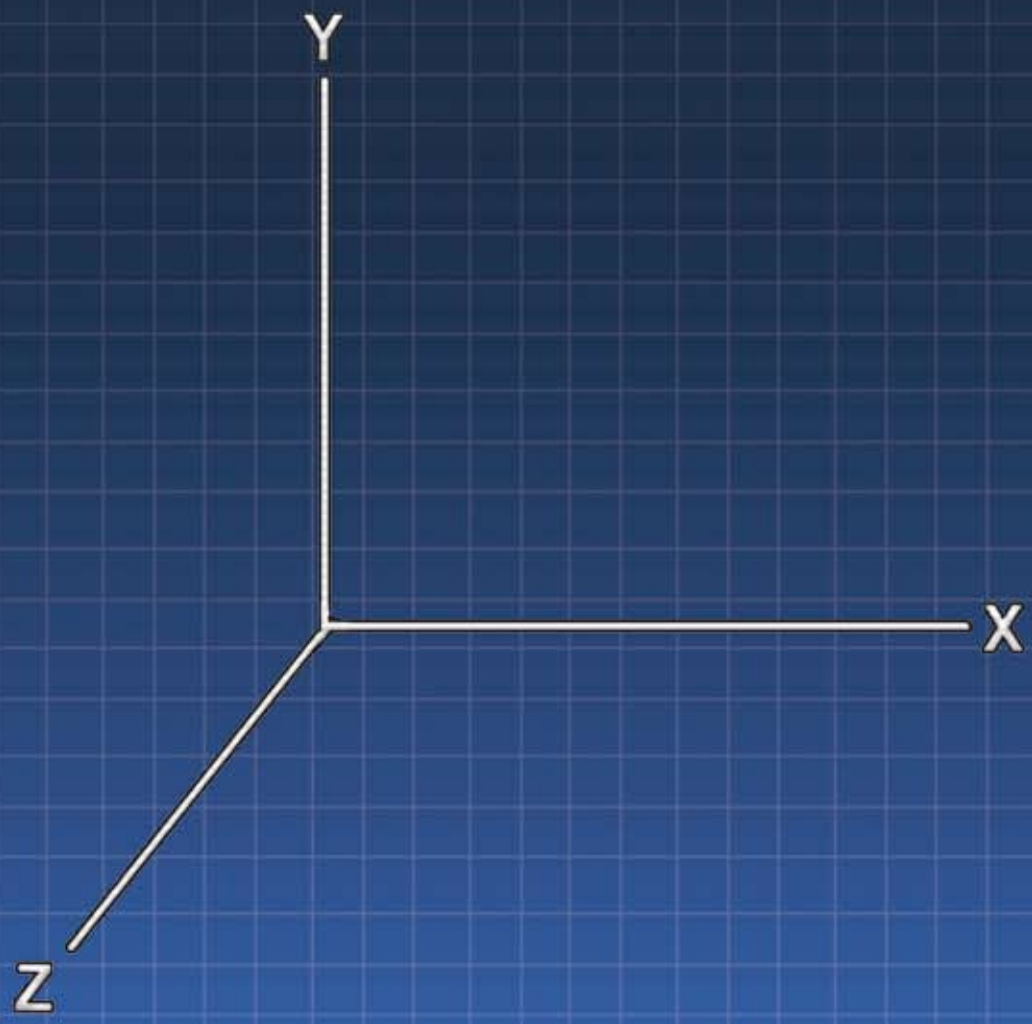
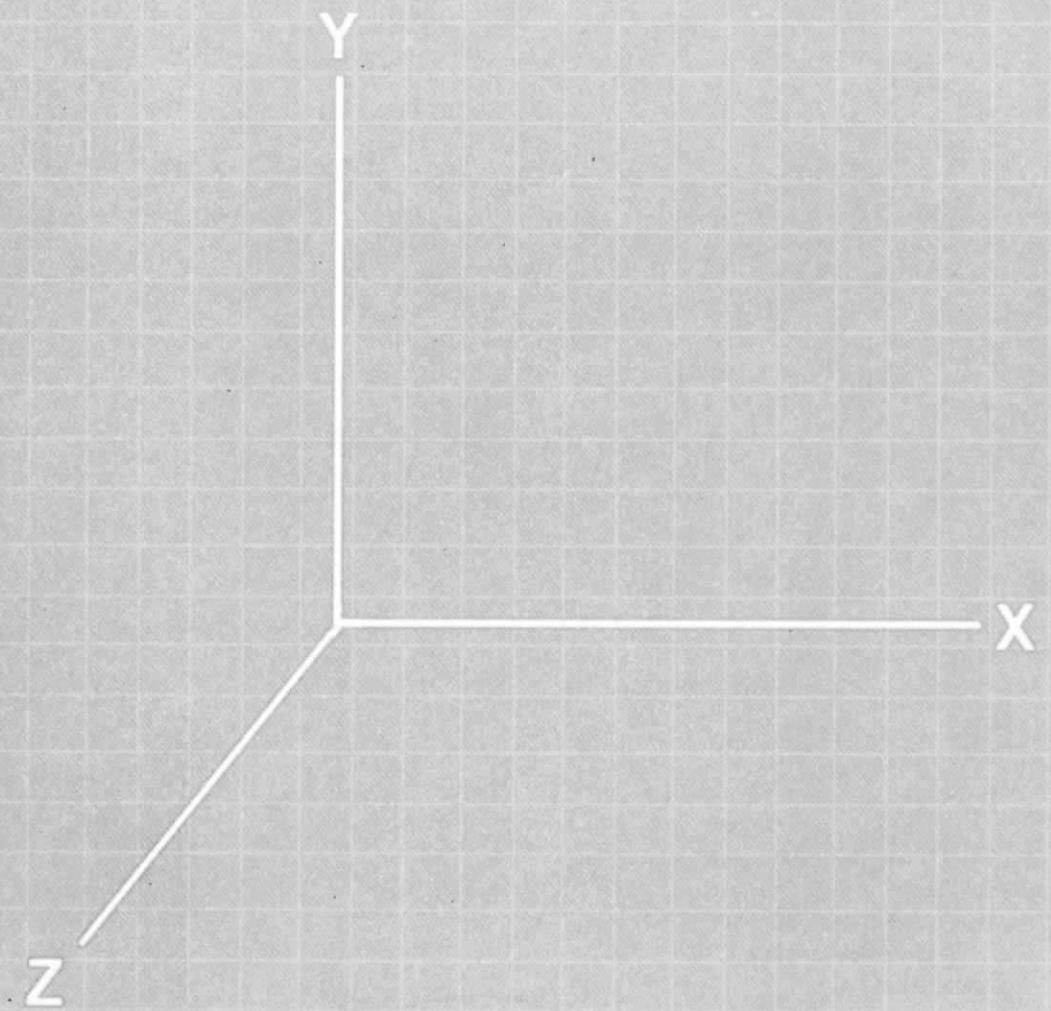
Memoria del XXIII Foro Nacional de Estadística



