UMA ILUSTRAÇÃO DO USO DA REGRESSÃO DE POISSON

Jean Carlos Cardoso[1], Everton da Costa[2], João Pedro Serenini[3] e Terezinha Aparecida Guedes[4]

[1]Universidade Estadual de Maringá, Departamento de Estatística, Maringá, PR, Brasil. E-mail: jeancarlos.card@gmail.com

[2]Universidade Estadual de Maringá, Departamento de Estatística, Maringá, PR, Brasil. E-mail: everto.cost@gmail.com

[3]Universidade Estadual de Maringá, Departamento de Estatística, Maringá, PR, Brasil. E-mail: jaopedro9@gmail.com

[4]Universidade Estadual de Maringá, Departamento de Estatística, Maringá, PR, Brasil. E-mail: taguedes@uem.br

RESUMO

O modelo de regressão de Poisson é utilizado frequentemente para modelar dados de contagem. O principal objetivo deste trabalho é mostrar, a partir de um exemplo da literatura, como fazer um ajuste usando o modelo de regressão de Poisson. Exibiremos passo a passo como fazer as análises de multicolinearidade, resíduos e diagnóstico de influência. Foi realizada uma aplicação cujo objetivo era verificar se o número esperado de clientes de uma determinada área de uma cidade é influenciado por variáveis como número de domicílios, renda, idade, distância ao concorrente mais próximo e distância até a loja. De acordo com os resultados, observamos que, em média, quanto maior a distância ao concorrente maior o número de clientes e, quanto maior a renda, idade, e distância até a loja, menor o número de clientes.

Palavras chave: Regressão de Poisson; Modelos Lineares Generalizados; Estimação via Máxima Verossimilhança; Contagem.

1 INTRODUÇÃO

O modelo de regressão de Poisson é um tipo específico de modelo linear generalizado (MLG) que teve origem por volta de 1970, quando Wedderburn (1974) desenvolveu a teoria da quasi-verossimilhança. A variável resposta de uma regressão de Poisson deve seguir uma distribuição de Poisson e os dados devem possuir igual dispersão (a média da variável resposta deve ser igual à variância). Frequentemente a suposição de igualdade entre média e variância é violada, porém mesmo nesses casos é possível se aplicar o modelo de regressão de Poisson, realizando alguns ajustes. Neste trabalho será apresentado a expressão matemática do modelo de Poisson destacando-se a função de ligação que conecta o regressor ao parâmetro correspondete a média na distribuição de Poisson. Será averiguado a qualidade de ajuste de tal modelo por intermédio de uma análise de resíduos e diagnóstico e, posteriormente, será realizada a interpretação dos parâmetros inerentes ao mesmo.

2 METODOLOGIA

Consideremos os dados apresentados em Neter *et al* (1996) sobre o perfil dos clientes de uma determinada loja oriundos de 110 áreas de uma cidade. Neste exemplo, será feita uma análise de regressão de Poisson afim de avaliar a relação entre o número esperado de clientes nas 110 áreas. O conjunto de dados é composto pela variável resposta Y: 'número esperado de clientes em cada área' e as covariáveis X_i como seguem: número de domicílios (em mil), renda média anual (em mil USD), idade média dos domicílios (em anos), distância ao concorrente mais próximo (em milhas) e distância à loja (em milhas).

O interesse dessa análise é responder: Quais fatores influenciam no aumento ou decaimento do número esperado de clientes em cada área? Existe relação entre o número esperado de clientes em cada área e os fatores econômicos envolvidos na análise?

Para responder tais perguntas, será utlizada a regressão de Poisson. Supondo que $Y \sim Poisson(\lambda)$ cuja função de probabilidade é dada por

$$Pr(Y = y) = \frac{e^{\lambda}\lambda^{y}}{y!},\ y = 0, 1, 2, \ldots, \tag{1}$$

utilizaremos a função de ligação canônica

$$\ln(\lambda_i) = \eta_i = x_i\beta \iff \lambda_i = e^{x_i\beta}. \tag{2}$$

3 RESULTADOS E DISCUSSÕES

É apresentado na tabela abaixo um resumo das medidas do conjunto de dados corresponde a 110 observações sobre o perfil dos clientes de uma determinada loja oriundos de 110 áreas de uma determinada cidade

