Triangular	0,986
Triweight	0,943
Normal	0,951
Uniforme	0,930

Tabela 3) Eficiência Relativa de Funções Núcleo

De facto, de acordo com a tabela, a escolha de função núcleo não assume um papel tão importante, quanto a este critério, como a escolha da largura de banda h, essa sim crítica, conforme foi referido na secção 3.1. A perda de eficiência, que se traduz em 4,9 por cento, é portanto, despicienda.

3) Em vez de se estimar o modelo com recurso à regressão de núcleo, pode-se recorrer ao alisamento por *splines*, com os ganhos respectivos, a nível da

exigência computacional. Em Messer (1991), é efectuada uma comparação entre as estimativas conseguidas através de ambas as metodologias são similares, que poderão diferir apenas, de certa forma, no comportamento de fronteira de ordem superior. Mais recentemente, Aydin (2007) consegue inclusive resultados com estimadores de regressão por alisadores de *spline*, do que os obtidos através de regressão de núcleo, em dados reais do PIB da Turquia, e dados referentes a transplantes de coração efectuados em Stanford. Os critérios de comparação naquele trabalho foram o erro quadrático médio (MSE) ou a raiz deste (RMSE), o erro absoluto médio (MAE) e o erro absoluto médio percentual (MAPE), definidos no trabalho da seguinte forma:

$$MSE = \frac{1}{T}\sum_{t=1}^{T}\left(y_t - \hat{y}_t\right)^2 \text{ ou } RMSE = \sqrt{MSE}$$

$$MAE = \frac{1}{T}\sum_{t=1}^{T}\left|y_t - \hat{y}_t\right|$$

$$MAPE = \frac{1}{T}\sum_{t=1}^{T}\frac{\left|y_t - \hat{y}_t\right|}{\left|y_t\right|}\left(\%100\right)$$

4) O método de Newon-Raphson é o método iterativo menos exigente computacionalmente para resolução de sistemas de equações não lineares, sendo, no entanto, complicado no que respeita à distância do ponto de inicialização do ponto da verdadeira solução, sendo por isso uma sua grande desvantagem. O método de Newton modificado (que consiste em manter constante a jacobiana a cada iteração), tem uma convergência mais lenta (a convergência deixa de ser quadrática), exigindo mais iterações, embora a exigência operacional de cada iteração seja inferior. As alternativas, como o algoritmo do maior declive descendente, ou do gradiente conjugado, são algoritmos mais pesados.

O algoritmo de Newton-Raphson não convergiu para nenhum vector de soluções iniciais, nem para nenhum ponto de interesse das amostras. Foi tentado inicialmente nos pontos das amostras de teste. Algum problema computacional surgiu, existindo diversas possibilidades, nomeadamente, a não

reunião das condições suficientes de convergência por ambas as equações em simultâneo (embora sejam necessárias), muito possivelmente devido à difícil proximidade ao vector da solução para cada ponto de interesse. As primeiras derivadas da função de log-verosimilhança em ordem aos parâmetros β_0 e β_1 são, respectivamente:

$$\frac{\partial \mathcal{L}}{\partial \beta_0} = h^{-1}\sum_{i=1}^{n}\left(-1+\frac{y_i}{\beta_0+\beta_1(x_i-x)}\right)K\left\{\frac{(x-x_i)}{h}\right\}$$
$$\frac{\partial \mathcal{L}}{\partial \beta_1} = h^{-1}\sum_{i=1}^{n}\left(-(x_i-x)+\frac{y_i(x_i-x)}{\beta_0+\beta_1(x_i-x)}\right)K\left\{\frac{(x-x_i)}{h}\right\}$$

As segundas derivadas da função de logverosimilhança em ordem aos parâmetros β_0 e β_1 são, respectivamente:

$$\frac{\partial^2 \mathcal{L}}{\partial \beta_0^2} = h^{-1}\sum_{i=1}^{n}\frac{-y_i}{\left(\beta_0+\beta_1(x_i-x)\right)^2}K\left\{\frac{(x-x_i)}{h}\right\}$$
$$\frac{\partial^2 \mathcal{L}}{\partial \beta_1^2} = h^{-1}\sum_{i=1}^{n}\frac{-y_i(x_i-x)^2}{\left(\beta_0+\beta_1(x_i-x)\right)^2}K\left\{\frac{(x-x_i)}{h}\right\}$$

Se atentarmos às primeira e segunda derivadas em ordem a β_1, estas anulam-se para cada ponto da amostra coincidente com o ponto de interesse, situação em que $x=x_i$. Acresce que na segunda derivada, esta poderá ter até dois zeros, quer no numerador quer no denominador, visto que são elevados ao quadrado, para os pontos não coincidentes com x_i (recorde-se que o domínio é ¡ . Há ainda o problema dos eventuais zeros do denominador, em que a derivada não está definida. Tal não se verifica na especificação de Santos (2005)). Estes factos dificultam a verificação **conjunta** das condições do teorema acima exposto, nomeadamente as dos pontos 2. e 3. e logo, a garantia de convergência do algoritmo de Newton-Raphson. Note-se que estas condições são suficientes mas não necessárias, pelo que o algoritmo pode, mesmo assim convergir para um vector de soluções. Na formulação de Santos (2005), a fracção é substituída por uma exponencial, nas derivadas. Ora a exponencial é

uma função que não se anula. A exponencial, apesar de ser mais complexa de tratar e programar. Todavia, ao não se anular, facilitará, porventura, a verificação das condições, e logo, a convergência do algoritmo para um vector de soluções.

Em síntese: quanto aos objectivos deste trabalho, comparar o desempenho do modelo aqui proposto com o modelo apresentado no ponto 5.2 de Santos (2005), uma vez que não foi obtida convergência para as soluções, não é possível concluir quanto à qualidade de ajustamento da metodologia aqui proposta. É de supor que, com recurso a uma ferramenta mais potente (possivelmente o Fortran, como no caso de Santos (2005)), se forem obtidas soluções, poder-se-á então avaliar o eventual ganho a nível da exigência computacional e a possível perda de qualidade de ajustamento de ambas as formulações. O intuito da utilização do *software R* é de que este é de acesso livre. É de todo o interesse conseguir, de forma não onerosa, acesso a estas metodologias, uma vez que a sua utilidade prática é transversal a diversas disciplinas e áreas científicas e de actividade. Fala-se aqui de Fortran, querendo dizer outos métodos de optimização mais poderosos e potentes sem a grande desvantagem do algoritmo de Newton-Raphson.

6.2 Trabalho Futuro

O sistema de equações referente às condições de primeira ordem de máxima verosimilhança local, não convergiu, na utilização do método de Newton-Raphson, com os pontos iniciais gerados por regressão de Poisson global ou estimada em janelas de diferentes larguras, no software *R*. Fica para trabalho futuro a utilização métodos algoritmos alternativos, com prejuízo de requisitos computacionais (conforme referido acima), possivelmente devolvam as soluções das equações de convergência de primeira ordem. Possivelmente se consiga a

convergência através de ferramentas mais potentes, como o Matlab ou o Fortran.

Neste trabalho, o modelo foi apenas ajustado a dados univariados, isto é, apenas com um regressor. Fica por testar o desempenho em dados com mais de um regressor, isto é, a modelos multivariados.

Sugere-se ajustar outros modelos paramétricos localmente, para modelar outros dados, nomeadamente a regressão logística, ou a binomial negativa, ou com base nestes.

Fica também a sugestão de ajustar, pelo exposto acima, um alisador de *spline*, no lugar da função núcleo, o que poderá trazer ganhos a nível de aderência do ajustamento e a nível computacional.

Pode ainda ajustar-se um modelo não paramétrico, sem qualquer componente paramétrica, isto é, sem especificar localmente o modelo de Poisson, recorrendo à regressão polinomial local, sem verosimilhança local, o que poderá trazer facilidades de ajustamento e de natureza computacional.

