REFINAMIENTO DE PRUEBAS DE HIPÓTESIS ASINTÓTICAS

Lozano Forero, Sébastien
cestadistico@libertadores.edu.co
Fundación Universitaria los Libertadores (Colombia)

RESUMEN

En esta conferencia serán presentadas las ideas fundamentales del refinamiento de pruebas de hipótesis basadas en poblaciones de tamaño pequeño y moderado, para los modelos en series de potencias no lineales generalizados, recientemente publicados en la literatura. Serán utilizadas las metodologías de corrección de Bartlett y corrección Bartlett-Bootstrap para la estadística de razón de verosimilitudes para pruebas de hipótesis sobre los parámetros que indexan dicho modelo de regresión. Se presentarán resultados de simulaciones de Monte Carlo que muestran que, efectivamente, la tasa de rechazo empírica de las estadísticas corregidas es más próxima a su respectivo nivel nominal que las estadísticas originales. Se pretende mostrar estas técnicas con el ánimo de dar a conocer a practicantes y docentes la existencia de alternativas para escenarios de tamaños muestra mediana y pequeño.

PALABRAS CLAVE

Bootstrap, Corrección de Bartlett, Monte Carlo, Pruebas de hipótesis

INTRODUCCIÓN

En Cordeiro, Andrade y De Castro (2009) fueron introducidos los Modelos en Series de Potencias no Lineales Generalizados (MSPNLGs) con el ánimo de unificar en una sola estructura conceptual varios de los principales modelos de regresión discreta, como por ejemplo Poisson generalizada, Binomial negativa generalizada, entre otros. Esta familia de distribuiciones discretas tiene una estructura bastante flexible para el modelamiento de datos discretos.

Para realizar pruebas de hipótesis en los MSPNLGs, pueden ser usadas las tradicionales estadísticas da razón de verosimilitudes (LR), escore (SR), Wald (W) y la recientemente explorada estadística gradiente (G), introducida en Terrell (2002). En este trabajo se dará un énfasis a la LR. Tal estadística, bajo la hipótesis nula tiene una distribución aproximada Chi-cuadrado, con error de orden hasta n^{-1}donde n es el tamaño de muestra. De esta forma, es importante obtener ajustes para la estadística LR, cuando el tamaño de muestra sea pequeño pues las pruebas asintóticas basadas en tamaños de muestra pequeños pueden conducir a resultados distorsionados. La idea es modificar dichas estadísticas por un factor de corrección, con el objetivo de producir nuevas estadísticas cuyo primer momento sea igual al de la distribución Chi-cuadrado de referencia, es decir, obtener estadísticas con distribución aproximada Chi-cuadrado con error de orden hasta n^{-2}.

Adicionalmente, fue considerada la técnica de remuestreo Bootstrap, presentada en Efron (1979) para estimar numéricamente dicho factor de corrección de Bartlett para la estadística LR, dicha idea fue inicialmente explorada en Rocke (1989).

MARCO DE REFERENCIA

Sean $Y_1, \ldots, Y_n$ variables aleatorias discretas independientes tal que Y_i sigue una familia de distribuciones con parámetros de media $\mu_i > 0$ y parámetro de dispersión $\phi > 0$, cuya función de probabilidad tiene la forma:

$$\pi(y; \mu_i, \phi) = \frac{a(y, \phi) g(\mu_i, \phi)^y}{f(\mu_i, \phi)}$$

En que $y \in A_p = \{p, p+1, \ldots\}$ y p es un entero positivo, $a(y, \phi)$ es positiva y las funciones $g(\mu_i, \phi)$ y $f(\mu_i, \phi)$ son positivas finitas y dos veces diferenciables. Se asume que el parámetro ϕ es conocido. Propiedades sobre esta familia de funciones pueden ser consultadas en Cordeiro et al. (2009).

El modelo de regresión está definido de forma que la media de Y_i está relacionada con el componente sistemático a través de una función de ligación de la forma $h(\mu_i) = \eta(x_i, \beta), i = 1, \ldots, n$ en que $h(\mu_i)$ es una función de ligación conocida y doblemente diferenciable. $\beta = (\beta_1, \beta_2, \ldots, \beta_p)^T$ es un vector de p $(p < n)$ parámetros desconocidos a ser estimados y $x_i = (x_{i1}, x_{i2}, \ldots, x_{k1})^T$ representa los valores de k variables explicativas y $\eta(x_i, \beta), i = 1, \ldots, n$ es una función posiblemente no lineal en el segundo argumento, continua y diferenciable respecto a las componentes de β.

