Ajuste do Modelo Binomial Negativo a um Conjunto de Dados Relacionado à Administração de Empresas

Larissa Bueno Fernandes[1]

[1]Programa de Pós-Graduação em Bioestatística/PBE - Universidade Estadual de Maringá/UEM

RESUMO

Este trabalho apresenta a caracterização do modelo Binomial Negativo por meio dos modelos lineares generalizados, ajustando o mesmo a um conjunto de dados referente ao número de diretores compartilhados com outras companhias de grandes empresas canadenses, em 1970, avaliando quais variáveis são significativas para descrever a variável resposta de interesse. Foi verificado que o modelo Binomial Negativo proposto, com ligação *log*, apesar de apresentar certa tendência nos resíduos, foi relativamente adequado para a descrição do número e diretores e executivos compartilhados com outras companhias, sendo que as variáveis nação de controle e ativo das coorporações foram significativas para o modelo.
Palavras chave: Modelos lineares generalizados, Binomial Negativa.

1 INTRODUÇÃO

A análise de regressão, termo introduzido por Francis Galton no século XIX [3], caracteriza-se como uma técnica amplamente utilizada em várias áreas para investigar a relação de dependência de uma variável resposta com uma ou mais variáveis preditoras. A regressão busca descobrir quais preditores são importantes e estimar o impacto da alteração do valor de uma variável preditora sobre o valor da variável resposta [7].

Durante muitos anos os modelos normais lineares foram utilizados na tentativa de descrever a maioria dos fenômenos aleatórios [6], até mesmo para os quais a suposição de normalidade da variável resposta não era razoável, sendo que a transformação dos dados era uma alternativa comum para alcançar a normalidade. Entretanto, a transformação, pela mudança de escala, altera a relação entre a variável resposta e as variáveis preditoras, comprometendo, muitas vezes, a interpretação dos parâmetros do modelo [2].

Buscando unificar os procedimentos de inferência para outras opções de distribuição da variável resposta e outras funções para ligar os parâmetros das distribuições a um preditor linear, [4] introduziram os modelos lineares generalizados (MLG).

Sejam $Y_1, ..., Y_n$ variáveis aleatórias independentes, cada uma com função densidade - ou função de probabilidades - pertencente a família exponencial, como abaixo

$$f(y_i; \theta_i, \phi) = \exp(\phi(y_i\theta_i - b(\theta_i)) + c(y_i, \phi)). \tag{1}$$

De acordo com [6], os MLG's são definidos pela Equação (1) e pela parte sistemática, definida a seguir:

$$g(\mu_i) = \eta_i, \tag{2}$$

em que $\eta_i = x_i^T\beta$ é o preditor linear, $\beta = (\beta_0, \beta_1, ..., \beta_p)^T$, $p < n$, é o vetor de parâmetros desconhecidos a serem estimados, $x_i = (x_{i1}, ..., x_{ip})^T$ são os valores de variáveis explicativas e $g(\Delta)$ é a função de ligação, monótona e diferenciável.

Em modelos cuja variável resposta representa o resultado de contagem (quantitativa discreta) variando no intervalo $[0, \infty)$, a distribuição mais comumente utilizada é a de Poisson, candidata ao ajuste dos dados.

Entretanto, um problema comumente observado na análise de dados reais é a sobredispersão da variável resposta, que não é atendida pelo modelo de Poisson. Uma possível solução é a distribuição binomial negativa, que apresenta a vantagem da estimação de um parâmetro de dispersão.

Neste contexto, o objetivo deste trabalho é ajustar o modelo binomial negativo, por meio dos modelos lineares generalizados, a um conjunto de dados referente ao número de diretores compartilhados com outras companhias de grandes empresas canadenses, em 1970, avaliando quais variáveis são significativas para descrever a variável resposta de interesse.

2 METODOLOGIA

2.1 Dados

O conjunto de dados Ornstein apresenta o número de diretores e executivos compartilhados com outras companhias [5]. As observações são das 248 maiores empresas canadenses com informações publicamente disponíveis em meados da década de 1970. As variáveis utilizadas são apresentadas a seguir:

- Ativos: ativos da corporação (em milhões de dolares) - variável quantitatica contínua;
- Setor: setor de operação - variável qualitativa nominal;
- Nação: nação de controle - variável qualitativa nominal;
- *Interlocks*: Número de diretores e executivos compartilhados com outras companhias - variável quantitativa discreta.