7. Anexos

Código R

```
# 1) Gerar uma amostra aleatória, de uma população com distribuição
uniforme,
#de dimensão n=50,100,500,1000

#fixar a semente
set.seed(23451)
#gerar a amostra com a semente da linha anterior
X50 <- runif(n=50, min=-2, max=2)
X50

#experiência, verificando que se gera sempre os mesmos dados com a
semente inicial
set.seed(23451)
Xi50 <- runif(n=50, min=-2, max=2)
Xi50
X50-Xi50

set.seed(23451)
X100 <- runif(n=100, min=-2, max=2)

set.seed(23451)
X500 <- runif(n=500, min=-2, max=2)

set.seed(23451)
X1000 <- runif(n=1000, min=-2, max=2)
```

```
# 2) Gerar os parâmetros lambda para cada uma das quatro amostras
n=50,100,500,1000
lambda50 <- exp(-2+2*X50)
lambda50

lambda100 <- exp (-2+2*X100)
lambda500 <- exp (-2+2*X500)
lambda1000 <- exp (-2+2*X1000)



# 3) Gerar uma amostra aleatória, de uma população com distribuição de
Poisson,
#de dimensão n=50,100,500,1000

set.seed(19671)
YP50 <- rpois(50,lambda50)
YP50

#ver se gera os valores cnforme pensamos
YP50E<-c()
for (i in 1:50){
set.seed(19671)
YP50E[i]<-rpois(1,exp(-2+2*X50[i]))
}
YP50-YP50E
#não dá a mesma coisa, por causa da semente



#experiência com lambda fixo - distribuição de probabilidade e não regressão
XPFixo<-rpois(25,2)
```

XPFixo

```
#experiência, verificando que se gera sempre os mesmos dados com a
semente inicial
set.seed(19671)
YP50i <- rpois(50,lambda50)
YP50i
YP50-YP50i

set.seed(19671)
YP100 <- rpois(100,lambda100)

set.seed(19671)
YP500 <- rpois(500,lambda500)

set.seed(19671)
YP1000 <- rpois(1000,lambda1000)

# 4) MODELO ZIP
# .1) gerar amostra aleatória uniforme
# .2) gerar respostas segundo o modelo ZIP


# 4.1) gerar amostras aleatórias, de uma população com distribuição ZIP, com
psi=0.2,
#de dimensões n=50,100,500,1000
#gerar uma variável auxiliar com distribuição uniforme em [0,1]

set.seed(45217)
amostraZIP50 <- runif(n=50)
amostraZIP50
```

```
set.seed(45217)
amostraZIP100 <- runif(n=100)

set.seed(45217)
amostraZIP500 <- runif(n=500)

set.seed(45217)
amostraZIP1000 <- runif(n=1000)

# 4.2) gerar respostas segundo o modelo ZIP
# n=50
i <- 1
YZIP50 <- YP50
for (i in 1:50){
if (amostraZIP50[i] <= 0.2) YZIP50[i] <- 0
else YZIP50[i] <- YP50[i]
i <- i+1
}
YZIP50

#n=100
i <- 1
YZIP100 <- YP100
for (i in 1:100){
if (amostraZIP100[i] <= 0.2) YZIP100[i] <- 0
else YZIP100[i] <- YP100[i]
i <- i+1
}
YZIP100

#n=500
```

```
i <- 1
YZIP500 <- YP500
for (i in 1:500){
if (amostraZIP500[i] <= 0.2) YZIP500[i] <- 0
else YZIP500[i] <- YP500[i]
i <- i+1
}
YZIP500

#n=1000
i <- 1
YZIP1000 <- YP1000
for (i in 1:1000){
if (amostraZIP1000[i] <= 0.2) YZIP1000[i] <- 0
else YZIP1000[i] <- YP1000[i]
i <- i+1
}
YZIP1000


#ESCOLHA DAS AMOSTRAS DE TESTE E DE TREINO

set.seed(46247)
SAM50 <- runif(n=50)
set.seed(46247)
SAM100 <- runif(n=100)
set.seed(46247)
SAM500 <- runif(n=500)
set.seed(46247)
SAM1000 <- runif(n=1000)
```

```
TESTRE50<-c()
for (i in 1:50) {
if (SAM50[i]<=sort(SAM50, decreasing = FALSE)[10]) TESTRE50[i]<-1
else TESTRE50[i]<-2
}
#confirmar
table(TESTRE50)

TESTRE100<-c()
for (i in 1:100) {
if (SAM100[i]<=sort(SAM100, decreasing = FALSE)[20]) TESTRE100[i]<-1
else TESTRE100[i]<-2
}
#confirmar
table(TESTRE100)

TESTRE500<-c()
for (i in 1:500) {
if (SAM500[i]<=sort(SAM500, decreasing = FALSE)[100]) TESTRE500[i]<-1
else TESTRE500[i]<-2
}
#confirmar
table(TESTRE500)

TESTRE1000<-c()
for (i in 1:1000) {
if (SAM1000[i]<=sort(SAM1000, decreasing = FALSE)[200]) TESTRE1000[i]<-1
else TESTRE1000[i]<-2
}
#confirmar
table(TESTRE1000)
```

```
#AMOSTRAS COM TODAS AS VARIÁVEIS E A VARIÁVEL
#INDICATRIZ DA AMOSTRA DE TESTE E DA AMOSTRA DE TREINO

# CRIAR AS MATRIZES

Matriz50 <- cbind(X50, YP50, YZIP50, TESTRE50)
Matriz50

Matriz100 <- cbind(X100, YP100, YZIP100, TESTRE100)
Matriz100

Matriz500 <- cbind(X500, YP500, YZIP500, TESTRE500)
Matriz500

Matriz1000 <- cbind(X1000, YP1000, YZIP1000, TESTRE1000)
YP1000-YZIP1000

#OBTER A MATRIZ DE TESTE E DE TREINO

Matriz50TE <- subset(Matriz50,subset =  TESTRE50==1)
Matriz50TE
Matriz50TR <- subset(Matriz50,subset =  TESTRE50==2)
Matriz50TR

Matriz100TE <- subset(Matriz100,subset =  TESTRE100==1)
Matriz100TE
Matriz100TR <- subset(Matriz100,subset =  TESTRE100==2)
Matriz100TR

Matriz500TE <- subset(Matriz500,subset =  TESTRE500==1)
Matriz500TE
Matriz500TR <- subset(Matriz500,subset =  TESTRE500==2)
```

```
Matriz500TR

Matriz1000TE <- subset(Matriz1000,subset =  TESTRE1000==1)
Matriz1000TE
Matriz1000TR <- subset(Matriz1000,subset =  TESTRE1000==2)
Matriz1000TR

# LARGURA DE BANDA h (com as amostras de treino)

h<-seq(.8,20,.1)
h



#n=500
X1000, YP1000, YZIP1000, TESTRE1000

xte<-c()
yte<-c()
xte<-Matriz500TE[,1]
yte<-Matriz500TE[,2]



#obter os valores de arranque para a solução das equações de MV

regpois<-glm(yte ~ xte, family=poisson())
regpois
str(regpois)
summary(regpois)
regpois$coefficients
regpois$coefficients[1]
regpois$coefficients[2]
```

```
library(rootSolve)
#b0b1<-matrix(data = NA,nrow=NROW(xte),ncol=2)
#for (j in 1:NROW(xte)){
j<-15
F1<-0
F2<-0
h<-2

eqmaxver<-function(x){
for (i in 1:NROW(xte)){
 F1<-((yte[i]/(x[1]+x[2]*(xte[i]-xte[j]))-1)*(2*pi)^(-1/2)*exp(-((xte[j]-
xte[i])/h)^2/2))+F1
 F2<-((yte[i]/(x[1]+x[2]*(xte[i]-xte[j]))-1)*(xte[i]-xte[j])*(2*pi)^(-1/2)*exp(-
((xte[j]-xte[i])/h)^2/2))+F2
}
F1=F1
F2=F2
c(F1=F1,F2=F2)
}

jfun<-function(x){
  J<-matrix(nrow=2,ncol=2)
  J[1,1]<-sum(-(2*pi)^(-1/2)*exp(-((xte[j]-xte)/h)^2/2)*yte/(x[1]+x[2]*(xte-
xte[j]))^2)
  J[1,2]<-sum(-(2*pi)^(-1/2)*exp(-((xte[j]-xte)/h)^2/2)*yte*(xte-
xte[j])/(x[1]+x[2]*(xte-xte[j]))^2)
  J[2,1]<-J[1,2]
  J[2,2]<-sum(-(2*pi)^(-1/2)*exp(-((xte[j]-xte)/h)^2/2)*yte*(xte-
xte[j])^2/(x[1]+x[2]*(xte-xte[j]))^2)
  J<-J/h
  return(J)
```

```
}

#(sb<-
multiroot(f=eqmaxver,start=c(regpois$coefficients[1],regpois$coefficients[2]),a
tol=c(1e-20,1e-20)))
(sb<-
multiroot(f=eqmaxver,start=c(regpois$coefficients[1],regpois$coefficients[2]),a
tol=1e-10,
            jactype="fullint",jacfunc=jfun))
}

eqmaxver(sb$f.root)
jfun(sb$f.root)

Newton<-function(f, start, rtol, ftol, nmax, jac) {
  x.antes <- start
  stop<-F
  n.iter <- 0
  while (!stop & n.iter<=nmax) {
     n.iter <- n.iter+1
     jacob <- jac(x.antes)
   print(jacob)
     b <- -f(x.antes)
     delta.x <- solve(jacob, b)
     x.depois <- x.antes+delta.x
     if ( max(abs(delta.x))/max(abs(x.antes)) < rtol & mean(abs(b))< ftol ) stop
<- TRUE
     x.antes <- x.depois
  } #while
  return(c(sol=x.depois, niter=n.iter))
}
```

```
res <- Newton(eqmaxver, start=c(regpois$coefficients[1]-
0.5,regpois$coefficients[2]+0.5), rtol=1e-7, ftol=1e-7,
            nmax=1000, jac=jfun)

res

### teste
g <- function(x) {
  g1 <- (x[1])^2+4*(x[2])^2-4
  g2 <- x[2]-(x[1])^2+2*x[1]-0.5
  return(c(g1,g2))
}

jacob.g <- function(x) {
  jac.g <- matrix(nrow=2,ncol=2)
  jac.g[1,1] <- 2*x[1]
  jac.g[1,2] <- 8*x[2]
  jac.g[2,1] <- -2*x[1]+2
  jac.g[2,2] <- 1
  jac.g
}

Newton(g, start=c(2, 0.4), rtol=1e-7, ftol=1e-7, nmax=100, jacob.g)
###############

j<-1
for (i in 1:NROW(xte)){
 (yte[i]/(-2+2(xte[i]-xte[j]))-1)*(2*pi)^(-1/2)*exp(-((xte[j]-xte[i])/h)^2/2)
}


# experiência
```

```
f<-0
f
for (i in 1:NROW(X50)) {
f<-((YP50[i]/(3+X50[i]))+f)
}
f

{layout(matrix(1:4, nrow = 2, ncol =2))
plot(X50,YP50)
title(main = "Modelo de Poisson na amostra de dimensão n=50")
plot(X100,YP100)
title(main = "Modelo de Poisson na amostra de dimensão n=100")
plot(X500,YP500)
title(main = "Modelo de Poisson na amostra de dimensão n=500")
plot(X1000,YP1000)
title(main = "Modelo de Poisson na amostra de dimensão n=1000")
}

{layout(matrix(1:4, nrow = 2, ncol =2))
plot(X50,YZIP50)
title(main = "Modelo ZIP na amostra de dimensão n=50")
plot(X100,YZIP100)
title(main = "Modelo ZIP na amostra de dimensão n=100")
plot(X500,YZIP500)
title(main = "Modelo ZIP na amostra de dimensão n=500")
plot(X1000,YZIP1000)
title(main = "Modelo ZIP na amostra de dimensão n=1000")
}
```

