Publicación Electrónica **ISSN 2718- 6636**

UNIVERSIDAD NACIONAL DE ROSARIO
FACULTAD DE CIENCIAS ECONÓMICAS Y ESTADÍSTICA
SECRETARIA DE CIENCIA Y TECNOLOGIA E INSTITUTOS DE INVESTIGACIONES

Resumen Ampliado

Jornadas Anuales
"Investigaciones en la Facultad"
Ciencias Económicas y Estadística

Harvey, Guillermina Beatriz
Boggio, Gabriela Susana
Instituto de investigaciones Teóricas y aplicadas de la Escuela de Estadística

MODELOS POISSON-TWEEDIE PARA DATOS DE CONTEO CON EXCESO DE CEROS. SU COMPARACIÓN CON EL MODELO BINOMIAL NEGATIVO[1]

Resumen

En muchos estudios que involucran el análisis de datos de conteo es común encontrar una gran cantidad de ceros. La sobredispersión que ello provoca ha sido tenida en cuenta en diferentes alternativas de modelización siendo el modelo Binomial Negativo la más utilizada. Recientemente Bonat *et al*. presentaron una nueva clase de modelos lineales generalizados basados en los modelos Poisson-Tweedie desarrollados por Jørgensen y Kokonendji para analizar este tipo de datos. En este trabajo se estudió el comportamiento de las estimaciones de los modelos Poisson-Tweedie y Binomial Negativo mediante un estudio por simulación. Se encontraron estimaciones de los coeficientes de regresión con sesgos muy pequeños en ambos casos y errores cuadráticos medios levemente menores para los modelos Poisson-Tweedie.

Palabras claves: datos de conteo; exceso de ceros; modelos Poisson-Tweedie

Abstract

Statistical analysis involving count data usually have to face the presence of a large number of zeros. This produces overdispersion which has been taken into account in different modeling strategies, among which the Binomial Negative model is the most frequently used. Recently Bonat *et al*. presented a new class of generalized linear models based on Jørgensen and Kokonendji's Poisson-Tweedie models. In this work we studied and compared the behavior of the estimators from the Binomial Negative model and the Poisson-Tweedie model through simulations. Regression coefficients presented considerably small biases in both cases and slightly lower mean square errors in the later.

Keywords: count data; zero-inflated data; Poisson-Tweedie models

Introducción

En estudios que involucran datos de conteo es común encontrar un número excesivo de ceros. Ello produce una sobredispersión en los datos de forma tal que el modelo de regresión paradigmático para datos de conteo, el modelo Poisson, no resulta apropiado. Es habitual en estos casos recurrir al modelo Binomial Negativo, encuadrado también en los modelos lineales gene-

[1]Trabajo elaborado en el marco del Proyecto ECO 215 titulado "Enfoques estadísticos alternativos para el estudio de la ocurrencia de eventos según tiempos de exposición", dirigido por Gabriela Boggio.

ralizados (MLG). En 2017 Bonat *et al*. proponen una nueva clase de MLG basados en los mode los Poisson-Tweedie presentados por Jørgensen y Kokonendji (2016). Esta familia de modelos está dada por una especificación jerárquica: $Y/Z \sim \text{Poisson}(Z)$ y a su vez $Z \sim \text{Tw}_p(\mu, \phi)$, con media μ, parámetro de potencia $p > 1$ y ϕ, parámetro de dispersión. El parámetro p juega un rol importante ya que permite abarcar un abanico de posibilidades para captar el exceso de ceros considerando diferentes valores en el intervalo (1,2]. El caso especial en que p es igual a 2 corresponde a la distribución gamma y, como es sabido, la mezcla Poisson-gamma conduce al conocido modelo Binomial Negativo.

En este sentido resulta de interés comparar las estimaciones del modelo Poisson-Tweedie más adecuado ante una amplia gama de datos sobredispersos debido al exceso de ceros con las obtenidas bajo el modelo Binomial Negativo, de manera de poder evaluar las ventajas de contar con esta nueva familia de modelos.

