Hachuel, Leticia
Boggio, Gabriela
Wojdyla, Daniel
Cuesta, Cristina
Instituto de Investigaciones Teóricas y Aplicadas, Escuela de Estadística.
Consejo de Investigación de la Universidad Nacional de Rosario.

ESTUDIO DEL COMPORTAMIENTO DE TESTS TIPO SCORE Y WALD PARA DATOS BINARIOS CORRELACIONADOS

1. INTRODUCCIÓN

El estudio de datos binarios correlacionados es un área de interés creciente en distintos campos científicos. Un análisis frecuente con este tipo de datos consiste en describir la relación entre una respuesta dicotómica y covariables a través de la aplicación de métodos de regresión.

Entre los métodos desarrollados para el análisis de regresión de datos binarios correlacionados, se destaca el enfoque de la ecuación de estimación generalizada (GEE) sugerido por Liang y Zeger (1986) y Zeger, Liang y Albert (1988). El enfoque GEE, sólo requiere especificar correctamente la esperanza marginal de las respuestas a fin de obtener estimadores de los coeficientes de regresión consistentes y asintóticamente normales, junto con un estimador consistente de la matriz de covariancias. Aunque las propiedades asintóticas de los estimadores GEE son aceptadas, sus propiedades para tamaños de muestra pequeños no son muy conocidas.

Rotnitzky y Jewell (1990) derivaron estadísticas de tipo Wald y score generalizadas para evaluar hipótesis sobre los parámetros de regresión estimados por GEE, también con propiedades a nivel asintótico.

Desde otro enfoque, el ajuste de modelos a datos provenientes de encuestas necesita incorporar aspectos relativos al diseño muestral y considerar la posible falta de independencia entre las observaciones. En este contexto, Rao, Scott y Skinner (1998) presentan el test de cuasi-score, análogo al test de score, pero que incorpora información sobre el diseño muestral complejo.

Una forma de evaluar el comportamiento de estimadores y tests bajo condiciones que pongan en duda las propiedades asintóticas es a través de estudios por simulación.

Teniendo en cuenta que las conclusiones de los estudios por simulación se reafirman en la medida que resulten consistentes a través de diferentes modelos de generación de datos, y en el marco del proyecto "Estudio del comportamiento de estadísticas para medidas repetidas en muestras pequeñas bajo escenarios múltiples"*, en una primera instancia se compararon diferentes algoritmos de generación de datos binarios correlacionados. Además se realizó una evaluación del comportamiento de un grupo de estadísticas que considera la falta de independencia de las observaciones mediante estudios de Montecarlo, que proporcionó resultados consistentes a partir de datos generados por dos algoritmos diferentes (Hachuel, et al., 2002).

El presente trabajo tiene por objeto completar el estudio del comportamiento de estadísticas, utilizando datos generados por el algoritmo de Servy et al (1999), a través de la evaluación del control del error tipo I y II de las mismas.

2. METODOLOGÍA

2.1. Presentación de estadísticas

La generalización de las ecuaciones de estimación para el análisis de medidas repetidas tiene en cuenta la correlación intragupo en la estimación de los parámetros de regresión de un modelo para la esperanza marginal y se supone que la distribución marginal de la variable pertenece a la familia exponencial (Liang y Zeger,1986; Zeger, Liang y Albert, 1988). Dichas ecuaciones se conocen como Ecuaciones de Estimación Generalizadas (GEE).

A partir de los resultados sobre propiedades y distribución asintótica de los estimadores $\hat{\beta}$ obtenidos por GEE, Rotniztky y Jewell (1990) derivan generalizaciones a los clásicos tests chi-cuadrado para probar hipótesis sobre los parámetros de regresión. En particular, definen el test de score generalizado y el test de Wald generalizado.

Se supone para ello, disponer de una muestra de n conglomerados de tamaño variable k_m (m=1,..,n) para comprobar hipótesis del tipo Ho: $\beta_2 = \beta_{20}$, cuando se considera una partición del vector β de dimensión px1 de coeficientes de regresión $\beta' = (\beta_1', \beta_2')$ siendo β_1 un vector (p-q)x1 que contiene las (p-q) primeras componentes de β y β_2 un vector de dimensión qx1.