Outputs para a amostra de dimensão n=500

Gerar uma amostra aleatória, de uma população com distribuição uniforme, de dimensão n=500

```
  [1]  1.5552610597 -0.2906678710  0.7821525708  0.2544105314  0.4614689006
  [6] -1.3661415344  0.6380086020 -0.5030691838 -1.9433566500 -0.2770328065
 [11]  1.9698722595 -1.2914058585  0.2853929000  0.2268352425  1.2733249441
 [16] -1.0944956085 -0.1241523046  1.5338596450  0.4378028363  1.9448204972
 [21] -0.7252396205 -0.3539301902 -0.2397365710 -0.7523612129 -1.0394747872
 [26] -1.6748485332 -0.7498433320  0.8802219508  1.7796558179 -1.0909403013
 [31] -1.3286445914  0.0238757236  0.8756370163 -0.7580434429 -0.7556381337
 [36]  1.4702994013 -0.1514449380 -1.2055959012 -1.1644607233 -1.9012938198
 [41]  1.5743736885  0.0190868704 -0.2411939967 -0.0262584174 -1.8878938193
 [46]  1.3335224753 -1.5566991204 -0.3955557486  1.6342669651  1.8749891529
 [51]  0.5106365345 -1.9840682838 -0.2929715365  1.2393192323  1.4307431020
 [56]  1.9336214233 -1.0066466229 -0.9231554857  1.1962311724  1.4732144279
 [61] -0.7231314881  1.3845976843  1.6542292386 -0.1485005151  0.4713229798
 [66] -1.6021703053 -0.5620127786  0.4316961151  0.2796088904 -0.1023562774
 [71] -0.7330317618  0.0557810236  0.9223792721 -0.0863437094 -0.7466074638
 [76] -1.4310262278  1.3177421466 -0.0603120560  0.2607739633  1.3964582561
 [81]  1.3111572713  1.1794723859  1.6835904941 -0.7328105513  1.4603884025
 [86] -1.9019501200 -0.8944221847  0.2792096166 -0.3345340639 -0.7311268561
 [91] -1.5944776759  0.0451569725 -1.6243047696  0.4580390360 -1.2525752289
 [96]  0.3257659608 -0.2654970568 -1.8959115390  0.6438409872 -1.8535263510
[101] -0.5222443361 -1.3424130762 -0.3608502708 -0.7390973428 -0.6407765159
[106]  1.7951614568  1.8846833128  0.0069091786 -1.2316931020 -1.9106408907
[111]  1.8938468890  1.4861562904 -1.9196180329  0.3010218330 -1.5243457416
[116] -0.5123414453  1.0619886331 -0.9540340668  1.2428393783  0.4639820829
[121] -0.0589395165  0.2351112422 -0.5078593101 -1.3486550804  0.3177595558
[126] -1.4988324828  1.8787411870 -1.8873504633 -1.8435044484  0.9179295227
[131]  0.9349761261 -0.5680217538 -1.2636075895  1.1968823802  0.7540375851
[136]  1.4128147177 -0.5346070258  1.9200324295  0.8186209183 -0.5146972826
[141] -1.2590122344  0.2058222955 -0.1486567408 -0.1037130905 -0.8627460822
[146]  0.2648915732  1.7388573661  0.6853713142 -0.6358503057 -0.1832473092
[151] -1.1887838319  0.4438125538 -0.9197471943  0.8990565157 -0.3155816663
```

```
[156]  0.8431073539 -1.2545678187 -1.8790090838 -1.1029655775  1.7261561742
[161]  1.9017274398  0.9111602008  1.7625917234 -1.5389626343  1.0614139568
[166]  1.5506159766  0.2057790039  0.6122750966  1.5548630431 -1.5087773781
[171]  1.3586392011 -0.5805416778 -0.2765890611 -0.5231985040  1.3784781154
[176]  0.9114765115  1.1323931729 -0.1657847846 -0.5477402220  1.0447829692
[181] -1.0325127598  0.1548267230 -1.8291341290  1.0588260684 -0.6212843042
[186]  0.9065069947 -1.1291594002 -1.0303278025 -0.8134907130  0.9654068518
[191]  0.8705616938 -0.3465562854 -0.7596423086  0.4176257523  0.3318414083
[196] -1.9876730954  0.7005367596  0.6690432625 -0.2870262656  1.2433747621
[201]  1.6986686103 -1.2188685359  1.1531054387  1.0360450596  1.5258220648
[206]  0.2018047730 -0.7506901119  0.5191077013  1.9093325902 -1.6607604334
[211]  0.1048358018 -0.2267035432  0.5120013971  1.9222490545 -0.5889518932
[216]  0.5785189960  0.3987571793 -0.8824971206  0.0652618464 -0.9210442230
[221]  1.0518621979  1.0078494083  1.2949817237 -0.8335859561 -0.5705325166
[226]  0.8611034108  0.2308024680 -0.6971622044  0.2145224828  0.7673702640
[231]  1.9865768449  0.9758534804  0.7329310887  0.5603367984 -1.7111623324
[236]  1.0547962468 -1.1564273424  1.0797651028  0.6856718902 -0.1492668856
[241]  0.9518587003  1.7249991531 -0.3948462950 -0.0246702712  0.5120298900
[246]  1.2788756285 -1.4905847544  0.1230181791  1.7395690074 -1.4334884081
[251]  1.1666485751  1.6988572329  1.1277482817  0.5535167065 -1.0926391874
[256] -1.5137981968  0.1608362189 -0.5666429913 -1.0425745025  1.3527375264
[261] -1.8406786611  1.7232444407 -1.7553690709  0.6588565055  0.8085244186
[266] -0.3798097949 -1.2845160859 -1.4463388165  0.2122144578  1.7239841921
[271]  1.5032988777 -0.2589602154 -0.5906142620  0.8882685583 -0.2113147918
[276] -0.3899352113 -0.2186396942 -1.3700812040 -0.4300882751  0.4285404235
[281]  1.4246193022  1.0879064780 -1.9888945371  0.5596687905 -0.6744020656
[286]  1.0987153938 -0.8051366759 -1.3605517279  0.9023747304  1.9569525337
[291] -1.7781735603 -0.8266688492 -1.0343094384  0.1149372803  0.3829342723
[296]  0.6434984393  1.0600393843 -1.1091988534 -0.4876024853  1.1914211009
[301]  0.0755594745  0.5138042821  0.5886177691 -1.0183729492 -0.2676783632
[306] -0.9493664214  0.6315865200  1.3957214328 -0.5416508904 -1.3539909925
[311] -0.2710118070 -0.5133050038 -0.4258985519  0.0721159521 -0.4064935641
[316] -0.5310153794  0.2260659961  0.1492355848 -0.5203227159 -0.6210526619
[321]  0.8930117730  1.5288454704 -1.9628244461 -0.4972227300  1.6951045077
[326]  0.4320750069  1.8477624832  0.9742371105  1.8808851903 -1.3339967774
[331]  1.9305161480 -0.6918117329  1.5751549192 -1.2159697581 -1.1892141849
[336]  0.0298828175 -0.9688815335  0.8648492992  1.3821047507  1.4968735110
[341] -1.3747336315  0.0030836957  0.1756704906  0.4425562751  0.5341751510
[346] -0.0007391293  0.9372286620 -1.3566575842 -0.8929946721 -1.0575518571
```

```
[351]  0.2164735533  1.1040358683 -0.8915083567  1.2703047311  0.9842019258
[356] -1.2475652024 -1.2973024724 -0.9133094363  1.3716988964  1.5548243206
[361]  1.5557315527 -0.9494122583 -1.4536209926  0.2468270995  1.2068235157
[366] -0.5961641017 -1.5902829096 -0.8186844504  1.2616111003  1.8002401888
[371]  0.3849349078  0.4924053010  1.4780560965  1.5728569077 -0.5153074516
[376] -1.8392150039 -0.8753285399 -1.2711635288  1.8721122062  1.5690716691
[381] -0.5089239739 -1.8420301797 -1.5844304124  1.7106058728 -1.5719678998
[386] -0.0466788039  1.7251089420 -1.7083596047  1.2073833821  1.5032830266
[391] -1.2617922621 -0.1211008485  0.3465107754 -0.4104193924  0.5263767075
[396] -0.8529571220 -0.6913358849 -0.9339801082 -1.3142890846 -0.1544003040
[401]  1.1687351698  1.7920104675 -1.0515553439  1.4420533404 -0.4938950399
[406]  0.1380602680 -0.1683841553  0.4914529314  0.9543143520 -1.1657101391
[411] -0.7944911020 -0.8158273790  0.0167965908 -0.4342621574 -0.6512672715
[416] -0.5769445589  1.0147974398  0.0389460977 -1.8705347730 -1.1435327418
[421] -0.9855729463 -1.7087521199 -0.2144808397 -0.4946054909 -1.2090629078
[426]  0.3710660944 -1.3761981772 -0.8095626999 -0.2030151039 -0.0169207407
[431]  0.2717077704 -0.6321320152  1.0377139198 -1.3364802375  1.6746739848
[436]  0.6920534652  1.4363250779 -0.6535932822  0.7462261058 -1.2767785676
[441]  1.2993008960 -1.1281138984  1.4593920810 -0.3007488912 -0.7117043780
[446]  1.2557431664  1.7235232880  0.4027466755  1.0636155596  1.4091287302
[451]  0.9132058285  0.2055824455 -1.7157374145  0.9831067026  0.3129122313
[456] -0.6597113525 -1.0852333652 -0.8569227373  1.6005729800  1.3059215853
[461] -0.5405174699 -1.4667036943 -1.7015448781  0.3873199532  0.3026662311
[466]  0.3825496230  0.1968805082 -1.6627365118 -1.2224188512  0.8050202075
[471] -1.2426965227  0.0604878031  1.7943919981  1.6542728571  0.1080579618
[476] -0.2080082223 -1.2980971504  1.0079940977  0.8640699517 -1.8877342325
[481] -0.2639799165 -1.3523944784 -0.9639573516 -1.6520704851 -0.2232643915
[486] -1.7897216463  1.0463432446 -0.1391797913 -0.1725199074 -1.9629208837
[491] -0.0168064972 -1.5085722078 -0.8720469344  1.5401114700 -1.4535376197
[496]  0.8361455658 -1.8476202199  1.5220045801  0.0434279665 -0.5488056745
```