Metodología

La comparación del ajuste de ambos modelos se realiza mediante un estudio por simulación. Se generan datos de conteo con exceso de ceros siguiendo el algoritmo presentado por Bonat *et al*. (2017), el cual procede a extraer en forma aleatoria valores de conteo Y correspondientes a una distribución Poisson-Tweedie, con valores medios que satisfagan un determinado modelo loglineal y con parámetro ϕ fijado de manera que el índice de dispersión (DI), $Var(Y_i)/E(Y_i) = 1 + \phi {\mu_i}^{p-1}$, tome los valores 2, 5, 10 y 20. Los valores de DI elegidos representan escenarios con dispersión baja, moderada, alta y muy alta respectivamente. Luego, los escenarios considerados quedan definidos por las combinaciones de estos cuatro valores del DI combinados con dos valores del parámetro p: 1,1 y 1,6.

El estudio por simulación para cada escenario consiste en generar 1000 muestras de tamaño n = 100 y en cada una de ellas ajustar el modelo Binomial Negativo y el modelo Poisson-Tweedie. Los parámetros del modelo Binomial Negativo se estiman por máxima verosimilitud mientras que los correspondientes al modelo Tweedie, a través de un enfoque de estimación desarrollado por Bonat *et al*. (2017) similar al enfoque de Ecuaciones de Estimación Generalizadas de los MLG marginales e implementado en el entorno computacional R. Se calcula el sesgo relativo promedio y el error cuadrático medio de los coeficientes de regresión del modelo.

Resultados

A partir del estudio por simulación realizado no se encontraron diferencias entre las densidades correspondientes a las estimaciones de los coeficientes derivadas del ajuste de ambos modelos. En los distintos escenarios, las densidades están centradas en los valores teóricos con una variabilidad que va en aumento a medida que se incrementa el DI. Asimismo, se observaron sesgos relativos muy pequeños, del orden del 1% del valor de los parámetros y errores cuadráticos medios algo menores para el caso del modelo Poisson-Tweedie y en especial en los escenarios con p=1,6.

Se evidenciaron problemas en la estimación del parámetro de dispersión en ambos modelos, siendo muy notorios para el modelo Binomial Negativo en el caso de alta sobredispersión debida a un marcado exceso de ceros (DI=20 y p=1,1).

A modo de síntesis, si bien los resultados hallados a partir de ambos modelos son en general muy similares, el modelo Poisson-Tweedie se comporta mejor cuando tanto la dispersión como

el exceso de ceros son muy altos.

En futuros trabajos se prevé generar datos sobredispersos a partir de un algoritmo más general. De este modo se podría ampliar la comparación de las estimaciones de los coeficientes de regresión a modelos para el tratamiento de esta clase de datos que incluyen un componente adicional para describir el exceso de ceros: los denominados modelos cero inflados y los modelos “hurdle” (Lambert, 1992; Ridout *et al*., 1998).

REFERENCIAS BIBLIOGRÁFICAS

Bonat, W.H., Jørgensen, B., Kokonendji, C.C., Hinde, J., y Demetrio, C.G.B. (2017) Extended Poisson–Tweedie: Properties and regression models for count data. Statistical Modelling, 18, 1–26.

Jørgensen, B. y Kokonendji, C.C. (2016) Discrete dispersion models and their Tweedieasymptotics. Advances in Statistical Analysis, 100, 43–78.

Lambert, D. (1992). Zero-Inflated Poisson Regression, with an Application to Defects in Manufacturing. *Technometrics*, *34* (1), 1–14.

R Core Team (2016) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Ridout, M.S., Demetrio, C.G.B., y Hinde, J.P. (1998) Models for count data with many zeros. *Proceedings of the XIXth International Biometrics Conference*, Cape Town, South Africa.