Análise de dados de contagem correlacionados através da distribuição Poisson bivariada

Jonas B Alonso[*], **Luiz K Hotta**[†], **Jorge A Achar**[‡]

Resumo

Neste trabalho foi proposto um Modelo de Regressão de Poisson Bivariado com covariáveis. O algoritmo EM foi utilizado para a estimação dos parâmetros pelo método de máxima verossimilhança. O modelo foi aplicado a dados bivariados de contagens do número de internações hospitalares decorrentes de doenças Hipertensivas e doenças Cerebrovasculares do sistema único de saúde da cidade de São Paulo.

Palavras Chave: Distribuição Poisson Bivariada,Correlação,Internações

1 Introdução

A análise de dados pareados de que exibam correlações podem ser encontradas nas mais diversas situações, e a distribuição Poisson Bivariada tem se destacado nos casos em que esses dados pareados são de contagem. No esporte Karlis e Ntzoufras (2003) e (2005) verificaram o número de gols marcados entre times competidores. Na área da saúde Kocherlakota e Kocherlakota (2001) investigaram o número de ataques epilépticos em pacientes antes e depois da aplicação de uma determinada droga e Karlis e Ntzoufras (2005) avaliaram o número de consultas ao médico ou especialista e o número total de medicamentos prescritos aos pacientes. Na indústria Ho e Singer (2001) observaram a contagem de dois tipos de defeitos na fabricação de fibras têxteis.

Kawamura (1972) foi um dos primeiros autores a desenvolver a distribuição Bivariada. O autor obteve a distribuição Poisson Bivariada através da soma de variáveis aleatórias Bernoulli Bivariadas e do Teorema de Poisson. Em outro trabalho Kawamura (1976) utilizou o lema da redução trivariada, com a distribuição Poisson Bivariada, para provar que as variáveis aleatórias univariadas independentes seguem distribuição Poisson. Esse lema ajudou a simplificar o processo de estimação e também é utilizado para o cálculo dos erros padrões do bootstrap e será enunciado formalmente no capítulo seguinte.

Karlis e Ntzoufras (2003) observaram que o alto custo computacional para a estimação dos parâmetros impediu uma larga utilização da distribuição nas décadas de 70 e 80. A partir da decada de 90, com o desenvolvimento de recursos computacionais, surgiram os primeiros modelos de regressão Poisson Bivariado com Jung e Winkelmann (1993), Kocherlakota e Kocherlakota (2001), Ho e Singer (2001), Karlis e Ntzoufras (2003) e (2005) dentre outros.

Neste trabalho, foram analisados dados de contagem diária de hospitalizações decorrentes de doenças Hipertensivas e doenças Cerebrovasculares, sendo investigada a relação das internações com as variáveis metereológicas pressão atmosférica, umidade relativa do ar e da temperatura, máximas e mínimas, da cidade de São Paulo, Brasil, no período de 01/01/2002 a 31/12/2005.

[*]Mestrando do Programa de Pós-Graduação em Estatística-UNICAMP. E-mail: jonasbodini@gmail.com

[†]Departamento de Estatística, Instituto de Matemática, Estatística e Computação Científica - IMECC, Universidade Estadual de Campinas - UNICAMP, Caixa Postal 6065, CEP: 13083-859, Campinas, SP, Brasil. E-mail: hotta@ime.unicamp.br

[‡]Departamento de Medicina Social, Faculdade de Medicina de Ribeirão Preto - FMRP, Universidade de São Paulo - USP, CEP: 14048-900, Ribeirão Preto, SP, Brasil. achcar@fmrp.usp.br

A divisão deste estudo foi feita da seguinte maneira: Na seção 2 está definida a distribuição Poisson bivariada, na seção 3 encontra-se o método de estimação utilizado e uma aplicação aos dados em questão encontra-se na seção 4; por fim na seção 5 são apresentadas as conclusões do trabalho.