Gerar uma amostra aleatória, de uma população com distribuição de Poisson de dimensão n=500

```
  [1] 3 0 2 0 0 0 1 0 0 0 6 0 0 1 0 0 0 2 1 8 0 1 0 0 0
 [26] 0 0 1 7 0 0 0 1 0 0 5 0 0 0 0 0 0 0 0 0 0 0 0 3 3
 [51] 0 0 0 1 1 6 0 0 1 4 0 1 3 0 0 0 0 1 0 1 1 0 1 0 0
```

[76] 0 0 0 0 1 1 0 4 0 1 0 0 0 0 0 0 0 0 1 0 0 0 0 1 0
[101] 0 0 0 0 0 6 6 0 0 0 4 2 0 0 0 0 1 0 2 0 0 0 0 0 0
[126] 0 7 0 0 0 0 0 0 2 0 1 0 5 0 0 0 0 0 1 0 0 3 1 0 0
[151] 0 0 0 1 0 0 0 0 0 2 4 1 5 1 0 4 0 0 3 0 4 0 0 0 1
[176] 1 2 0 0 0 0 1 0 0 0 0 0 0 0 2 0 0 0 0 1 0 0 1 0 1
[201] 6 0 1 0 2 0 0 0 4 0 0 0 0 11 0 0 0 0 0 0 3 1 1 0 0
[226] 1 0 0 0 0 5 0 1 0 0 2 0 2 0 0 1 8 1 0 1 5 1 0 5 0
[251] 0 1 1 2 0 0 0 0 0 2 0 3 0 0 0 0 0 0 0 7 6 0 0 1 0
[276] 0 0 0 0 0 1 3 0 0 0 1 0 0 1 7 0 0 0 0 0 0 2 0 0 2
[301] 0 1 0 0 0 0 0 1 0 0 1 0 0 0 0 0 1 0 0 0 0 2 0 0 0
[326] 0 8 3 8 0 3 0 3 0 0 0 0 1 1 3 0 0 1 0 0 0 0 0 0 0
[351] 1 1 0 0 0 0 0 0 4 4 4 0 0 0 3 0 0 0 1 6 1 0 2 5 0
[376] 0 0 0 4 6 0 0 0 3 0 0 2 0 1 3 0 1 0 0 0 0 0 0 0 0
[401] 3 5 0 5 0 0 0 1 1 0 0 0 0 0 0 0 2 0 0 0 0 0 0 0 0
[426] 0 0 0 0 0 0 0 2 0 7 0 2 0 1 0 5 0 0 0 0 1 8 1 0 2
[451] 1 0 0 2 0 0 0 0 4 3 0 0 0 1 0 0 0 0 0 2 0 1 6 7 0
[476] 0 0 0 1 0 0 0 0 0 0 0 2 1 0 0 0 0 0 3 0 0 0 1 1 0

Gerar amostra aleatória de uma população com distribuição ZIP, com ψ=0.2 de dimensão n=500

[1] 3 0 2 0 0 0 1 0 0 0 0 0 0 1 0 0 0 2 0 8 0 1 0 0 0
[26] 0 0 1 7 0 0 0 1 0 0 5 0 0 0 0 0 0 0 0 0 0 0 0 3 0
[51] 0 0 0 1 1 6 0 0 0 4 0 1 3 0 0 0 0 1 0 1 1 0 1 0 0
[76] 0 0 0 0 1 1 0 4 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
[101] 0 0 0 0 0 6 6 0 0 0 4 2 0 0 0 0 1 0 2 0 0 0 0 0 0
[126] 0 7 0 0 0 0 0 0 2 0 1 0 5 0 0 0 0 0 1 0 0 3 1 0 0
[151] 0 0 0 0 0 0 0 0 0 2 4 1 5 1 0 4 0 0 3 0 4 0 0 0 1
[176] 1 2 0 0 0 0 0 0 0 0 0 0 0 0 2 0 0 0 0 1 0 0 1 0 0
[201] 6 0 1 0 2 0 0 0 4 0 0 0 0 11 0 0 0 0 0 0 3 0 1 0 0
[226] 1 0 0 0 0 5 0 1 0 0 0 0 0 0 0 1 8 0 0 1 5 1 0 0 0
[251] 0 1 1 0 0 0 0 0 0 2 0 3 0 0 0 0 0 0 0 7 6 0 0 0 0

```
[276] 0 0 0 0 0 0 3 0 0 0 0 0 0 0 0 0 0 0 0 0 0 2 0 0 0
[301] 0 1 0 0 0 0 0 1 0 0 1 0 0 0 0 0 1 0 0 0 0 0 0 0 0
[326] 0 8 3 8 0 0 0 0 0 0 0 0 1 1 3 0 0 1 0 0 0 0 0 0 0
[351] 1 1 0 0 0 0 0 0 4 4 4 0 0 0 0 0 0 0 1 6 1 0 2 5 0
[376] 0 0 0 4 6 0 0 0 0 0 0 0 2 0 1 0 0 1 0 0 0 0 0 0 0 0
[401] 3 5 0 5 0 0 0 1 1 0 0 0 0 0 0 0 2 0 0 0 0 0 0 0 0
[426] 0 0 0 0 0 0 0 2 0 0 0 2 0 1 0 0 0 0 0 0 1 8 0 0 2
[451] 1 0 0 2 0 0 0 0 4 0 0 0 0 1 0 0 0 0 0 2 0 1 6 0 0
[476] 0 0 0 1 0 0 0 0 0 0 0 0 2 1 0 0 0 0 0 3 0 0 0 0 0 0
```