DESARROLLO

El logaritmo de la función de verosimilitud para los MSPNLGs, dado un vector de observaciones $y = (y_1, y_2, \ldots, y_n)^T$ es definida como:

$$\ell(\beta; y) = \sum_{i=1} [\log\{a(y_i, \phi)\} + y_i \log\{g(\mu_i, \phi)\} - \log\{f(\mu_i, \phi)\}]$$

Los elementos del vector escore para el parámetro β se definen como las derivadas de la función de log-verosimilitud respecto a las componentes de β, queda dado por $U_\beta = X(Ty - Q)^T$ donde:

$$- X = \frac{\partial \eta}{\partial \beta}$$

- $T = diag\{t_1, t_2, \ldots, t_n\}$ es una matriz diagonal de dimensión $n \, x \, n$ y la i-ésima entrada está dada por $t_i = \frac{g'(\mu_i, \phi)}{g(\mu_i, \phi) h'(\mu_i)}$

- Q= $(q_1, q_2, \ldots, q_n)^T$ es un vector de dimensión $n \, x \, 1$ cuya i-ésima entrada está dada por $q_i = \frac{f'(\mu_i, \phi)}{f(\mu_i, \phi) h'(\mu_i)}$.

Adicionalmente, la matriz de información de Fischer para β con ϕ conocido está dada por

$$K = X^T W X \text{ donde } W = diag\{w_1, \dots, w_n\} \text{ y } w_i = \left(q_i' - \frac{f'(\mu_i,\phi)g(\mu_i,\phi)}{f(\mu_i,\phi)g'(\mu_i,\phi)} t_i'\right) \frac{1}{h\prime(\mu_i)}.$$

Así, siendo $y = (y_1, y_2, \dots, y_n)^T$ un vector de observaciones de tamaño n y cuya función de log-verosimilitud dado anteriormente depende del parámetro de vectores desconocidos $\beta = (\beta_1, \beta_2, \dots, \beta_p)^T$. Se asume que $\ell(\beta; y)$ sea regular respecto a los componentes de β hasta cuarto orden.

Considerando que β puede ser descompuesto como $\beta = (\beta_1^T, \beta_2^T)^T$ siendo $\beta_1 = (\beta_1, \beta_2, \dots, \beta_q)^T$ el vector de parámetros de interés y $\beta_2 = (\beta_{q+1}, \beta_{q+2}, \dots, \beta_p)^T$el vector de parámetros de perturbación. El interés es poder probar la hipótesis $H_0: \beta_1 = \beta_1^{(0)}$ vs $H_1: \beta_1 \neq \beta_1^{(0)}$ donde $\beta_1^{(0)}$ es un vector conocido de dimensión q $(q < p)$. La estadística de razón de verosimilitud se define como:

$$LR = 2\{\ell(\hat{\beta}; y) - \ell(\tilde{\beta}; y)\}$$

Donde $\hat{\beta}$ es el estimador de máxima verosimilitud sobre H_1 (estimador irrestricto) y $\tilde{\beta}$ es el estimador de máxima verosimilitud sobre H_0 (estimador restricto). La estadística LR tiene, bajo H_0, una distribución asintótica χ_q^2. Por tanto, la regla de rechazo consiste en rechazar H_0, a un nivel de significancia α si $LR > \chi_{(\alpha,q)}^2$ donde $\chi_{(\alpha,q)}^2$ es el percentil $1-\alpha$ de la distribución χ_q^2 y q es el número de restricciones.

Para garantizar la convergencia de la distribución de la estadística de prueba LR a una χ_q^2, es necesario un tamaño grande de muestra. En escenarios de muestra de tamaño pequeño, tal convergencia podría no ser satisfactoria. Buscando mejorar esta aproximación, Bartlett (1937) propuso multiplicar la estadística LR por un factor de corrección $\frac{1}{1+d}$, con el objetivo de obtener la estadística corregida LR^*. La obtención del factor de corrección $\frac{1}{1+d}$ se sigue del resultado $d = (\epsilon_p - \epsilon_{p-q})/q$ donde ϵ_p y ϵ_{p-q} son expresiones que envuelven cumulantes y derivadas de hasta cuarta orden de la función de log-verosimilitud. Formalmente, se define la estadística $LR^* = \frac{LR}{1+d}$.

Para el caso de los MSPNLGs la corrección de Bartlett para pruebas de hipótesis basados en la estadística LR, es decir LR*, la obtención del factor de corrección puede ser consultado en Lozano, Silva, Cysneiros y Cordeiro (2015). Se omite aquí su escritura por considerar irrelevante la presentación explícita del mismo en este medio.