* Proyecto de Investigación de la Secretaría de Ciencia y Tecnología de la UNR (PID 2001/2003).

Sea el estimador obtenido por GEE, $\hat{\beta}_G$, la solución de:

$$S_G = \sum_{m=1}^{n}\left(\frac{\partial \mu_m}{\partial \beta}\right)' V_m^{-1}(Y_m - \mu_m) = 0 \qquad (2.1.1)$$

donde $V_m = A_m^{1/2} R_m(\alpha) A_m^{1/2}$ con $A_m = \text{diag}\{\text{var}(y_{mv})\}$ m=1,..,n ; v=1,..,k_m. Bajo condiciones débiles de regularidad, Liang y Zeger demuestran que $n^{1/2}\left(\hat{\beta}_G - \beta\right)$ tiene distribución asintótica normal con media cero y variancia asintótica dada por:

$$V_\beta = W_\beta \Omega W_\beta, \qquad (2.1.2)$$

donde:

$W_\beta = n\left(\sum_{m=1}^{n} D_m' V_m^{-1} D_m\right)^{-1}$, siendo $D_m = \frac{\partial \mu_m}{\partial \beta}$ y

$\Omega = \sum_{m=1}^{n} D_m' V_m^{-1} \text{cov}(Y_m) V_m^{-1} D_m$, siendo $\text{cov}(Y_m)$ la verdadera matriz de covariancia de Y_m, la cual se estima por $\left\{Y_m - \mu_m\left(\hat{\beta}_G\right)\right\}\left\{Y_m - \mu_m\left(\hat{\beta}_G\right)\right\}'$.

A partir de estos resultados Rotnitzky y Jewell definen el test de score generalizado de la siguiente manera:

$$X_{SG}^2 = n^{-1} \tilde{S}_{G_2}' \tilde{\Sigma}_2^{-1} \tilde{S}_{G_2}, \qquad (2.1.3)$$

siendo:

$$S_{G_2} = \sum_{m=1}^{n}\left(\frac{\partial \mu_m}{\partial \beta_2}\right)' V_m^{-1}(Y_m - \mu_m), \qquad (2.1.4)$$
$$\Sigma_2 = W_{\beta_2}^{-1} V_{\beta_2} W_{\beta_2}^{-1}$$

V_{β_2} es la sub-matriz principal de dimensión qxq de la matriz de covariancias del vector de estimadores de β, V_β, y W_{β_2} es la sub-matriz principal de dimensión qxq de W_β.

En (2.1.3), las expresiones están valorizadas en $\tilde{\beta} = (\tilde{\beta}_1, \beta_{20})$, donde $\tilde{\beta}_1$ es solución de:

$\tilde{S}_{G1} = S_{G1}(\tilde{\beta}) = 0$; es decir, $\tilde{S}_{G2} = S_{G2}(\tilde{\beta})$ y $\tilde{\Sigma}_2$ es el estimador de Σ_2 valorizado en $\tilde{\beta}$.

La estadística X_{SG}^2 tiene una distribución asintótica chi-cuadrado con q grados de libertad bajo condiciones débiles de regularidad y suponiendo una correcta especificación del modelo para la esperanza marginal.

Los autores definen además el test de Wald generalizado, que para la misma hipótesis resulta:

$$X^2_{WG} = (\hat{\beta}_{G_2} - \beta_{20})' \hat{V}_{\beta_2}{}^{-1} (\hat{\beta}_{G_2} - \beta_{20}), \qquad (2.1.5)$$

donde $\hat{\beta}_{G_2}$ es el sub-vector de dimensión qx1 del vector de estimadores $\hat{\beta}_G$; $\hat{V}_{\beta_2}$ es la submatriz qxq de $\hat{V}_\beta$, correspondiente a las covariancias de $\hat{\beta}_{G_2}$.

Esta estadística se distribuye bajo Ho con distribución chi-cuadrado con q grados de libertad.