Matriz com o regressor (X500), resposta segundo o modelo de Poisson (YP500), segundo o modelo ZIP (YZIP500), e amostras de teste (TESTRE500=1) e de treino (TESTRE500=2)

```
              X500 YP500 YZIP500 TESTRE500
 [1,]  1.5552610597     3       3        2
 [2,] -0.2906678710     0       0        1
 [3,]  0.7821525708     2       2        2
 [4,]  0.2544105314     0       0        2
 [5,]  0.4614689006     0       0        2
 [6,] -1.3661415344     0       0        2
 [7,]  0.6380086020     1       1        2
 [8,] -0.5030691838     0       0        2
 [9,] -1.9433566500     0       0        2
[10,] -0.2770328065     0       0        2
[11,]  1.9698722595     6       0        2
[12,] -1.2914058585     0       0        2
[13,]  0.2853929000     0       0        2
[14,]  0.2268352425     1       1        2
[15,]  1.2733249441     0       0        1
[16,] -1.0944956085     0       0        1
[17,] -0.1241523046     0       0        1
```

[18,]	1.5338596450	2	2	1
[19,]	0.4378028363	1	0	2
[20,]	1.9448204972	8	8	2
[21,]	-0.7252396205	0	0	2
[22,]	-0.3539301902	1	1	2
[23,]	-0.2397365710	0	0	2
[24,]	-0.7523612129	0	0	1
[25,]	-1.0394747872	0	0	1
[26,]	-1.6748485332	0	0	2
[27,]	-0.7498433320	0	0	2
[28,]	0.8802219508	1	1	2
[29,]	1.7796558179	7	7	2
[30,]	-1.0909403013	0	0	2
[31,]	-1.3286445914	0	0	2
[32,]	0.0238757236	0	0	2
[33,]	0.8756370163	1	1	2
[34,]	-0.7580434429	0	0	2
[35,]	-0.7556381337	0	0	2
[36,]	1.4702994013	5	5	2
[37,]	-0.1514449380	0	0	2
[38,]	-1.2055959012	0	0	1
[39,]	-1.1644607233	0	0	2
[40,]	-1.9012938198	0	0	2
[41,]	1.5743736885	0	0	2
[42,]	0.0190868704	0	0	2
[43,]	-0.2411939967	0	0	2
[44,]	-0.0262584174	0	0	1
[45,]	-1.8878938193	0	0	2
[46,]	1.3335224753	0	0	2
[47,]	-1.5566991204	0	0	1
[48,]	-0.3955557486	0	0	2
[49,]	1.6342669651	3	3	2

[50,]	1.8749891529	3	0	1
[51,]	0.5106365345	0	0	2
[52,]	-1.9840682838	0	0	2
[53,]	-0.2929715365	0	0	1
[54,]	1.2393192323	1	1	2
[55,]	1.4307431020	1	1	2
[56,]	1.9336214233	6	6	1
[57,]	-1.0066466229	0	0	1
[58,]	-0.9231554857	0	0	2
[59,]	1.1962311724	1	0	2
[60,]	1.4732144279	4	4	1
[61,]	-0.7231314881	0	0	2
[62,]	1.3845976843	1	1	2
[63,]	1.6542292386	3	3	2
[64,]	-0.1485005151	0	0	1
[65,]	0.4713229798	0	0	2
[66,]	-1.6021703053	0	0	2
[67,]	-0.5620127786	0	0	2
[68,]	0.4316961151	1	1	1
[69,]	0.2796088904	0	0	2
[70,]	-0.1023562774	1	1	2
[71,]	-0.7330317618	1	1	2
[72,]	0.0557810236	0	0	1
[73,]	0.9223792721	1	1	1
[74,]	-0.0863437094	0	0	2
[75,]	-0.7466074638	0	0	2
[76,]	-1.4310262278	0	0	2
[77,]	1.3177421466	0	0	2
[78,]	-0.0603120560	0	0	2
[79,]	0.2607739633	0	0	2
[80,]	1.3964582561	1	1	2
[81,]	1.3111572713	1	1	2

```
 [82,]  1.1794723859    0    0    2
 [83,]  1.6835904941    4    4    1
 [84,] -0.7328105513    0    0    1
 [85,]  1.4603884025    1    1    2
 [86,] -1.9019501200    0    0    2
 [87,] -0.8944221847    0    0    1
 [88,]  0.2792096166    0    0    2
 [89,] -0.3345340639    0    0    2
 [90,] -0.7311268561    0    0    2
 [91,] -1.5944776759    0    0    2
 [92,]  0.0451569725    0    0    2
 [93,] -1.6243047696    0    0    2
 [94,]  0.4580390360    1    0    2
 [95,] -1.2525752289    0    0    2
 [96,]  0.3257659608    0    0    2
 [97,] -0.2654970568    0    0    2
 [98,] -1.8959115390    0    0    1
 [99,]  0.6438409872    1    0    2
[100,] -1.8535263510    0    0    2
[101,] -0.5222443361    0    0    2
[102,] -1.3424130762    0    0    2
[103,] -0.3608502708    0    0    1
[104,] -0.7390973428    0    0    2
[105,] -0.6407765159    0    0    2
[106,]  1.7951614568    6    6    2
[107,]  1.8846833128    6    6    2
[108,]  0.0069091786    0    0    2
[109,] -1.2316931020    0    0    2
[110,] -1.9106408907    0    0    2
[111,]  1.8938468890    4    4    2
[112,]  1.4861562904    2    2    2
[113,] -1.9196180329    0    0    2
```

[114,]	0.3010218330	0	0	2
[115,]	-1.5243457416	0	0	2
[116,]	-0.5123414453	0	0	2
[117,]	1.0619886331	1	1	2
[118,]	-0.9540340668	0	0	1
[119,]	1.2428393783	2	2	2
[120,]	0.4639820829	0	0	2
[121,]	-0.0589395165	0	0	1
[122,]	0.2351112422	0	0	2
[123,]	-0.5078593101	0	0	2
[124,]	-1.3486550804	0	0	2
[125,]	0.3177595558	0	0	2
[126,]	-1.4988324828	0	0	1
[127,]	1.8787411870	7	7	2
[128,]	-1.8873504633	0	0	1
[129,]	-1.8435044484	0	0	2
[130,]	0.9179295227	0	0	2
[131,]	0.9349761261	0	0	1
[132,]	-0.5680217538	0	0	2
[133,]	-1.2636075895	0	0	2
[134,]	1.1968823802	2	2	2
[135,]	0.7540375851	0	0	2
[136,]	1.4128147177	1	1	1
[137,]	-0.5346070258	0	0	2
[138,]	1.9200324295	5	5	2
[139,]	0.8186209183	0	0	2
[140,]	-0.5146972826	0	0	1
[141,]	-1.2590122344	0	0	2
[142,]	0.2058222955	0	0	2
[143,]	-0.1486567408	0	0	2
[144,]	-0.1037130905	1	1	1
[145,]	-0.8627460822	0	0	2

[146,]	0.2648915732	0	0	1
[147,]	1.7388573661	3	3	2
[148,]	0.6853713142	1	1	2
[149,]	-0.6358503057	0	0	2
[150,]	-0.1832473092	0	0	2
[151,]	-1.1887838319	0	0	2
[152,]	0.4438125538	0	0	2
[153,]	-0.9197471943	0	0	2
[154,]	0.8990565157	1	0	2
[155,]	-0.3155816663	0	0	2
[156,]	0.8431073539	0	0	2
[157,]	-1.2545678187	0	0	2
[158,]	-1.8790090838	0	0	2
[159,]	-1.1029655775	0	0	2
[160,]	1.7261561742	2	2	2
[161,]	1.9017274398	4	4	2
[162,]	0.9111602008	1	1	2
[163,]	1.7625917234	5	5	1
[164,]	-1.5389626343	1	1	2
[165,]	1.0614139568	0	0	2
[166,]	1.5506159766	4	4	2
[167,]	0.2057790039	0	0	1
[168,]	0.6122750966	0	0	2
[169,]	1.5548630431	3	3	2
[170,]	-1.5087773781	0	0	1
[171,]	1.3586392011	4	4	2
[172,]	-0.5805416778	0	0	2
[173,]	-0.2765890611	0	0	1
[174,]	-0.5231985040	0	0	2
[175,]	1.3784781154	1	1	2
[176,]	0.9114765115	1	1	2
[177,]	1.1323931729	2	2	1

[178,]	-0.1657847846	0	0	2
[179,]	-0.5477402220	0	0	1
[180,]	1.0447829692	0	0	2
[181,]	-1.0325127598	0	0	2
[182,]	0.1548267230	1	0	2
[183,]	-1.8291341290	0	0	1
[184,]	1.0588260684	0	0	2
[185,]	-0.6212843042	0	0	2
[186,]	0.9065069947	0	0	2
[187,]	-1.1291594002	0	0	2
[188,]	-1.0303278025	0	0	2
[189,]	-0.8134907130	0	0	2
[190,]	0.9654068518	2	2	2
[191,]	0.8705616938	0	0	2
[192,]	-0.3465562854	0	0	2
[193,]	-0.7596423086	0	0	2
[194,]	0.4176257523	0	0	2
[195,]	0.3318414083	1	1	2
[196,]	-1.9876730954	0	0	1
[197,]	0.7005367596	0	0	2
[198,]	0.6690432625	1	1	1
[199,]	-0.2870262656	0	0	2
[200,]	1.2433747621	1	0	2
[201,]	1.6986686103	6	6	2
[202,]	-1.2188685359	0	0	2
[203,]	1.1531054387	1	1	2
[204,]	1.0360450596	0	0	2
[205,]	1.5258220648	2	2	2
[206,]	0.2018047730	0	0	2
[207,]	-0.7506901119	0	0	2
[208,]	0.5191077013	0	0	2
[209,]	1.9093325902	4	4	2