INSTITUTO NACIONAL
DE ESTADÍSTICA Y GEOGRAFÍA

AME

Memoria del XXIII Foro Nacional de Estadística



INSTITUTO NACIONAL
DE ESTADÍSTICA Y GEOGRAFÍA

AME

310.4 Foro Nacional de Estadística (23° : 2008 : Boca del Río, Veracruz).
Memoria del XXIII Foro Nacional de Estadística / Instituto Nacional de Estadística y Geografía, Asociación Mexicana de Estadística.-- México : INEGI, c2009

viii, 208 p. : il.

ISBN 978-607-494-003-9

"Universidad Veracruzana. Boca del Río, Veracruz del 10 al 12 de septiembre del 2008"

1. Estadística – Alocuciones, Ensayos, Conferencias. I. Instituto Nacional de Estadística y Geografía. II. Asociación Mexicana de Estadística.

 Instituto Nacional de Estadística y Geografía
Edificio Sede
Av. Héroe de Nacozari Sur Núm. 2301
Fracc. Jardines del Parque, CP 20276
Aguascalientes, Ags.

www.inegi.org.mx
atencion.usuarios@inegi.org.mx

Memoria del
XXIII Foro Nacional
de Estadística

Impreso en México
ISBN 978-607-494-003-9

Presentación

En estas memorias publicamos los resúmenes de algunas contribuciones libres o carteles presentados durante el XXIII Foro Nacional de Estadística. La institución sede fue la Universidad Veracruzana y el evento tuvo lugar en Boca del Río, Veracruz, del 10 al 12 de septiembre del 2008.

El volumen está integrado por dos secciones:

I. Tesis de licenciatura y maestría, y

II. Tesis doctorales y trabajos de investigación (metodológicos y/o aplicados).

Los trabajos fueron sometidos a un proceso de arbitraje coordinado por la mesa directiva de la Asociación Mexicana de Estadística. En este proceso, todos los artículos fueron revisados en su forma y contenido; siguiendo en todo momento criterios mínimos para evaluar la calidad en sus propuestas, resultados y aplicaciones, con énfasis en la originalidad para los trabajos de la Sección II.

Agradecemos profundamente a todos los autores por su entusiasmo y por la calidad de los trabajos presentados. Agradecemos además a todos aquellos colegas que nos apoyaron participando como árbitros, pues con su esfuerzo contribuyen a la calidad académica de estas memorias. En nombre de la Asociación Mexicana de Estadística expresamos también nuestra gratitud a la Universidad Veracruzana por el apoyo en la realización de este Foro, y al Instituto Nacional de Estadística y Geografía por patrocinar la edición e impresión de esta obra.

El Comité Editorial:

Elida Estrada Barragán,
Asael F. Martínez Martínez,
Luis Enrique Nieto Barajas,
Carlos Cuevas Covarrubias.

Índice general

Sección I. Tesis de Licenciatura y Maestría

Sección II. Tesis Doctorales y Trabajos de Investigación (Metodológicos y/o Aplicados)

Sección I

Tesis de Licenciatura y Maestría

Memoria FONE XXIII (2009), Sección I, 3–8

Estimaciones de intervalo en el análisis de máximos por bloques. Un estudio basado en simulaciones

Alejandro Cruz-Marcelo[a]
Rice University

Joaquín Ortega Sánchez
Centro de Investigación en Matemáticas

1. Introducción

Sea $\{X_i\}_{i\geq 1}$ una sucesión de variables aleatorias independientes e idénticamente distribuidas, cuya distribución común es desconocida. Definimos $M_n = \max\{X_1, \ldots, X_n\}$, para $n \geq 2$. En la modelación de valores extremos, es de interés determinar el comportamiento estadístico de M_n. El teorema de tipos para extremos afirma que las únicas distribuciones límite posibles para M_n (cuando n tiende a infinito) son las distribuciones generalizadas de valores extremos ($DGVE$) que tienen la forma

$$G(z) = \begin{cases} \exp -[1+\xi(\frac{z-\mu}{\sigma})]^{-\frac{1}{\xi}}, & \xi \neq 0; \\ \exp[-\exp -(\frac{z-\mu}{\sigma})], & \xi = 0, \end{cases}$$

donde $\mu \in \mathbf{R}$, $\sigma > 0$ y está definida en

$$\begin{cases} \{z : 1+\xi(\frac{z-\mu}{\sigma}) > 0\}, & \xi \neq 0; \\ -\infty < z < \infty, & \xi = 0. \end{cases}$$

Usando la distribución límite como una aproximación para muestras finitas, es posible argumentar que M_n puede ser aproximado por alguna $DGVE$. Por tanto, es de interés *estudiar el desempeño de diferentes técnicas estadísticas para ajustar una DGVE a una muestra dada.*

[a] alejandro@rice.edu

2. Métodos de estimación

En este trabajo comparamos tres tipos de estimadores de intervalo para cuantiles grandes que describimos a continuación (para una explicación más detallada consultar Cruz-Marcelo (2008)).

Método de Máxima Verosimilitud (MMV). Bajo este enfoque, obtenemos por máxima verosimilitud estimadores puntuales para los parámetros de la $DGVE$. Después, usando la normalidad asintótica de dichos estimadores, la cual Smith (1985) demostró es cierta cuando $\xi > -0.5$, obtenemos intervalos de confianza aproximada para cuantiles grandes . Dado que es posible expresar los cuantiles de una $DGVE$ como una función de sus parámetros, entonces usando el método delta podemos obtener intervalos de confianza aproximada para los cuantiles.