Distintos autores están investigando realizar ajustes a la estadística de Wald para tener en cuenta el número de grupos o conglomerados con el fin de producir estadísticas con mejores propiedades para tamaños de muestras moderadas. Shah, Holt y Folsom (1977) presentaron la siguiente modificación de la estadística de Wald, X^2_{WG}, basada en una transformación semejante a la T^2 de Hotelling:

$$X^2_{WG_C} = \frac{n-c}{nc} X^2_{WG}, \qquad (2.1.6)$$

donde n es la cantidad de conglomerados y c son los grados de libertad del contraste. Esta estadística se distribuye según una Fde Snedecor con c y (n-c) grados de libertad.

Rao, Scott y Skinner (1998) parten de un enfoque cuasi-verosímil similar al de Liang y Zeger para derivar el test de cuasi-score. El mismo incluye elementos de la teoría de muestreo de población finita ya que considera que la muestra observada proviene de una población finita que a la vez se supone constituye una muestra aleatoria de una superpoblación infinita.

Entonces, para comprobar la hipótesis: Ho: $\beta_2 = \beta_{20}$, se plantean en primer término las ecuaciones de estimación:

$$S_C(\beta) = \sum_{l=1}^{N} u_l(\beta) = 0, \qquad (2.1.7)$$

donde la l-ésima componente de $u_l(\beta)$ es: $u_{lh} = (\partial \mu_l / \partial \beta_h)(y_l - \mu_l)/V_{ol}$ y N es el tamaño de la población finita.

Como $S_C(\beta)$ es un vector de totales poblacionales para β fijo, un estimador de ese vector resulta:

$\hat{S}_C(\beta) = \sum_{\gamma \in s} w_\gamma u_\gamma(\beta)$ donde w_γ son las ponderaciones muestrales.

(2.1.8)

Sea $\hat{S}_C = (\hat{S}'_{C1}, \hat{S}'_{C2})'$ una partición de $\hat{S}'_C$ compatible con la de $\beta' = (\beta_1', \beta_2')$. La solución de:

$$\hat{S}_{C1}(\tilde{\beta}) = 0 \qquad (2.1.9)$$

se denomina $\tilde{\beta}' = (\tilde{\beta}_1', \beta_{20}')$.

El test de cuasi-score, análogo al test de score, se basa en la estadística:

$$X^2_{CS} = \tilde{S}'_{C2} \tilde{V}^{-1}_{C2} \tilde{S}_{C2}, \qquad (2.1.10)$$

donde $\tilde{S}_{C2} = \hat{S}_{C2}(\tilde{\beta})$ y $\tilde{V}_{C2}$ es un estimador consistente de $\mathrm{Cov}(\tilde{S}_{C2})$.

Se demuestra que $\tilde{S}_{C2}$ tiene una distribución asintótica normal con media cero y matriz de covariancia $\mathrm{Cov}(\tilde{S}_{C2})$, por lo que la estadística X^2_{CS}, construida según (2.1.10), es asintóticamente una variable chi cuadrado con q grados de libertad, bajo Ho.

El cálculo de la estadística de cuasi-score X^2_{CS} requiere de la estimación de $\mathrm{Cov}(\tilde{S}_{C2})$ Un estimador posible es el estimador por linearización de Taylor que se simboliza con $\tilde{V}_{L2}$ (Rao, 1996) y es el utilizado en el presente trabajo.

2.2. Algoritmo de generación de datos

El modelo de simulación, presentado en Servy et al. (1998), fue diseñado originalmente para generar muestras de conglomerados cuyos elementos son pares de valores de dos variables categóricas, de forma tal que finalmente la muestra puede presentarse bajo la forma de una tabla de contingencia bivariada. Los conglomerados se generan por los k pasos de una cadena de Markov. Si se ignora una de las variables, dichos conglomerados se transforman en univariados y si esa variable es binaria, el algoritmo se puede utilizar para generar muestras de conglomerados univariados o de vectores de respuestas binarias correlacionadas. El modelo fija inicialmente el valor de la probabilidad de respuesta igual a 1, π, el tamaño k del conglomerado y especifica la matriz de transición M de la cadena de Markov. A partir de estos valores, se determina el vector de probabilidades iniciales

resolviendo un sistema de ecuaciones. Es posible determinar, una vez especificados dichos parámetros, las correlaciones entre las respuestas en pares de posiciones diferentes, ρ_{vj}.