2.2 Modelo Binomial Negativo

Seja X uma variável aleatória que representa o número de tentativas para se obter k sucessos, em n ensaios de Bernoulli, cada uma com probabilidade p de sucesso. Neste caso, a variável $X \sim NB(p, k)$, isto é, segue uma distribuição binomial negativa, com função de probabilidade dada por:

$$P(X = x) = \begin{pmatrix} x-1 \\ k-1 \end{pmatrix} p^k (1-p)^{x-k}, \tag{3}$$

em que $x = k, k+1, \ldots$.

A média de X é dada por $\mu = k/p$, sendo que a função de probabilidada apresentada na Equação (3) pode ser reparametrizada em função de μ. Assim, o modelo binomial negativo, ajustado aos dados referentes ao número de diretores e executivos compartilhados com outras companhias, é descrito a seguir.

Componente aleatório: $Y_i \sim NB(k, \mu_i)$, com $i = 1, 2, ..., 248$;

Componente sitemática: $\eta_i = \beta_0 + \beta_1 Setor_i + \beta_2 Nacao_i + \beta_3 Ativos_i$

Função de ligação: $\eta_i = log(\mu_i)$

Para identificar quais variáveis foram significativas para descrever o número de diretores compartilhados realizadas, foi utilizada a análise do Tipo 3. Esta análise testa a significância de cada variável explicativa, sob o pressuposto de que todas as outras variáveis inseridas na equação do modelo estão presentes.

Para avaliar qualidade do ajuste do MLG, também foi utilizado o quociente entre a estatística *Scaled Deviance* e seus respectivos graus de liberdade. A *Scaled Deviance* é dada por:

$$D^* = 2(l_{max} - l(\theta(\hat{\beta}))) \tag{4}$$

em que l_{max} é o valor máximo da função de verossimilhança e $l(\theta(\hat{\beta}))$ é a verossimilhança considerando os estimadores de máxima verossimilhança (EMV) dos parâmetros da regressão. Sob certas condições de regularidade, D^* segue uma distribuição limite do qui-quadrado, com graus de liberdade iguais ao número de observações menos o número de parâmetros estimados. Assim, o quociente entre D^* e seus respectivos graus de liberdade devem ser próximos a 1, como esperado para uma variável aleatória com distribuição qui-quadrado.

2.3 Análise de resíduos

De acordo com [6], a análise dos resíduos é constituída por um conjunto de técnicas para avaliar se a distribuição em estudo é apropriada para descrever o comportamento da variável resposta e ainda identificar a presença de possíveis pontos extremos no conjunto de dados.

Um resíduo R_i descreve a distância entre o valor observado y_i e o seu respectivo valor ajustado $\widehat{\mu}_i$, dado por $R_i = h_i(y_i, \widehat{\mu}_i)$, com h_i uma função conhecida e de fácil interpretação [1]. Assim, existem vários possíveis tipos de resíduos que podem ser obtidos, de acordo com a escolha (adequada) de h_i.

Seis tipos de resíduos serão utilizados neste trabalho para analisar a presença de tendência. Sabe-se que os resíduos devem ser independentes dos valores ajustados ou de uma combinação linear dos mesmos. Assim, a dispersão entre os resíduos e o preditor linear não deve apresentar nenhum tipo de tendência se as suposições do modelo forem verificadas.

Ainda, o envelope de probabilidade foi utilizado como a ferramenta para a análise da Normalidade dos resíduos. O nível de confiança do envelope é de 95%, e o tipo de resíduo utilizado foi o componente de desvio padronizado.

3 RESULTADOS E DISCUSSÕES

Tabela 1: Estatística da razão de verossimilhança para análise Tipo 3 do modelo Binomial Negativo ajustado.

Fonte	Graus de liberdade	χ^2	Valor p
Setor	9	11,71	0,2301
Nação	3	35,08	$< 0,0001$
Ativos	1	66,88	$< 0,0001$

De acordo com a Tabela 1, verifica-se que tanto a nação de controle quanto o ativo das coorporações foram sigificativos para explicar o comportamento do número de diretores e executivos compartilhados com outras companhias, ao nível de 5% de significância.

O quociente entre a estatística *Scaled Deviance* e seus respectivos graus de liberdade foi de $296,52/234 = 1,27 \approx 1$, como esperado para uma variável aleatória com distribuição qui quadrado.