Tabela 1: Análise descritiva dos dados.

	nclientes	domic	renda	idade	dist1	dist2
N	110.00	110.00	110.00	110.00	110.00	110.00
mean	11.20	647.76	48836.78	27.43	3.07	6.83
Std.Dev.	6.64	263.03	18531.06	16.68	1.50	2.29
min	0.00	19.00	19673.00	1.00	0.34	0.87
Q1	7.00	472.25	35160.25	13.00	1.93	5.59
median	10.00	647.00	44564.50	27.00	2.93	7.28
Q3	14.00	825.25	58369.00	41.75	4.28	8.67
max	32.00	1289.00	120065.00	58.00	6.61	9.90

De acordo com a tabela 1, podemos observar que o número esperado de clientes em cada área é de 11.20 com um desvio-padrão de 6.64. Além disso, cerca de 50% das áreas observadas têm um número esperado de clientes que varia entre 7 e 14. A área com o maior número esperado de clientes, 32, é a área 19, que tem 877 domicílios, uma renda anual de 51707, a idade média dos domicílios é de 27 anos, a distância até o concorrente mais próximo é de 5.19 milhas e a distância à loja é de 3.66 milhas (1 milha $\approx$ 1.609 km). As áreas 7, 37 e 45 não tiveram clientes.

Seja Y_i o número de clientes da i-ésima área que foram a loja em um determinado período. Suponhamos então que $Y_i \sim Poisson(\mu_i)$ com parte sistemática dada por:

$$log(\mu_i) = \beta_0 + \beta_1 domic + \beta_2 renda + \beta_3 idade + \beta_4 dist1 + \beta_5 dist2. \tag{3}$$

O ajuste deste modelo foi feito usando o comando *glm* do software R e os respectivos resultados estão descritos na tabela abaixo:

Tabela 2: Estimativas dos parâmetros do modelo log-linear de Poisson.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	2.9424	0.2072	14.20	0.0001
domic	0.0006	0.0001	4.26	0.0001
renda	-1.2e-05	2.1e-06	-5.53	0.0001
idade	-0.0037	0.0018	-2.09	0.0365
dist1	0.1684	0.0258	6.53	0.0001
dist2	-0.1288	0.0162	-7.95	0.0001

Observando o teste z de significância dos coeficientes, apresentado na tabela 2, temos que todas as estimativas são altamente significativas. O desvio do modelo foi de $D(y, \widehat{\mu}) = 114.99$ com 104 graus de liberdade que equivale a um nível descritivo $P = 0.35$ indicando um ajuste adequado.

Apartir dos resultados indicados na tabela 2 podemos concluir que o número esperado de clientes que frequentam a loja: cresça com o aumento do número de domicílios na área, cresça com a distância ao concorrente mais próximo, diminua com o aumento da renda médias, diminua com o aumento da idade média dos domicílios, diminua com o aumento da distância da área à loja.

Na tabela abaixo está descrito a análise de variância (ANOVA). Notamos ao observar esta tabela, que por meio do teste F, as variáveis domicílio (domic), distância até o primeiro concorrente (dit1) e distância até à loja (dist2), são significativas na regressão, visto que o p-valor de ambas é menor que 0.05, enquanto as variáveis renda e idade tem um p-valor superior a 0.05. Pelo teste F do ajuste global temos que o p-valor = 0.001, o que implica na não rejeição da hipótese nula, isto é, o modelo de regressão proposto é adequado.

Tabela 3: Anova.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
domic	1	476.53	476.53	46.17	0.0000
renda	1	14.20	14.20	1.38	0.2435
idade	1	3.63	3.63	0.35	0.5545
dist1	1	2239.14	2239.14	216.95	0.0000
dist2	1	998.72	998.72	96.77	0.0000
Residuals	104	1073.38	10.32		

O coeficiente de determinação $R^2 = 0.7766$, isto é, aproximadamente 77.66% da variabilidade total observada para a variável resposta número de clientes (nclientes) é explicada pela regressão. Há uma outra medida para avaliar a variabilidade total observada dos dados, essa medida é a R^2_{adj}. Para esse conjunto de dados temos que o $R^2_{adj} = 0.7659$ e sua interpretação se dá como a anterior, 76.59% da variabilidade dos dados é explicada pela regressão.