2 A Distribuição de Probabilidade

Sejam $X_k, k = 1, 2, 3$ variáveis aleatórias independentes com distribuição Poisson com parâmetros $\lambda_k > 0$. Então as variáveis aleatórias $X = X_1 + X_3$ e $Y = X_2 + X_3$ seguem conjuntamente uma distribuição Poisson Bivariada $BP(\lambda_1, \lambda_2, \lambda_3)$ com função densidade de probabilidade dada por

$$P_{X,Y}(x,y) = exp(-(\lambda_1+\lambda_2+\lambda_3))\frac{\lambda_1^x}{x!}\frac{\lambda_2^y}{y!}\sum_{i=1}^{min(x,y)}\binom{x}{k}\binom{y}{k}k![\frac{\lambda_3}{\lambda_1\lambda_2}]^k \quad (1)$$

onde $\lambda_k > 0$ e $X, Y = 0, 1, ...$

Marginalmente cada variável aleatória segue uma distribuição Poisson com $EX = \lambda_1+\lambda_3$ e $EY = \lambda_2+\lambda_3$. E a Covariância é dada por $Cov(X, Y) = \lambda_3$. Ou seja, λ_3 é uma medida de dependência entre as duas variáveis aleatórias. Se $\lambda_3 = 0$ então a distribuição bivariada acima é reduzido ao caso do produto de duas distribuições Poisson independentes, que em Karlis e Ntzoufras (2003) denominam distribuição Poisson Dupla, denominação que também será utilizada neste trabalho para esse caso.

No lema apresentado abaixo Kawamura (1976) demonstra a volta da técnica de redução trivariada para a distribuição Poisson Bivariada.

Lema 1 *Se um vetor aleatório (X, Y) segue uma distribuição Poisson bivariada dada em (1) então ele pode ser determinado de forma única por*

$$X = X_1 + X_3$$
$$Y = X_2 + X_3$$

onde X_1, X_2 e X_3 são variáveis aleatórias Poisson univariadas com parâmetros λ_1, λ_2 e λ_3 respectivamente.

3 Estimação dos Parâmetros

A estimação dos parâmetros de interesse pelo método de máxima verossimilhança levaria a uma fórmula muito complicada que envolveria produtos de somatórios. Isso levou alguns autores a utilizar métodos alternativos de estimação. Jung e Winkelman (1993) e Kocherlakota e Kocherlakota (2001) utilizaram o método de Newton-Raphson, Ho e Singer (2001) utilizaram o método de mínimos quadrados Generalizados e Karlis e Ntzoufras (2003) e (2005) utilizaram o algoritmo EM. Karlis e Ntzoufras (2005) apontam como vantagem do algoritmo EM o fato de ele contornar os problemas de convergência devido a valores iniciais que ocorrem com o método de Newton-Raphson.

Para a i-ésima observação, o modelo de regressão bivariado toma a forma geral:

$$(X_i, Y_i) \sim BP(\lambda_{1i}, \lambda_{2i}, \lambda_{3i})$$
$$log(\lambda_{ki}) = w_{ki}\beta_k$$

para $i = 1, ..., n$ e $k = 1, 2, 3$. w_{ki} denota o vetor de variáveis explanatórias usadas para a i-ésima observação usada no modelo, λ_{ki} e β_k denota o correspondente vetor de coeficientes da regressão. Ou seja, não é obrigatório

e nem sempre ocorre de as variáveis aleatórias X e Y terem as mesmas variáveis explicativas, consequentemente nesses casos terão vetores de parâmetros de tamanhos diferentes.

A idéia básica da construção do algoritmo EM consiste em fazer uso da técnica da redução trivariada da distribuição Poisson Bivariada. Para cada observação são introduzidas as variáveis latentes X_{1i},X_{2i} e X_{3i} para as quais é assumido que elas seguem distribuição Poisson com parâmetros λ_1,λ_2 e λ_3 respectivamente. É assumido também que $X_i = X_{1i} + X_{3i}$ e $Y_i = X_{2i} + X_{3i}$.

O algoritmo EM estima as variáveis não observadas X_{1i},X_{2i} e X_{3i} via esperança condicional no passo E e então maximiza a verossimilhança no passo M. Ou seja, no passo E obtemos a esperança a posteriori de X_{1i},X_{2i} e X_{3i} dado as observações e os valores correntes dos parâmetros e então maximizamos a verossimilhança completa ajustando três modelos Poisson. Em seguida são estimados os parâmetros $\beta_k, k = 1, 2, 3$. Em Karlis e Ntzoufras (2005) é apresentado o algoritmo EM utilizado.