2.3. Estudio de simulación

Se estudia el comportamiento de las estadísticas X^2_{CS}, X^2_{SG} , X^2_{WG} y $X^2_{WG_C}$ anteriormente presentadas para el caso particular de poner a prueba la hipótesis de nulidad del parámetro de regresión en un modelo logit con una única variable explicativa dicotómica aplicado a datos binarios correlacionados. El modelo logit:

$$\ln\frac{\pi}{1-\pi} = \beta_1 + \beta_2 X \quad ; X = 0,1 \qquad (3.7)$$

se ajusta a datos generados por el algoritmo de Servy et al. bajo diferentes escenarios. En esta oportunidad se eligen escenarios con esquemas de correlación semejante pero con probabilidad de respuesta igual a 1(π) baja, media y alta. Para ello, se elige una misma matriz de transición M para el modelo de generación de datos igual a $\begin{pmatrix} 0.9 & 0.1 \\ 0.1 & 0.9 \end{pmatrix}$, asociada a los valores de π que se presentan en la Tabla 1. Como consecuencia se obtienen las correlaciones de a pares también presentadas en esa tabla.

Tabla1: Escenarios paramétricos

Escenario	π	R=(ρ_{12};ρ_{13};ρ_{23})
1	0.3	(0.76, 0.59, 0.77)
2	0.5	(0.80, 0.64, 0.80)
3	0.8	(0.67, 0.49, 0.73)

Bajo estas condiciones, se simulan muestras simples aleatorias de n= 15, 30, 50, 70 y 100 conglomerados de tamaño fijo k=3 y se observan todos los individuos dentro de los mismos. En cada muestra se calculan las estadísticas ya definidas utilizando dos especificaciones diferentes para la matriz de correlación de trabajo $R(\alpha)$ en las estadísticas de score y Wald generalizadas. Ellas son Independencia y AR(1) por ser esta última una buena aproximación a la forma de dependencia que presentan los datos generados.

Para el estudio del control del error de tipo I del test Ho: $\beta_2 = 0$, las muestras se generan bajo el mismo escenario para X=0 y X=1 con asignación aproximadamente

balanceada en ambas categorías de X. Para cada muestra se calculan las estadísticas presentadas y se decide el rechazo o no de la Ho a un nivel del 5%. Se repite este procedimiento mil veces para cada tamaño de muestra y se calcula el porcentaje de rechazo real, o nivel de significación real.

Para el estudio del control del error de tipo II del test Ho: $\beta_2 = 0$, las muestras se generan bajo dos escenarios distintos según sea X=0 o X=1, con asignación aproximadamente balanceada en ambas categorías de X. En particular, si para X=0 se generan datos con los valores paramétricos correspondientes al escenario 1 y para X=1 los datos se generan a partir del escenario 2, el valor de β_2 resulta igual a 0.85. Un segundo valor de β_2, más alejado del postulado en la Ho, esto es: $\beta_2 = 2.23$ se obtiene generando para X=0 datos de acuerdo a los parámetros del escenario 1 y utilizando los parámetros del escenario 3 para X=1.

Para cada una de las alternativas descriptas y para cada muestra se calculan las estadísticas presentadas y se decide el rechazo o no de la Ho a un nivel del 5%. Se repite este procedimiento mil veces para cada tamaño de muestra y se calcula el porcentaje de rechazo real, el cual representa la potencia empírica de cada test para la alternativa esfecífica tenida en cuenta.

3. RESULTADOS

De acuerdo a los resultados presentados en las Tablas 2, 3 y 4, referidos al nivel de significación real, si bien coinciden en sus tendencias generales en los tres escenarios estudiados, presentan algunas diferencias atribuibles al valor estipulado para π = P(Y=1/ x). Así, es posible observar un comportamiento en general liberal de la estadística de cuasi-score para n menor que 50, pero con un acercamiento hacia el 5% especialmente en el escenario 2 (π=0.5) cuando el tamaño de la muestra aumenta.