Ao calcularmos os valores de inflação de variância podemos perceber que não há presença de multicolinearidade, visto que todos os valores são menores que 10.

Outro teste usado para tentar captar multicolinearidade é feito por meio da análise dos autovalores da matriz rxx. Considerando os autovalores dessa matriz, temos que os números de condições da matriz rxx são dados por $k = \frac{\lambda_{max}}{\lambda_{min}}$. Como todos os valores de k são menores do que 100, não há presença de multicolinearidade entre as variáveis regressoras, porém, o determinante da matriz rxx é 0.3073184, mostrando evidência fraca de linearidade entre as regressoras.

Consideremos agora o algoritmo de pesquisa que permite simplificar um modelo de regressão, sem precisar analisar todos os possíveis submodelos. Usamos o algoritmo de exclusão sequencial (backward), para verificar os AICs dos modelos, adicionando ou removendo variáveis. Como já era esperado, devido ao teste z para significância dos coeficientes e o teste F global, o modelo com menor AIC, selecionado pelo comando 'step' do software R, foi o modelo completo.

De acordo com o teste de $Shapiro - Wilk$ para a normalidade dos resíduos obtivemos um $p - valor = 0.1184$, aproximadamente 12%. O que leva a conclusão a um nível de significância de 5% a não rejeição da hipótese nula, isto é, os resíduos provém de uma distribuição normal satisfazendo o primeiro pressuposto do modelo em questão.

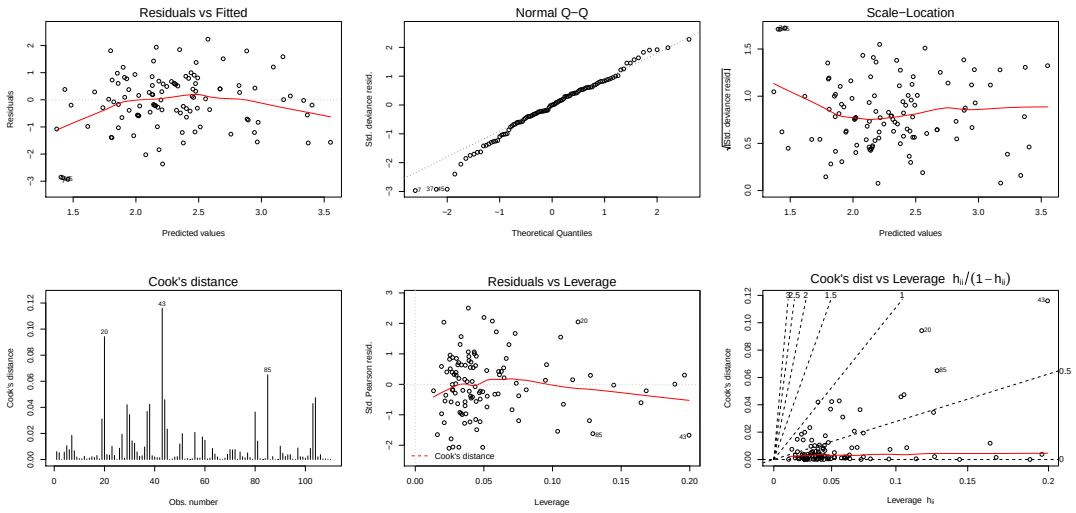

Figura 1: Gráficos da análise de Resíduos.

Realizando o diagnóstico de pontos de alavancagem de Hoaglin e Welsh (1978) concluímos que o rank da matriz H é 5 e $h_{ii} = 0.1$ o que indica que as observações $15, 20, 37, 47, 89$ e 94 são possíveis pontos de alavanca. Agora, ao analisarmos a influência dos coeficientes da regressão proposto por Belsley, Kuh e Welsch (1980), temos que $DFBeta = 0.1906925$, isto é, o ponto de corte usado para comparar os $DFBeta_{j,i}$ é de 0.20. Ainda, analisando a influência dos valores ajustados, temos que $DFFit = 0.4264014$, isto é, o ponto de corte usado para comparar os $DFFit_i$ é de 0.43. Por fim, analisando a influência na precisão da estimação de Belsley et al.(1980), temos que o limite superior e inferior do $CovRatio_i$ são 1.136364 e 0.8636364 o que nos diz que se o i-ésimo ponto satisfazer $CovRatio_i < 1.14$ ou $CovRatio_i > 0.87$ então ele deve ser um ponto influente.