Na estimação dos erros padrões Jung e Winkelman (1993) utilizaram a matriz de informação de Fisher, Karlis e Ntzoufras (2003) e (2005) fizeram uso do método bootstrap e apontam como vantagem do método a velocidade de convergência.

A Estimação dos parâmetros nesse trabalho foi feita pelo algoritmo EM descrito em Karlis e Ntzoufras (2005). As análises foram realizadas pela função biv.pois() do programa R 2.8.0 que pode ser encontrado em `http://www.sciviews.org/_rgui/`, e o programa bootstrap para o cálculo dos erros padrões são baseados no programa de Karlis e Ntzoufras (2005).

4 Aplicação

Para a analise dos dados de internação, utilizou-se o modelo de regressão de Poisson Bivariado descrito nas seções anteriores considerando como variáveis respostas o número total de internações diárias por doenças Hipertensivas e doenças Cerebrovasculares, definidas conforme a Classificação Estatística Internacional de Doenças - CID-10 que pode ser obtida no endereço `http://www.datasus.gov.br/cid10/v2008/cid10.htm`. As variáveis explicativas utilizadas são pressão máxima, pressão mínima, umidade máxima, umidade mínima, temperatura máxima, temperatura mínima, estação do ano e fim de semana.

As contagens diárias de hospitalizações referem-se aos pacientes do Sistema Único de Saúde (SUS) da cidade de São Paulo e foram fornecidas pela prefeitura do município. Os dados das variáveis climáticas foram fornecidos pelo instituto de Astronomia e Geofísica da Universidade de São Paulo (IAG-USP) e são referentes à estação de medição do local. As variáveis explicativas estações do ano e fim de semana foram consideradas no modelo como variáveis *dummies*. No caso da variável estação do ano, a estação verão foi considerada como referência, e para a variável fim de semana foi considerado 1 para os dias sábado e domingo.

Na tabela 1 é apresentado o ajuste de três modelos Bivariados Poisson, O modelo 1 corresponde ao ajuste da distribuição Dupla Poisson $\lambda_3 = 0$, no modelo 2 não foram utilizadas variáveis explicativas para o parâmetro λ_3 (covariância constante) e finalmente no modelo 3 também foram utilizadas as cováriaveis para estimar o parâmetro λ_3. As estimativas dos erros padrões dos modelos foram calculados utilizando 200 replicações de bootstrap, conforme apresentado por Karlis e Ntzoufras (2005) e os efeitos das covariáveis foram avaliados utilizando testes t assintóticos.

Quando são comparados os modelos 1 e 2, objetiva-se saber se o parâmetro da covariância é estatisticamente significante no modelo. Podemos observar que as estimativas dos parâmetros e a significância destes divergiram um pouco devido à presença do parâmetro de correlação λ_3, cujo valor é estatisticamente significante.

No caso dos modelos 2 e 3 objetivou-se saber se as covariáveis têm influencia no efeito da covariância dos parâmetros. Verificou-se que, com exceção da umidade mínima, todas as covariáveis têm efeito sobre o parâmetro da covariância. Ou seja, podemos observar diferentes estimativas de covariâncias sobre as médias de internações das doenças Hipertensivas e Cerebrovasculares .

