2.2. TALLER A2

TÉCNICAS DE REGRESIÓN PARA DATOS DE RECUENTO

MSc: Euclides Díaz MSc: Andrés Jaramillo
Facultad de Ingeniería Industrial
Universidad Tecnológica de Pereira, Pereira Risaralda, Colombia
e-mail: edarcos52@hotmail.com Andresjaramillo32@gmail.com

XIV Coloquio Regional de Matemáticas y IV Simposio de Estadística
Universidad de Nariño, Pasto Colombia
9 al 11 de mayo de 2018

Resumen

En este trabajo se presenta un estudio sobre la evaluación de Modelos de Regresión: Poisson, Binomial Negativo, Poisson Inflado con Ceros y Binomial Negativo Inflado con Ceros. Las variables explicativas consideradas para realizar el ajuste de los modelos son: sexo, rango de edad, región del hecho y temporada. La variable observada Y es el número de denuncias de violencia sexual que ocurren en una determinada población de un vector $N_{m\times 1}$, donde $N_{m\times 1}$ es la variable de control o variable de exposición, de esta forma lo que se modela en este caso no es el conteo, sino la tasa que se expresa como y_i/N_i. Se logró establecer que las variables que mejor explican la variable respuesta son: rango de edad, región del hecho y sexo de la víctima. Se calcularon las tasas de incidencia IRR y se mostró que los las mujeres son aproximadamente 10 veces más vulnerables que los hombres a estos sucesos. De las cinco regiones en las que se categorizó la población de estudio, la región con mayor riesgo es la central, además con respecto a la edad, el grupo que se encuentra entre 10 y 19 años son los de mayor riesgo. En cuanto a la temporada, la tasa de delitos sexuales fue levemente decreciendo conforme pasa el tiempo. Haciendo uso de las pruebas de bondad de ajuste y tomando en cuenta los criterios de selección AIC y BIC, se logra seleccionar el modelo de regresión binomial negativo (MRBN) como el mejor modelo que se acerca a la representación de los datos.

Palabras claves

Modelo lineal generalizado, Modelo inflado con ceros, Datos de recuento, Función de enlace, Variable *offset*.

1. Objetivo y presentación

El propósito del taller es presentar la teoría de los modelos de regresión Poisson, Binomial Negativo, Poisson Inflado con Ceros y Binomial Negativo Inflado con Ceros, para utilizarlos en el ajuste de datos tipo recuento. Los datos se relacionan con el número de denuncias de violencia sexual presentadas en el departamento de Nariño en una determinada temporada.

Los recuentos se definen como el número de sucesos o eventos que ocurren en una misma unidad de observación durante un intervalo temporal o espacial definido (ver [2]). Para nuestro estudio la unidad de observación es el tamaño de una población específica de $N_{m\times 1}$, siendo $N_{m\times 1}$ un vector con tamaños de individuos. Las variables de recuento se caracterizan por su naturaleza discreta y no negativa. Es decir, si Y es una variable aleatoria de recuento entonces los valores que toma son $0, 1, 2, \ldots$.

Generalmente las variables de recuento acumulan una gran cantidad de ceros en las observaciones, por tal motivo es importante recurrir en estos casos a modelos de regresión que consideren este fenómeno. De igual manera, en sucesos reales muchas veces se presenta sobredispersión en los datos, es decir, cuando la varianza supera el valor esperado en la variable respuesta. Este acontecimiento mas general que el primero y que muchas veces conlleva a exceso de ceros, se puede modelar con distribuciones de probabilidad que consideran esta coyuntura. Uno de los modelos de regresión que consideran este asunto es el Modelo de Regresión Binomial Negativo, en el cual la variable respuesta Y sigue una distribución binomial negativa (BN).

Mientras que en los modelos lineales (ML) se produce una relación de identidad entre los valores ajustados y el predictor lineal, $\mu = \eta$, en los MLG la linealidad se establece en la escala del predictor lineal pero no en la escala de los valores ajustados. No se da, por tanto, la identidad entre valores ajustados y valores predichos, sino que entre ellos media una función que los relaciona, *la función de enlace* (ver [1, 2, 6, 8–11]):

$$g(\lambda) = \ln(\lambda) = \eta = X\beta\,,$$

donde η es el predictor lineal, X es una matriz con p variables explicativas y β es un vector columna con p coeficientes de regresión, los cuales se asocian con su respectiva variable independiente. El valor λ se encuentra en la escala de la variable respuesta.

Los modelos de regresión de Poisson y Binomial Negativo, a diferencia del Modelos Inflados con Ceros, pertenecen a la familia de los modelos lineales generalizados (ver [6, 7]).

2. Contenido

Los modelos de regresión Poisson y Binomial Negativo tanto estándar como inflados con ceros, no solamente se utilizan para modelar conteos, sino también tasas, es decir, cuando existe una unidad de exposición $N = N_{m\times 1}$, denominada variable *Offset* y definida como:

$$Offset = \ln N\,.$$

Esta variable permite establecer la población a riesgo en cada observación. La tasa se define como el recuento y_i dividido por alguna unidad exposición N_i, y_i/N_i.