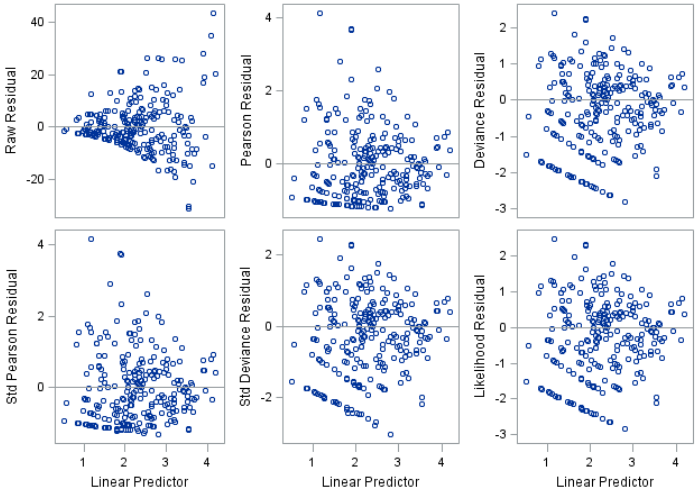

Figura 1: Gráficos dos resíduos contra os valores ajustados para o preditor linear.

Analisando a Figura 1, observamos uma certa tendência no comportamento dos resíduos brutos (*raw residuals*) de acordo com a variação de $\widehat{\eta}$, sendo que, a medida que o valor do preditor linear aumenta, a variabilidade do resíduo também aumenta. Já os resíduos *Deviance* aparentam distribuir-se aleatoriamente de acordo com a variação de $\widehat{\eta}$.

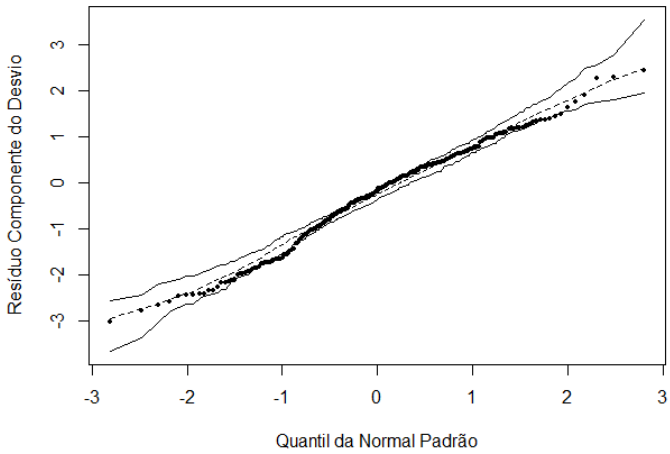

Figura 2: Envelope simulado dos resíduos componentes do desvio padronizados para o modelo de Binomial Negativo.

Observa-se que nenhum ponto (resíduo) ficou fora das bandas de confiança do envelope simulado para um ajuste da distribuição Binomial Negativa representado pela Figura 2, idicando que o modelo se ajustou bem aos dados e pode ser adequado para descrevê-los.

4 CONCLUSÃO

Os modelos lineares generalizados caracterizam-se como uma importante ferramenta para a modelagem de dados de diversas naturezas, como dados de contagem. Para o conjunto de dados considerados, Ornstein, verificou-se que o modelo Binomial Negativo proposto, com ligação *log*, apesar de apresentar certa tendência nos resíduos, foi relativamente adequado para a descrição do número e diretores e executivos compartilhados com outras companhias, sendo que todos as variáveis nação de controle e ativo das coorporações foram significativas para o modelo.

Referências

[1] COX, D.R.; SNEL, E.J., A general definition of residuals, Journal of the Royal Statistical Society B, v. 30, p. 248-275, 1968.

[2] FERRARI, S.; CRIBARI-NETO, F. Beta regression for modelling rates and proportions. Journal of Applied Statistics, Taylor & Francis, v. 31, n. 7, p. 799–815, 2004.

[3] GALTON, F. Regression towards mediocrity in hereditary stature. The Journal of the Anthropological Institute of Great Britain and Ireland, JSTOR, v. 15, p. 246–263, 1886.

[4] NELDER, J. A.; BAKER, R. J. Generalized linear models. Wiley Online Library, 1972.

[5] ORNSTEIN, Michael D. The boards and executives of the largest Canadian corporations: Size, composition, and interlocks. Canadian Journal of Sociology/Cahiers canadiens de sociologie, p. 411-437, 1976.

[6] PAULA, G. A. Modelos de regressão: com apoio computacional. IME-USP São Paulo, 2004.

[7] WEISBERG, S. Applied linear regression. John Wiley & Sons, 2005.