[210,]	-1.6607604334	0	0	2
[211,]	0.1048358018	0	0	2
[212,]	-0.2267035432	0	0	2
[213,]	0.5120013971	0	0	2
[214,]	1.9222490545	11	11	2
[215,]	-0.5889518932	0	0	2
[216,]	0.5785189960	0	0	2
[217,]	0.3987571793	0	0	2
[218,]	-0.8824971206	0	0	2
[219,]	0.0652618464	0	0	2
[220,]	-0.9210442230	0	0	2
[221,]	1.0518621979	3	3	2
[222,]	1.0078494083	1	0	1
[223,]	1.2949817237	1	1	1
[224,]	-0.8335859561	0	0	2
[225,]	-0.5705325166	0	0	1
[226,]	0.8611034108	1	1	2
[227,]	0.2308024680	0	0	2
[228,]	-0.6971622044	0	0	1
[229,]	0.2145224828	0	0	2
[230,]	0.7673702640	0	0	2
[231,]	1.9865768449	5	5	2
[232,]	0.9758534804	0	0	2
[233,]	0.7329310887	1	1	2
[234,]	0.5603367984	0	0	2
[235,]	-1.7111623324	0	0	2
[236,]	1.0547962468	2	0	2
[237,]	-1.1564273424	0	0	2
[238,]	1.0797651028	2	0	1
[239,]	0.6856718902	0	0	1
[240,]	-0.1492668856	0	0	1
[241,]	0.9518587003	1	1	2

[242,]	1.7249991531	8	8	2
[243,]	-0.3948462950	1	0	2
[244,]	-0.0246702712	0	0	2
[245,]	0.5120298900	1	1	1
[246,]	1.2788756285	5	5	2
[247,]	-1.4905847544	1	1	2
[248,]	0.1230181791	0	0	2
[249,]	1.7395690074	5	0	2
[250,]	-1.4334884081	0	0	2
[251,]	1.1666485751	0	0	2
[252,]	1.6988572329	1	1	2
[253,]	1.1277482817	1	1	2
[254,]	0.5535167065	2	0	1
[255,]	-1.0926391874	0	0	1
[256,]	-1.5137981968	0	0	2
[257,]	0.1608362189	0	0	2
[258,]	-0.5666429913	0	0	1
[259,]	-1.0425745025	0	0	2
[260,]	1.3527375264	2	2	2
[261,]	-1.8406786611	0	0	2
[262,]	1.7232444407	3	3	2
[263,]	-1.7553690709	0	0	1
[264,]	0.6588565055	0	0	2
[265,]	0.8085244186	0	0	2
[266,]	-0.3798097949	0	0	2
[267,]	-1.2845160859	0	0	2
[268,]	-1.4463388165	0	0	2
[269,]	0.2122144578	0	0	1
[270,]	1.7239841921	7	7	2
[271,]	1.5032988777	6	6	2
[272,]	-0.2589602154	0	0	2
[273,]	-0.5906142620	0	0	2

[274,]	0.8882685583	1	0	1
[275,]	-0.2113147918	0	0	2
[276,]	-0.3899352113	0	0	2
[277,]	-0.2186396942	0	0	2
[278,]	-1.3700812040	0	0	2
[279,]	-0.4300882751	0	0	1
[280,]	0.4285404235	0	0	2
[281,]	1.4246193022	1	0	2
[282,]	1.0879064780	3	3	2
[283,]	-1.9888945371	0	0	2
[284,]	0.5596687905	0	0	1
[285,]	-0.6744020656	0	0	2
[286,]	1.0987153938	1	0	2
[287,]	-0.8051366759	0	0	1
[288,]	-1.3605517279	0	0	2
[289,]	0.9023747304	1	0	2
[290,]	1.9569525337	7	0	2
[291,]	-1.7781735603	0	0	2
[292,]	-0.8266688492	0	0	2
[293,]	-1.0343094384	0	0	2
[294,]	0.1149372803	0	0	2
[295,]	0.3829342723	0	0	2
[296,]	0.6434984393	0	0	2
[297,]	1.0600393843	2	2	2
[298,]	-1.1091988534	0	0	2
[299,]	-0.4876024853	0	0	2
[300,]	1.1914211009	2	0	2
[301,]	0.0755594745	0	0	1
[302,]	0.5138042821	1	1	1
[303,]	0.5886177691	0	0	2
[304,]	-1.0183729492	0	0	2
[305,]	-0.2676783632	0	0	1

[306,]	-0.9493664214	0	0	2
[307,]	0.6315865200	0	0	1
[308,]	1.3957214328	1	1	2
[309,]	-0.5416508904	0	0	2
[310,]	-1.3539909925	0	0	1
[311,]	-0.2710118070	1	1	2
[312,]	-0.5133050038	0	0	2
[313,]	-0.4258985519	0	0	1
[314,]	0.0721159521	0	0	2
[315,]	-0.4064935641	0	0	2
[316,]	-0.5310153794	0	0	2
[317,]	0.2260659961	1	1	2
[318,]	0.1492355848	0	0	2
[319,]	-0.5203227159	0	0	2
[320,]	-0.6210526619	0	0	2
[321,]	0.8930117730	0	0	1
[322,]	1.5288454704	2	0	2
[323,]	-1.9628244461	0	0	2
[324,]	-0.4972227300	0	0	2
[325,]	1.6951045077	0	0	2
[326,]	0.4320750069	0	0	2
[327,]	1.8477624832	8	8	2
[328,]	0.9742371105	3	3	2
[329,]	1.8808851903	8	8	2
[330,]	-1.3339967774	0	0	2
[331,]	1.9305161480	3	0	1
[332,]	-0.6918117329	0	0	1
[333,]	1.5751549192	3	0	2
[334,]	-1.2159697581	0	0	1
[335,]	-1.1892141849	0	0	2
[336,]	0.0298828175	0	0	2
[337,]	-0.9688815335	0	0	1

[338,]	0.8648492992	1	1	2
[339,]	1.3821047507	1	1	1
[340,]	1.4968735110	3	3	2
[341,]	-1.3747336315	0	0	2
[342,]	0.0030836957	0	0	2
[343,]	0.1756704906	1	1	2
[344,]	0.4425562751	0	0	1
[345,]	0.5341751510	0	0	2
[346,]	-0.0007391293	0	0	2
[347,]	0.9372286620	0	0	2
[348,]	-1.3566575842	0	0	1
[349,]	-0.8929946721	0	0	2
[350,]	-1.0575518571	0	0	2
[351,]	0.2164735533	1	1	2
[352,]	1.1040358683	1	1	2
[353,]	-0.8915083567	0	0	2
[354,]	1.2703047311	0	0	2
[355,]	0.9842019258	0	0	2
[356,]	-1.2475652024	0	0	2
[357,]	-1.2973024724	0	0	2
[358,]	-0.9133094363	0	0	2
[359,]	1.3716988964	4	4	2
[360,]	1.5548243206	4	4	2
[361,]	1.5557315527	4	4	2
[362,]	-0.9494122583	0	0	2
[363,]	-1.4536209926	0	0	2
[364,]	0.2468270995	0	0	2
[365,]	1.2068235157	3	0	2
[366,]	-0.5961641017	0	0	2
[367,]	-1.5902829096	0	0	2
[368,]	-0.8186844504	0	0	2
[369,]	1.2616111003	1	1	2

[370,]	1.8002401888	6	6	1
[371,]	0.3849349078	1	1	2
[372,]	0.4924053010	0	0	2
[373,]	1.4780560965	2	2	2
[374,]	1.5728569077	5	5	1
[375,]	-0.5153074516	0	0	2
[376,]	-1.8392150039	0	0	2
[377,]	-0.8753285399	0	0	1
[378,]	-1.2711635288	0	0	2
[379,]	1.8721122062	4	4	2
[380,]	1.5690716691	6	6	2
[381,]	-0.5089239739	0	0	2
[382,]	-1.8420301797	0	0	2
[383,]	-1.5844304124	0	0	2
[384,]	1.7106058728	3	0	2
[385,]	-1.5719678998	0	0	2
[386,]	-0.0466788039	0	0	2
[387,]	1.7251089420	2	2	2
[388,]	-1.7083596047	0	0	2
[389,]	1.2073833821	1	1	1
[390,]	1.5032830266	3	0	2
[391,]	-1.2617922621	0	0	2
[392,]	-0.1211008485	1	1	1
[393,]	0.3465107754	0	0	1
[394,]	-0.4104193924	0	0	2
[395,]	0.5263767075	0	0	1
[396,]	-0.8529571220	0	0	2
[397,]	-0.6913358849	0	0	1
[398,]	-0.9339801082	0	0	2
[399,]	-1.3142890846	0	0	2
[400,]	-0.1544003040	0	0	1
[401,]	1.1687351698	3	3	2