Otra alternativa para la obtención del factor de corrección de Bartlett es estimarlo numéricamente basado en las ideas de Rocke (1989). Para tal estimación es necesario

realizar el procedimiento computacional de Bootstrap obtenido a través de los siguientes pasos:

1. Construya B pseudo-muestras aleatorias $(y_1, y_2, \dots, y_B)^T$ sobre H_0, basando en la muestra original y.
2. Para cada una de las muestras anteriormente construidas y_i, calcule el valor de la estadística LR. Sea LR_i al valor obtenido.
3. Defina la estadística $LR^*{}_{boot}$, LR corregida vía Bootstrap Bartlett como:

$$LR^*{}_{boot} = q\frac{LR}{\overline{LR}}$$

Donde $\overline{LR} = \frac{1}{B}\sum_{i=1}^{B} LR_i$.

Se espera que este procedimiento produzca buenos resultados con apenas B=200 réplicas, aunque es recomendable realizar el procedimiento con al menos B=300 réplicas.

En Bayer y Cribari (2013) fueron propuestas versiones corregidas de la estadística LR, vía Bartlett y vía Bootstrap Bartlett en modelos de regresión Beta, llegando a concluir que ambas pruebas tienen un desempeño similar. Por lo anterior podemos concluir que, al menos en esa clase de modelos, como las correcciones de Barlett (teórica) y Bootstrap (numérica) tuvieron desempeños semejantes; es recomendable preferir la corrección de Bootstrap, pues no demanda cantidades extensas de cuentas algebraicas como sí lo requiere la corrección de Bartlett.

RESULTADOS

Se presentan los resultados de la simulación de Monte Carlo en un escenario no lineal para las distribuciones GPO (Generalized Poisson) con ϕ=1 y BNG (Binomial negativa generalizada) con ϕ=1 y ν=3, para el modelo:

$$\eta_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 x_{4i} + \beta_5 x_{5i} + \beta_6 x_{6i} + \exp(\beta_7 x_{7i}).$$

Es de interés probar la hipótesis $H_0: (\beta_5, \beta_6) = (0,0)$ vs $H_1: (\beta_5, \beta_6) \neq (0,0)$ bajo las siguientes condiciones:

1. La variable respuesta fue generada asumiendo $\beta_5 = \beta_6 = 0$ y $\beta_i = 0.05$ para $i = 1,2,3,4,7,8$.
2. Las respectivas covariadas fueron tomadas como muestras aleatorias de las distribuciones:

LN(0,1), F(2,5), Cauchy, χ^2_3, Beta (2,3), N(0,2), exp(1) y N(0,1), respectivamente.

3. Fueron utilizados tamaños de muestra $n = 20, 30, 40$ y 50.
4. Fueron utilizados los niveles nominales $\alpha = 1\%, 5\%, 10\%$.
5. Fueron utilizadas $B = 300$ réplicas Bootstrap.
6. Fueron utilizadas 10000 réplicas de Monte Carlo.

Para evaluar las estadísticas involucradas se utiliza la proximidad de la tasa de rechazo empírica (basada en réplicas Monte Carlo) con su respectivo nivel de significancia en escenarios de distintos tamaños de muestra y distintos niveles de significancia.

		Modelo GPO			Modelo BNG		
	α	LR	LR^*	LR^*_{boot}	LR	LR^*	LR^*_{boot}
n=20	1%	1,8	1,3	1	2,2	1,1	1,1
	5%	7,3	5,1	4,9	8,2	5,3	5,3
	10%	12,9	9,6	10,2	14,7	10,1	9,9
n=30	1%	1,4	1,1	0,9	1,5	1,1	1
	5%	6,3	5	5	6,4	5,1	5
	10%	11,9	10,1	10,1	12,1	10	10,2
n=40	1%	1,4	1,3	0,9	1,4	1,1	1,1
	5%	6,1	5,3	5	6,1	5,1	5
	10%	11,9	10,4	10,1	11,7	10	9,9
n=50	1%	1	0,9	1	1,1	0,9	1,1
	5%	5,5	4,9	4,9	5,5	4,8	5,1
	10%	1,08	10	9,8	11	9,8	10

Tabla 1. Tasas de rechazo de la hipótesis planteada para las estadísticas consideradas con varios tamaños de muestra y niveles nominales
Feunte: Elaboración propia

En la Tabla 1, es posible apreciar los siguientes hechos.

- En general, la estadística original LR tiene un comportamiento notablemente liberal (tasas de rechazo empíricas por encima de los respectivos niveles nominales).
- Dicha tendencia liberal se reduce conforme el tamaño de muestra aumenta.
- Las estadísticas LR^*_{boot} y LR^* presentan tasas de rechazo empíricas más próximas a los respectivos niveles nominales que la estadística original LR.