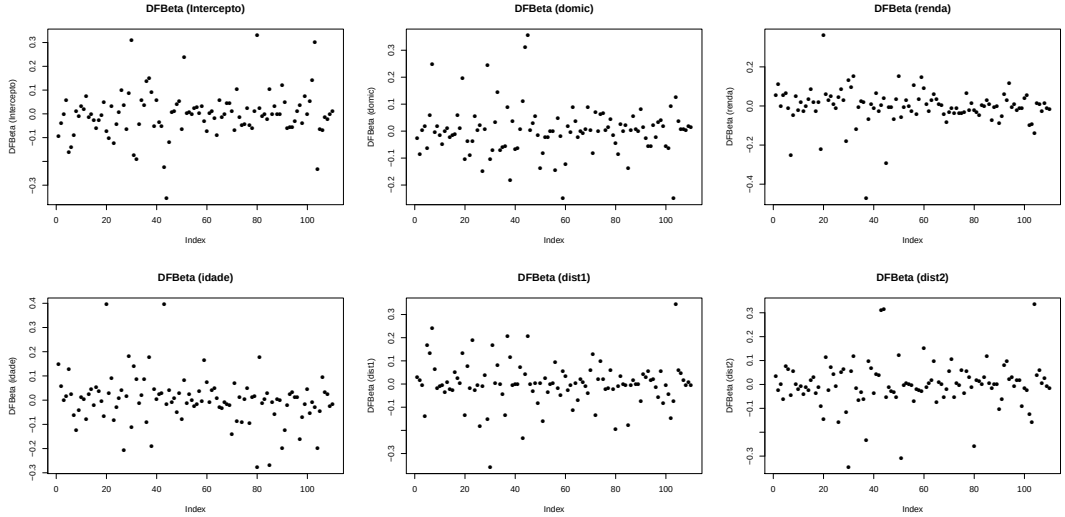

Figura 2: Gráficos dos BFBetas, DFFit, CovRatio e matriz hat.

4 CONCLUSÃO

Ao observamos nossos resultados podemos tirar algumas conclusões. Primeiramente as observações destacadas como influentes não mudam a inferência, pois não são significativas e não há indício que a função de ligação seja inapropriada. Por fim, não há indício de afastamentos importantes da suposição de distribuição de Poisson para o número de clientes que frequêntam a loja. Além disso, notamos que todos os fatores são significativos para a análise de regressão de Poisson. O fator que aumenta de forma significativa o número esperado de clientes em cada área é o 'dist1' (distância até o concorrente mais próximo). Os fatores que fazem com que haja um decaimento no valor esperado de clientes são: renda, idade e distância até a loja, sendo essa última a mais significativa. Existe uma forte relação entre as distâncias em geral e o número esperado de clientes em cada área.

Referências

[1] CHATTERJEE, Samprit. HADI, Ali S. **Regression Analysis by Exemple.** Wiley Series in Probability and Statistics. Ed 4. John Wiley & Sons, Inc. Hoboken: New Jersey, 1938.

[2] CZADO, Claudia. **Lectures 6: Poisson regression.** Technische Universitat Munchen: Gottingen, 2014.

[3] PAULA, Gilberto A. **Modelos de Regressão com apoio computacional.** Instituto de Matemática e Estatística, Universidade Estadual de São Paulo: São Paulo.

[4] RODRÍGUEZ, G. **Chapter 4: Poisson Models for Count data.** Revised September, 2007.

[5] TADANO, Yara de Souza. UGAYA, Cassia Maria Lie. FRANCO, Admilson Teixeira. **Método de Regressão de Poisson: Metodologia para avaliação do impacto da população atmosférica na saúde populacional.** Ambiente e Sociade, v. 12, n.2, p.241-255. Campinas, 2009.