Los modelos lineales generalizados (ver [6, 7]) presentan tres componentes:

Componente sitemático; Componente aleatorio y *Función de enlace.*

La función de enlace transforma el valor esperado a la escala del predictor lineal.

2.1. Modelo de regresión de Poisson

El Modelo de Regresión de Poisson (MRP) es adecuado cuando se cumple la propiedad $Var(Y) = E(Y) = \lambda$. Su formulación para la tasa es:

$$\ln(\lambda_i) = \ln(N_i) + \beta_1 + \sum_{j=2}^{p} \beta_j x_{ij} \tag{1}$$

La estimación de los parámetros β_j se realiza por el método de máxima verosimilitud.

2.2. Modelo de regresión binomial negativo

El MRBN se utiliza cuando $Var(Y) > E(Y)$. La formulación e interpretación del modelo es similar al MRP. La media y la varianza son, respectivamente:

$$E(Y) = \lambda \qquad\qquad Var(Y) = \lambda + \alpha\lambda^2,$$

donde α es el parámetro de dispersión. La estimación de los parámetros β_j se realiza por el método de máxima verosimilitud o el método de Newton Raphson.

2.3. Modelos inflados con ceros

Cuando la variable respuesta es un conteo los datos observados se deben modelar estadísticamente con distribuciones discretas como la Poisson y la Binomial Negativa. Sin embargo, no es raro que el número de ceros observados en la variable respuesta exceda a la frecuencia que se espera observar bajo la distribución que se ajusta o que la variable respuesta presente exceso de ceros y sobredispersión.

Por tal motivo, se han desarrollado modelos con inflado de ceros que consideran diversos escenarios, es decir, tanto exceso de ceros como exceso de ceros y sobredispersión en la variable respuesta. Entre estos modelos, el Modelo de Regresión de Poisson Inflado con Ceros, conocido como modelo ZIP y propuesto por Lamber en [10], el cual presenta un mejor ajuste a los datos que el MRP cuando la variable explicada presenta un número elevado de ceros. Sin embargo, el modelo ZIP no es apropiado cuando la parte no nula de la distribución esta sobredispersa con respecto a la distribución de Poisson. Entonces cuando la variable respuesta en un modelado presenta exceso de ceros y sobredispersión, el modelo mas apropiado que recomiendan algunos autores es el Modelo de Regresión Binomial Negativo Inflado con Ceros, denominado en esta literatura como modelo ZINB (ver [9, 11, 13]).

2.4. Pruebas de sobredispersión y bondad de ajuste

Las pruebas que permiten evaluar sobredispersión de los datos en los MLG se pueden consultar en [1,6,7]. De igual manera, las pruebas de bondad de ajuste para los modelos: MRP, MRBN, ZIP y ZINB se pueden observar en [1,6,7,12]. Finalmente los criterios de selección AIC y BIC se pueden consultar en cualquiera de las fuentes anteriormente mencionadas.

Referencias

[1] A. Agresti, Categorical data analysis, New York, Wiley, 2002.

[2] J.K. Lindsey, Modelling frequency and count data, Oxford science publications, Clarendon Press, 1995.

[3] R.B. Christopher and M.L. Thomas, Analysis of categorical data with R, Texts in Statistical Science, 2015.

[4] A.C. Cameron and P.K. Trivedi, Microeconometrics using stata, Stata Press, 2009.

[5] A.C. Cameron and P.K. Trivedi, Econometric models based on count data. comparisons and applications of some estimators and tests, Journal of Applied Econometrics 1 (1986), no. 1, Pages 29-53, Cited By :862.

[6] P. McCullagh and J.A. Nelder, Generalized linear models, Chapman and Hall, 1989.

[7] J.A. Nelder and R.W.M.Wedderburn, Generalized linear models, Journal of the Royal Statistical Society 135 (1972), no. 3, Pages 370-384.

[8] F. Felix and S.P. Karan, Zero-inflated generalized poisson regression model with an application to domestic violence data, Journal of Data Science 4 (2006), no. 1, Pages 117-130.

[9] J.M. Hilbe, Negative binomial regression, Cambridge University Press, 2011.

[10] D. Lambert, Zero-inflated poisson regression, with an application to defects in manufacturing, Technometrics 34 (1992), no. 1, Pages 1-14, cited By 1414.

[11] S.M. Mwalili, E. Lesaffre, and D. Declerck, The zero-inflated negative binomial regression model with correction for misclassification: An example in caries research, Statistical methods in medical research 17 (2008), no. 2, Pages 123-139, Cited By 44.

[12] Q.H. Vuong, Likelihood ratio tests for model selection and non-nested hypotheses, Econometrica 57 (1989), no. 2, Pages 307-333, cited By 2188.

[13] O.B. Yusuf, T. Bello, and O. Gureje, Zero inflated poisson and zero inflated negative binomial models with application to number of falls in the elderly, 1 (2017), no. 4, Pages 1-7.