Por tanto, es de resaltar que las estadísticas LR^*_{boot} y LR^*son, bajo las condiciones simuladas, son mejores alternativas para conducir pruebas de hipótesis en este tipo de modelos de regresión en escenarios de tamaño de muestra medianos y pequeños, que la estadística de prueba original LR.

APLICACIÓN

Finalmente, para dar a conocer la utilidad de las estadísticas propuestas, previamente definidas en un conjunto de datos reales, se estudia la cantidad de especies de peces en un lago (variable dependiente) y el logaritmo natural de la zona de tal lago, dado en km^2. Tales datos fueron estudiados inicialmente por Barbour y Brown (1974) y posteriormente por Rigby, Stansinopoulos y Akantziliotou (2008) y por Cordeiro et al. (2009). Ésta última investigación analiza la flexibilidad de los MSPNLGs en estos datos que emplean los siguientes predictores:

$$\eta_i = \beta_0 + \beta_1 \log(x_i)$$

y

$$\eta_i = \beta_0 + \beta_1 \log(x_i) + \beta_2 \{\log(x_i)\}^2$$

Con $i = 1, \ldots, 70$, donde $\eta_i = \log(\mu_i - m)$ y m denota el valor mínimo del soporte de la función correspondiente. Nótese que el primer modelo presentado es lineal mientras que el segundo es no lineal. Es de interés probar la hipótesis $H_0: \beta_2 = 0$ vs $H_1: \beta_2 \neq 0$, es decir, cuál de los dos modelos presentados se ajusta mejor los datos.

Para hacer esto, fueron tenidos en cuenta los modelos examinados por Cordeiro et al. (2009): Poisson, NB, GP, GNB y DB. Este último fue el modelo más adecuado para ajustar el número de las especies de peces, dado que tiene el menor valor de AIC (criterio de información de Akaike). Dichos criterios fueron de 610,9 y 614,1 para los predictores considerados. Los valores de las estadísticas de pruebas LR, LR^* y LR^*_{boot} para el modelo Delta Binomial son: 2,7447, 2,5624 y 2,2710, respectivamente. Observándose que para el nivel nominal del 10%, estas pruebas conducen a resultados divergentes y sólo el LR indica el rechazo de H_0.

CONCLUSIONES

Los modelos en Series de Potencias no Lineales Generalizados constituyen una alternativa atractiva para modelar fenómenos cuya variable tenga distribución discreta, toda vez que generalizan modelos de regresión tradicionales en esta área como el Poisson, Binomial Negativa Generalizada, Consul, entre otras.

Fueron presentadas versiones corregidas de la estadística de prueba LR, a saber: $LR^*(Bartlett)$ y $LR^*_{boot}(bootstrap\ Bartlett)$; que presentaron un comportamiento más estable y menos liberal que la estadística original en escenarios de tamaños de muestra medianos y pequeños. Las estadísticas LR^* y LR^*_{boot} convergen en distribución a la distribución de referencia χ^2_q más rápidamente que la estadística sin modificar.

Se espera que este trabajo sirva para ilustrar la necesidad imperante de investigación en este campo. El resultado principal será la maximización del uso de la información a partir de muestra de tamaño pequeño, toda vez que las estadísticas modificadas conducen a resultados más precisos que sus análogas sin modificar.

REFERENCIAS

Barbour, C. D. & Brown, J. (1974). Fish species diversity in lakes. *American Naturalist, 108*(962), 473–489.

Bartlett, M. S. (1937). Properties of sufficiency and statistical tests. Proceedings of the Royal Society of London. *Series A-Mathematical and Physical Sciences, 160*(901), 268–282.

Bayer, F.M. & Cribari, F. (2013). Bartlett corrections in beta regression models. *Journal of Statistical Planning and Inference, 143*(3), 531–547.

Cordeiro, G. M., Andrade, M. G. & De Castro, M. (2009). Power series generalized nonlinear models. *Computational Statistics & Data Analysis, 53*(4), 1155–1166.

Efron, B. (1979). Bootstrap methods: another look at the jackknife. *The annals of Statistics, 7*(1), 1–26

Lozano, S., Silva, P., Cysneiros, A. & Cordeiro, G. M. (2015). *Improved Likelihood Ratio Test in Power Series Generalized Nonlinear Models*. Submitted to Statistical papers.

Terrell, G. R. (2002). The gradient statistic. *Computing Science and Statistics, 34*, 206–215.

Rocke, D. (1989). Bootstrap Bartlett adjustment in seemingly unrelated regression. *Journal of the American Statistical Association, 84*(406), 598–601.

Rigby, R., Stansinopoulos, M. & Akantziliotou, C. (2008). A framework for modelling overdispersed count data, including the poisson-shifted generalized inverse Gaussian distribution. *Computational Statistics & Data Analysis, 53*(2), 381–393.