En las estadísticas derivadas de la estimación de β_2 por la metodología GEE, no se observan diferencias entre los valores obtenidos para las dos especificaciones de la matriz de correlación de trabajo, Independencia y AR(1). Dentro de este grupo de estadísticas, se observan comportamientos conservadores en X^2_{SG} y $X^2_{WG_C}$ para n=15 particularmente en los escenarios 1 y 3 y algo más atenuado en X^2_{WG}. Sus niveles de significación empíricos tienden al nivel de significación nominal a medida que el tamaño de la muestra crece. En

particular en el escenario 2 estos cambios son menos apreciables y en general se observan niveles de significación reales más estables y más cercanos al 5%, aún para n chico.

Tabla 2: Nivel de significación real de los tests en el Escenario 1 según el tamaño de la muestra

N	Estadísticas						
	X^2_{CS}	$X^2_{SG(ind)}$	$X^2_{WG(ind)}$	$X^2_{WG_C(ind)}$	$X^2_{SG(AR1)}$	$X^2_{WG(AR1)}$	$X^2_{WG_C(AR1)}$
15	7.70	3.00	4.60	2.90	2.70	4.40	2.90
30	6.50	5.40	5.10	4.40	4.70	4.70	4.00
50	6.60	6.00	5.70	5.00	6.20	6.00	5.20
70	6.40	5.90	5.80	5.00	5.40	5.50	4.80
100	6.50	6.00	6.00	5.60	6.10	6.30	6.00

Tabla 3: Nivel de significación real de los tests en el Escenario 2 según el tamaño de la muestra

	Estadísticas						
N	X^2_{CS}	$X^2_{SG(ind)}$	$X^2_{WG(ind)}$	$X^2_{WG_C(ind)}$	$X^2_{SG(AR1)}$	$X^2_{WG(AR1)}$	$X^2_{WG_C(AR1)}$
15	9.70	5.70	6.40	4.20	5.70	6.40	4.30
30	6.30	5.00	5.00	3.70	5.40	5.10	4.50
50	5.80	5.00	4.50	3.00	4.70	4.50	3.60
70	4.90	4.50	4.20	4.00	4.20	4.30	3.90
100	5.10	5.10	4.80	4.70	4.60	4.60	4.60

Tabla 4: Nivel de significación real de los tests en el Escenario 3 según tamaño de la muestra

n	Estadísticas						
	X^2_{CS}	$X^2_{SG(ind)}$	$X^2_{WG(ind)}$	$X^2_{WG_C(ind)}$	$X^2_{SG(AR1)}$	$X^2_{WG(AR1)}$	$X^2_{WG_C(AR1)}$
15	7.10	1.60	4.40	2.40	1.10	3.50	1.90
30	6.30	4.70	5.20	4.10	4.20	5.80	4.20
50	6.50	6.20	5.90	5.40	5.50	5.80	5.50
70	5.30	5.10	4.90	4.60	4.80	4.60	4.30
100	6.20	6.00	5.80	5.60	6.10	5.80	5.70

En relación al control del error de tipo II, se consideran los comportamientos de las estadísticas ante las dos hipótesis alternativas consideradas, teniendo en cuenta sólo las situaciones donde los tests presentaron un aceptable control del error de tipo I en el escenario 1. Por lo tanto, para tamaños de muestra mayores que 30 y cuando β_2 =0.85, se observan potencias empíricas bajas, alrededor de 35%, para n=50 y llegan sólo al 63% para n=100. En cambio, para la alternativa β_2 =2.23 más alejada del valor β_2 =0 postulada en Ho, las potencias ya alcanzan el 85% para n=30 y el 98% para un tamaño de muestra mayor o igual a 50 (Tablas 5 y 6).

Para ambas alternativas no se aprecian diferencias en las potencias empíricas entre las estadísticas bajo consideración, pero se debe recordar el comportamiento algo liberal de la estadística de cuasi-score en su control del error de tipo I.