Tabela 1: Estimação dos Parâmetros

	Modelo 1		Modelo 2		Modelo 3	
lambda1						
Intercepto	8,4557	2,5486	8,2696	3,5993	13,8235	3,6328
PRESS MAX	0,0248	0,0070	0,0215	0,0097	-0,0358	0,0094
PRESS MIN	-0,0312	0,0066	-0,028	0,0087	0,0253	0,0089
UMID MAX	0,0039	0,0021	0,0026	0,0033	-0,0079	0,0035
UMID MIN	-0,0001	0,0010	0,0006	0,0016	-0,0018	0,0015
TEMP MAX	-0,0098	0,0046	-0,0075	0,0071	-0,0034	0,0068
TEMP MIN	0,0068	0,0048	0,0044	0,0066	-0,0212	0,0071
outo	0,0359	0,0198	0,0378	0,0273	-0,0114	0,0308
inv	0,0883	0,0230	0,0837	0,0313	-0,273	0,0326
prim	-0,0034	0,0218	0,0117	0,0291	-0,0044	0,0316
fim de semana	-0,3807	0,0152	-0,5341	0,0341	-0,2931	0,0275
lambda2						
Intercepto	4,5587	1,9589	3,4241	2,4584	7,1637	2,6853
PRESS MAX	0,0120	0,0056	0,0072	0,0066	-0,0258	0,0076
PRESS MIN	-0,0137	0,0054	-0,0078	0,0062	0,0223	0,0069
UMID MAX	0,0036	0,0018	0,0028	0,0022	-0,0035	0,0024
UMID MIN	-0,0005	0,0009	-0,0002	0,0011	-0,0016	0,0013
TEMP MAX	-0,0099	0,0039	-0,0087	0,0049	-0,0064	0,0056
TEMP MIN	0,0094	0,0039	0,0086	0,0051	-0,0067	0,0053
outo	0,0260	0,0167	0,0248	0,0216	-0,0032	0,0219
inv	0,0674	0,0183	0,0599	0,0216	-0,1473	0,0244
prim	0,0563	0,0170	0,0778	0,0199	0,0612	0,0213
fim de semana	-0,3069	0,0133	-0,3695	0,0179	-0,247	0,0168
lambda3						
Intercepto			1,5456	0,0938	-28,3096	2,6035
PRESS MAX					0,2297	0,0077
PRESS MIN					-0,2073	0,0069
UMID MAX					0,0664	0,0024
UMID MIN					0,0016	0,0012
TEMP MAX					-0,0485	0,0053
TEMP MIN					0,1493	0,0056
outo					0,1805	0,0225
inv					1,4075	0,024
prim					-0,0259	0,0239
fim de semana					-0,8225	0,0176

5 Conclusão

Neste trabalho a distribuição Poisson Bivariada mostrou-se um modelo útil para avaliar dados de contagens com correlação, mostrando que as estimativas dos parâmetros das correlações foram significativas e que as médias marginais das variáveis são influenciadas pelos efeitos das covariáveis.

Algumas extensões desse trabalho,talvez a mais natural delas, seria a extensão para o caso multivariado. Nesse caso uma expansão direta da técnica da redução trivariada e consequentemente algumas modificações sobre o algoritmo EM do caso bivariado seriam necessárias para essa abordagem.

Outra forma de extensão é considerar uma abordagem bayesiana para o problema. Ao considerar distribuições a priori para os parâmetros da distribuição Poisson Bivariada. Nesse caso as distribuições poderiam ser independentes ou uma priori conjunta para os parâmetros. Essa abordagem possibilita que o parâmetro da correlação assuma valores negativos.

6 Referências Bibliográficas

Referências

[1] Ho, L. e Singer, J. (2001) *Generalized Least Squares Methods for Bivariate Poisson Regression* Communications in Statistics - Theory and Methods, **30**, 263-277

[2] Jung, R. e Winkelmann, R (1993) *Two aspects of Labor Mobility: A Bivariate Poisson Regression Approach* Empirical Economics,**18**, 543-556

[3] Karlis, D. e Ntzoufras, I. (2003) *Analysis of Sports Data by Using Bivariate Poisson Models*, Journal of Royal Statistical Society,**52**, 381-393

[4] Karlis, D. e Ntzoufras, I. (2005) *Bivariate Poisson and Diagonal Inflated Bivariate Poisson Regression Models in R* Journal of Statistical Software, Vol 14,

[5] Kawamura, K. (1972) *The structure of Bivariate Poisson Distribution* Kodai Math,

[6] Kawamura, K. (1976) *The structure of Trivariate Poisson Distribution* Kodai Math,

[7] Kocherlakota,S e Kocherlakota,K (2001) *Regression in the Bivariate Poisson Distribution* Communications in Statistics - Theory and Methods, **30**, 815-827