[402,]	1.7920104675	5	5	2
[403,]	-1.0515553439	0	0	2
[404,]	1.4420533404	5	5	2
[405,]	-0.4938950399	0	0	2
[406,]	0.1380602680	0	0	2
[407,]	-0.1683841553	0	0	2
[408,]	0.4914529314	1	1	2
[409,]	0.9543143520	1	1	2
[410,]	-1.1657101391	0	0	2
[411,]	-0.7944911020	0	0	2
[412,]	-0.8158273790	0	0	2
[413,]	0.0167965908	0	0	2
[414,]	-0.4342621574	0	0	2
[415,]	-0.6512672715	0	0	2
[416,]	-0.5769445589	0	0	1
[417,]	1.0147974398	2	2	2
[418,]	0.0389460977	0	0	2
[419,]	-1.8705347730	0	0	2
[420,]	-1.1435327418	0	0	1
[421,]	-0.9855729463	0	0	2
[422,]	-1.7087521199	0	0	2
[423,]	-0.2144808397	0	0	1
[424,]	-0.4946054909	0	0	2
[425,]	-1.2090629078	0	0	1
[426,]	0.3710660944	0	0	1
[427,]	-1.3761981772	0	0	2
[428,]	-0.8095626999	0	0	1
[429,]	-0.2030151039	0	0	1
[430,]	-0.0169207407	0	0	2
[431,]	0.2717077704	0	0	2
[432,]	-0.6321320152	0	0	2
[433,]	1.0377139198	2	2	2

[434,]	-1.3364802375	0	0	2
[435,]	1.6746739848	7	0	2
[436,]	0.6920534652	0	0	2
[437,]	1.4363250779	2	2	2
[438,]	-0.6535932822	0	0	2
[439,]	0.7462261058	1	1	2
[440,]	-1.2767785676	0	0	1
[441,]	1.2993008960	5	0	2
[442,]	-1.1281138984	0	0	2
[443,]	1.4593920810	0	0	2
[444,]	-0.3007488912	0	0	2
[445,]	-0.7117043780	0	0	2
[446,]	1.2557431664	1	1	2
[447,]	1.7235232880	8	8	1
[448,]	0.4027466755	1	0	2
[449,]	1.0636155596	0	0	2
[450,]	1.4091287302	2	2	1
[451,]	0.9132058285	1	1	2
[452,]	0.2055824455	0	0	2
[453,]	-1.7157374145	0	0	2
[454,]	0.9831067026	2	2	2
[455,]	0.3129122313	0	0	2
[456,]	-0.6597113525	0	0	2
[457,]	-1.0852333652	0	0	2
[458,]	-0.8569227373	0	0	2
[459,]	1.6005729800	4	4	2
[460,]	1.3059215853	3	0	2
[461,]	-0.5405174699	0	0	2
[462,]	-1.4667036943	0	0	2
[463,]	-1.7015448781	0	0	2
[464,]	0.3873199532	1	1	2
[465,]	0.3026662311	0	0	2

```
[466,]  0.3825496230    0    0    2
[467,]  0.1968805082    0    0    2
[468,] -1.6627365118    0    0    2
[469,] -1.2224188512    0    0    2
[470,]  0.8050202075    2    2    2
[471,] -1.2426965227    0    0    2
[472,]  0.0604878031    1    1    2
[473,]  1.7943919981    6    6    2
[474,]  1.6542728571    7    0    1
[475,]  0.1080579618    0    0    2
[476,] -0.2080082223    0    0    2
[477,] -1.2980971504    0    0    1
[478,]  1.0079940977    0    0    2
[479,]  0.8640699517    1    1    2
[480,] -1.8877342325    0    0    2
[481,] -0.2639799165    0    0    2
[482,] -1.3523944784    0    0    2
[483,] -0.9639573516    0    0    2
[484,] -1.6520704851    0    0    1
[485,] -0.2232643915    0    0    2
[486,] -1.7897216463    0    0    2
[487,]  1.0463432446    2    2    2
[488,] -0.1391797913    1    1    2
[489,] -0.1725199074    0    0    1
[490,] -1.9629208837    0    0    2
[491,] -0.0168064972    0    0    1
[492,] -1.5085722078    0    0    1
[493,] -0.8720469344    0    0    2
[494,]  1.5401114700    3    3    1
[495,] -1.4535376197    0    0    2
[496,]  0.8361455658    0    0    2
[497,] -1.8476202199    0    0    2
```

```
[498,]  1.5220045801    1    0    2
[499,]  0.0434279665    1    0    2
[500,] -0.5488056745    0    0    1
```

8. Bibliografia

Antoniadis, A., I. Gijbels e M. Nikolova (2011). "Penalized likelihood regression for generalized linear models with non-quadratic penalties." ANNALS OF THE INSTITUTE OF STATISTICAL MATHEMATICS **63**(3): 585-615.

Aydin, D. (2007). A Comparison of the Nonparametric Regression Models using Smoothing Spline and Kernel Regression. Conference of the World-Academy-of-Science-Engineering-and-Technology, Banguecoque.

Besag, J. (1975). "Statistical-analysis of non-lattice data." STATISTICIAN **24**(3): 179-195.

Cleveland, W. (1979). " Robust locally weidhted regression and smoothing scatterplots." JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION **74**(368): 829-836.

Cottet, R., R. Kohn e D. Nott (2008). "Variable selection and model averaging in semiparametric overdispersed generalized linear models." JOURNAL OF THE AMERICAL STATISTICAL ASSOSSIATION **103**(482): 661-671.

Coxe, S., S. West e L. Aiken (2009). "The Analysis of Count Data: A Gentle Introduction to Poisson Regression and its Alternatives." JOURNAL OF PERSONALITY ASSESSMENT **91**(2): 121-136.

Dean, C., F. Nathoo e J. Nielsen (2006). Spatial and mixture models for recurrent event processes. 3rd International Workshop on Spatio-Temporal Modelling. ENVIRONMETRICS. Pamplona, Espanha, JOHN WILEY & SONS LTD, THE ATRIUM, SOUTHERN GATE, CHICHESTER PO19 8SQ, W SUSSEX, ENGLAND. **18:** 713-725.

Deng, D. e S. Paul (2005). "Score tests for zero-inflation and over-dispersion in generalized linear models." STATISTICA SINICA **15**(1): 257-276.

Eguchi, S., T. Kim e B. Park (2003). " Local likelihood method: A bridge over parametric and nonparametric regression." JOURNAL OF NONPARAMETRIC STATISTICS **15**(6): 665-683.

Eubank, R. (1999). *Nonparametric Regression and Smoothing Spline*, Nova York, NY. Marcel Dekker.

Fan, J., M. Farmen e I. Gijbels (1998). "Local maximum likelihood estimation and inference." JOURNAL OF THE ROYAL STATISTICAL SOCIETY, SERIE B (STATISTICAL METHODOLOGY) **60**(3): 591-608.

Garay, A., E. Hashimoto, E. Ortega e V. Lachos (2011). "On estimation and influence diagnostics for zero-inflated negative binomial regression models." COMPUTATIONAL STATISTICS & DATA ANALYSIS **55**(3): 1304-1318.

Green, P. e B. Silverman (1994). *Nonparametric Regression and Generalized Linear Models. A Roughness Penalty Approach*. Londres, Chapman and Hall.

Gschlossl, S. e C. Czado (2008). "Modelling count data with overdispersion and spatial effects." STATISTICAL PAPERS **49**(3): 531-552.

Guikema, S. e J. Goffelt (2008). "A flexible count data regression model for risk analysis." RISK ANALYSIS **28**(1): 213-223.

Hall, D. (2000). "Zero-inflated Poisson and binomial regression with random effects: A case study." BIOMETRICS **56**(4): 1030-1039.

Hall, D. e K. Berenhaut (2002). "Score tests for heterogeneity and overdispersion in zero-inflated Poisson and binomial regression models." CANADIAN JOURNAL OF STATISTICS-REVUE CANADIENNE DE STATISTIQUE **30**(3): 415-430.

Hall, D. e J. Shen (2010). "Robust Estimation for Zero-Inflated Poisson Regression." SCANDINAVIAN JOURNAL OF STATISTICS **37**(2): 237-252.

Hohle, M. e M. Paul (2008). "Count data regression charts for the monitoring of surveillance time series." COMPUTATIONAL STATISTICS & DATA ANALYSIS **52**(9): 4357-4368.

Huang, C., M. Wang e Y. Zhang (2006). "Analysing panel count data with informative observation times." BIOMETRIKA **93**(4): 763-775.