Tabla 5: Potencia empíricas medias de los tests para la alternativa β_2 =0.85 según tamaño de la muestra

N	Estadísticas						
	X^2_{CS}	$X^2_{SG(ind)}$	$X^2_{WG(ind)}$	$X^2_{WG_C(ind)}$	$X^2_{SG(AR1)}$	$X^2_{WG(AR1)}$	$X^2_{WG_C(AR1)}$
15	19.00	11.50	13.40	9.60	12.50	14.70	10.50
30	24.40	21.10	20.30	18.10	21.60	21.20	18.40
50	37.00	34.60	34.60	32.80	36.00	35.80	34.40
70	48.30	46.70	46.60	45.40	48.00	48.50	47.40
100	62.40	61.40	61.50	60.40	63.00	62.80	61.90

Tabla 6: Potencia empíricas medias de los tests para la alternativa β_2 = 2.23 según tamaño de la muestra

N	Estadísticas						
	X^2_{CS}	$X^2_{SG(ind)}$	$X^2_{WG(ind)}$	$X^2_{WG_C(ind)}$	$X^2_{SG(AR1)}$	$X^2_{WG(AR1)}$	$X^2_{WG_C(AR1)}$
15	65.00	52.00	54.60	46.40	53.30	57.50	49.20
30	90.40	88.30	88.80	86.60	89.20	89.80	87.70
50	98.60	98.50	98.60	98.10	98.60	98.60	98.60
70	99.9	99.90	99.90	98.90	99.90	99.90	99.90
100	99.9	99.90	99.90	99.90	99.90	99.90	99.90

4. DISCUSIÓN

Los resultados hallados permiten concluir acerca de una aparente influencia del valor de la probabilidad de respuesta igual a 1, π, en el comportamiento de las estadísticas comparadas. Es posible observar un mejor y más estable comportamiento cuando π es cercano a 0.5; sin embargo esta observación debería ser motivo de investigaciones futuras.

Respecto de las estadísticas estudiadas, se encuentra un comportamiento algo liberal para cualquier tamaño de muestra de la estadística de cuasi-score, resultado consistente con el presentado en Hachuel (2001), hecho que la hace menos potente respecto de las de score y Wald generalizadas. Por el contrario estas últimas resultan conservadoras para muestras muy pequeñas. No se detecta, además, la liberalidad esperada en la estadística generalizada de Wald por lo que parece innecesaria su corrección.

En síntesis, los resultados alertan acerca de los recaudos a considerar a la hora de extraer conclusiones válidas con muestras de tamaño pequeño o moderado en base a algunas de las estadísticas estudiadas.

Bibliografía

HACHUEL, L. “Estadísticas tipo score para modelos logit con diseños muestrales complejos”. Tesis para acceder al grado de Magister en Estadística Aplicada. Universidad Nacional de Córdoba. 2001.

HACHUEL, L.; BOGGIO, G ; WOJDYLA, D.; CUESTA, C.: “Evaluación del comportamiento de estadísticas tipo score para datos binarios correlacionados bajo escenarios múltiples”. Actas en CD:5º Congreso Latinoamericano De Sociedades de Estadística (CLATSE). Caseros, 2002.

LIANG, K. Y.; ZEGER, S. L. “Longitudinal data analysis using generalized linear models”. *Biometrika* 73, 13-22, 1986.

RAO, J. N. K "Developments in sample survey theory: an appraisal. *Canadian Journal Statistics* 25, 1-21, 1996.

RAO, J. N. K.; SCOTT, A. J.; SKINNER, C. J. "Quasi-score tests with survey data". *Statistica Sinica*, 8, 1059-1070, 1998.

ROTNITZKY, A.; JEWELL, N. “Hypothesis testing of regression in semiparametric generalized linear models for cluster correlated data”. *Biometrika* 77, 485-497, 1990.

SHAH, B.; HOLT, M.; FOLSOM, R. “Inference about regression models from sample survey data” *Bulletin of the International Statistical Institute*, 47, 43-57, 1977.

SERVY, E.; HACHUEL, L.; WOJDYLA, D. “Análisis de tablas de contingencia para muestras de diseño complejo”. *Cuadernos IITAE,* 4. Escuela de Estadística. UNR, 1998.

SERVY, E.; HACHUEL, L.; WOJDYLA, D. "A simulation study for analizing the perfomance of tests of independence under cluster sampling". *Bulletin of the International Statistical Institute.51st Session Istanbul*. Book 2: 411, 1997.

ZEGER, S. L.; LIANG, K. Y.; ALBERT, P. S. “Models for Longitudinal Data: A Generalized Estimating Equation Approach”. *Biometrics* 44, 1049-1060, 1988.