Ishwaran, H. e L. James (2004). "Computational methods for multiplicative intensity models using weigthed gamma processes: proportional hazards, market point processes, and count panel data." JOURNAL OF THE AMERICAL STATISTICAL ASSOSSIATION **99**(465): 175-190.

Jansakul, N. e J. Hinde (2009). "Score Tests for Extra-Zero Models in Zero-Inflated Negative Binomial Models." COMMUNICATIONS IN STATISTICS-SIMULATION AND COMPUTATION **38**(1): 92-108.

Joe, H. e R. Zhu (2005). "Generalized Poisson distribution: the property of mixture of Poisson and comparison with negative binomial distribution." BIOMEDICAL JOURNAL **47**(2): 219-229.

Jung, B., M. Jhun e J. Lee (2005). "Bootstrap tests for overdispersion in a zero-inflated Poisson regression model." BIOMETRICS **61**(2): 626-628.

Karlis, D. e E. Xekalaki (2005). "Mixed Poisson distributions." INTERNATIONAL STATISTICAL REVIEW **73**(1): 35-58.

Kim, Y. (2006). "Analysis of panel count data with dependent observation times." COMMUNICATIONS IN STATISTICS - SIMULATION AND COMPUTATION **35**(4): 983-990.

Kohler, M. e A. Krzyzak (2007). "Asymptotic confidence intervals for Poisson regression." JOURNAL OF MULTIVARIATE ANALYSIS **98**(5): 1072-1094.

Lam, K., H. Xue e Y. Cheung (2006). "Semiparametric analysis of zero-inflated count data." BIOMETRICS **62**(4): 996-1003.

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

Lee, A., K. Wang, J. Scott, K. Yau e G. McLachlan (2006). "Multi-level zero-inflated Poisson regression of correlated count data with excess zeros." STATISTICAL METHODS IN MEDICAL RESEARCH **15**(1): 47-61.

Lee, J., B. Jung e S. Jin (2009). "Tests for zero inflation in a bivariate zero-inflated Poisson model." STATISTICA NEERLANDICA **63**(4): 400-417.

McCarthy, M., S. Zeger, R. Ding, D. Aronsky, N. Hoot e G. Kelen (2008). "The challenge of predicting demand for emergency department services." ACADEMIC EMERGENCY MEDICINE **15**(4): 337-345.

Messer, K. (1991). "A Comparison of a Spline Estimate to its Equivalent Kernel Estimate " THE ANNALS OF STATISTICS **19**(2): 817-829.

Min, Y. e A. Agresti (2005). "Random effect models for repeated measures of zero-inflated count data." STATISTICAL MODELLING **5**(1): 1-19.

Mir, K. (2011). "Estimation in truncated Poisson distribution." MATHEMATICA SLOVACA **61**(2): 289-296.

Moghimbeiji, A., M. Eshraghian, K. Mohammad e B. McArdle (2008). "Multilevel zero-inflated negative binomial regression modeling for over-dispersed count data with extra zeros." JOURNAL OF APPLIED STATISTICS **35**(10).

Mukhopadhyay, K. e L. Marsh (2006). "An approach to nonparametric smoothing techniques for regressions with discrete data." APPLIED ECONOMICS **38**(301-305).

Naya, H., J. Urioste, Y. Chang, M. Rodrigues-Motta e D. Gianola (2008). "A comparison between Poisson and zero-inflated Poisson regression models with an application to number of black spots in Corriedale sheep." GENETICS SELECTION EVOLUTION **40**(4).

Nelder, J. e D. Pregibon (1987). "An extended quasi-likelihood function." BIOMETRIKA **74**(2): 221-232.

Nielsen, J. e C. Dean (2008). "Adaptative functional mixed NHPP models for the analysis of recurrent event panel data." COMPUTATIONAL STATISTICS & DATA ANALYSIS **52**(7): 3670-3675.

Nielsen, J. e C. Dean (2008). "Clustered mixed nonhomogenous Poisson process spline models for the analysis of recurrent event panel data." BIOMETRICS **64**(3): 751-761.

Paulsen, J. e K. Stubo (2011). "On Maximum Likelihood and Pseudo-Maximum Likelihood Estimation in Compound Insurance Models With Deductibles." ASTIN BULLETIN **41**(1): 1-28.

Podlich, H., M. Faddy e G. Smith (2004). "Semi-parametric extended Poisson process models for count data." STATISTICS AND COMPUTING **14**(1): 311-321.

Ridout, M., C. Demetrio e J. Hinde (1998). Models for count data with many zeros. XIXth International Biometric Conference, Cidade do Cabo.

Ridout, M., J. Hinde e C. Demetrio (2001). "A score test for testing a zero-inflated Poisson regression model against zero-inflated negative binomial alternatives." BIOMETRICS **57**(1): 219-223.

Rigby, R., D. Stanisopoulos e C. Akantziliotou (2008). "A framework for modelling overdispersed count data including the Poisson-shifted generalized inverse Gaussian distribution." COMPUTATIONAL STATISTICS & DATA ANALYSIS **53**(2): 381-393.

Saha, K. K. (2008). "Semiparametric estimation for the dispersion parameter in the analysis of over- or underdispersed count data." Journal of Applied Statistics **35**(11-12): 1383-1397.

Santos, J. A. e M. M. Neves (2008). "A local maximum likelihood estimator for Poisson regression." METRIKA **68**(3): 257-270.

Santos, J. A. R. A. (2005). Regressão Não Paramétrica em Modelos de Regressão de Dados de Contagem com Excesso de Zeros. **Doutoramento**.

Scheid, F. (1988). *Schaum's Outline of Theory and Problems of NUMERICAL ANALYSIS, 2nd edition*; trad. "Análise Numérica" António César de Freitas: Lisboa, McGraw-Hill Portugal, 2ª edição (2000).

Silva, J. e S. Tenreyro (2011). "Further simulation evidence on the performance of the Poisson pseudo-maximum likelihood estimator." ECONOMICS LETTERS **112**(2): 220-222.

Silverman, B. (1986). *Density estimation for statistics and data analysis*. Oxford, Chapman and Hall.

Simonoff, J. (1996). *Smoothing methods in Statistics*. Nova Iorque, Springer.

Stone, C., P. Bickel e L. Breiman (1977). " Consistent nonparametric regression." ANNALS OF STATISTICS **5**(4): 595-645.

Sun, J., D. Park, L. Sun e X. Zhao (2005). "Semiparametric regression analysis of longitudinal data with informative observation times." JOURNAL OF THE AMERICAL STATISTICAL ASSOSSIATION **100**(471): 882-889.

Tibshirani, R. e T. Hastie (1987). "Local Likelihood Estimation." JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION **82**(398): 559-567.

Tsou, T. (2006). "Robust Poisson regression." JOURNAL OF STATISTICAL PLANNING AND INFERENCE **136**(9): 3173-3186.

van Duijn, M. A., K. Gile e M. Handcock (2007). Comparison of Maximum Pseudo Likelihood and Maximum Likelihood Estimation of Exponential Family Random Graph Models. Seattle, WA, Center for Statistics and the Social Sciences - University of Washington.

Wahba, G. (1990). Spline Model for Observational Data. Filadelfia, Pa, SIAM.

Wand, M. e M. Jones (1995). *Kernel Smoothing*. Londres, Chapman and Hall.

Wang, N., R. Carrol e X. Lin (2005). "Efficient semiparametric marginal estimation for longitudinal/clustered data." JOURNAL OF THE AMERICAN STATISTICAL ASSOSSIATION **100**(469): 147-157.

Wedderburn, R. (1974). " Quasi-likelihood functions, generalized linear-models, and Gauss-Newton method." BIOMETRIKA **61**(3): 439-447.

Winkelmann, R. (2000). *Econometric Analysis of Count Data*. Berlim, Springer.

Xie, F., B. Wei e J. Lin (2008). "Assessing influence in pharmaceutical data in zero-inflated generalized Poisson mixed models." STATISTICS IN MEDICINE **27**(18): 3656-3673.

Yang, Z., J. Hardin e C. Addy (2009). "Testing overdispersion in the zero-inflated Poisson model." JOURNAL OF STATISTICAL PLANNING AND INFERENCE **139**(9): 3340-3353.

Yang, Z., J. Hardin e C. Addy (2010). "Score Tests for Zero-Inflation in Overdispersed Count Data." COMMUNICATIONS IN STATISTICS-THEORY AND METHODS **39**(11): 2008-2030.

Yang, Z., J. Hardin e C. Addy (2010). "Some Remarks on Testing Overdispersion in Zero-Inflated Poisson and Binomial Regression Models." COMMUNICATIONS IN STATISTICS-THEORY AND METHODS **39**(15): 2743-2752.

Yip, K. e K. Yau (2005). "On modeling claim frequency data in general insurance with extra zeros." INSURANCE MATHEMATICS & ECONOMICS **36**(2): 153-163.

Zhao, Q. e J. Sun (2006). "Semiparametric and nonparametric analysis of recurrent events with observation gaps." COMPUTATIONAL STATISTICS & DATA ANALYSIS **51**(3): 1924-1933.