Método de Momentos Pesados por Probabilidad (MPP). Este método es una generalización del método de momentos. En el cálculo de los estimadores de MPP, los momentos teóricos a considerar para una variable aleatoria X, con función de distribución $F(x) = P[X \leq x]$, están definidos como $M_{p,r,s} = E\left[X^p \left\{F(X)\right\}^r \left\{1 - F(X)\right\}^s\right]$, donde p, r, y s son números reales. Para la $DGVE$, Hosking et al. (1985) mostraron que cuando $\xi < 1$, los estimadores de MPP existen y son asintóticamente normales, por lo que es posible obtener intervalos de confianza aproximada tanto para los parámetros como para los cuantiles. Finalmente, y al igual que ocurre con los estimadores de momentos, los estimadores de MPP pueden ser no factibles, es decir, que el soporte de la distribución ajustada no contenga todos los elementos de la muestra. De acuerdo a Dupuis (1996), la probabilidad de calcular estimadores no factibles es más alta cuando $\xi < -2$, pudiendo llegar hasta 0.2.

Intervalos de Verosimilitud–Confianza. En este enfoque, los estimadores de intervalo para parámetros y cuantiles corresponden a intervalos de verosimilitud calculados usando la verosimilitud perfil. Ejemplos de este procedimiento aplicado a la DGVE aparecen en Coles (2001). Es posible asociar a un intervalo de verosimilitud un nivel aproximado de confianza a través de la aproximación χ^2 (ver Serfling (1980), p. 158). Los intervalos resultantes de esta combinación son conocidos como intervalos de verosimilitud-confianza.

3. Estudio basado en simulaciones

Para comparar el desempeño de los estimadores de intervalo descritos en la sección anterior cuando son aplicados a muestras pequeñas, elaboramos un estudio basado en simulaciones con las siguientes características. Para cada combinación de los valores

$$\begin{aligned}\xi \;=\; & -0.4, -0.35, -0.3, -0.25, -0.2, -0.175, -0.15, -0.125, -0.1, -0.09, -0.08, -0.07, \\ & -0.06, -0.05, -0.04, -0.03, -0.02, -0.01, 0, 0.01, 0.02, 0.03, 0.04, 0.05, 0.06, \\ & 0.07, 0.08, 0.09, 0.1, 0.125, 0.15, 0.175, 0.2, 0.25, 0.3, 0.35, 0.4,\end{aligned}$$

y tamaños de muestra $n = 25, 50, 100$, generamos $10,000$ muestras de la correspondiente $DGVE$ con $\mu = 0$ y $\sigma = 1$. Para cada muestra calculamos estimadores de intervalo para cuantiles del 95 % y 99 %, respectivamente, siguiendo los tres métodos descritos en la sección anterior. En todos los casos consideramos un nivel de confianza del 95 %.

Comenzamos el análisis de los resultados calculando el porcentaje de casos en que no fue posible obtener los estimadores cuando usamos el MMV y los estimadores de MPP, respectivamente. Dichos porcentajes aparecen en la Figura 1 y en ambos casos incluyen las muestras cuyas estimaciones del parámetro de forma no se encuentran en el rango donde la normalidad asintótica es cierta. Para los estimadores de MPP, dicho porcentaje también incluye los casos en que los estimadores fueron no factibles.

Para comparar los estimadores calculamos su cobertura empírica, las cuales aparecen en las Figuras 2 y 3. La cobertura empírica fue calculada tanto como el porcentaje con respecto a: (a) el número total de simulaciones y (b) el total de muestras para las cuales fue posible calcular los estimadores. Los resultados sugieren que los intervalos de verosimilitud-confianza tienen mejor rendimiento pues presentan coberturas cercanas al 95 % en casi todos los casos. En contraste, los estimadores obtenidos por el MMV y el método de MPP, respectivamente, presentan una reducción en su cobertura empírica que varía dependiendo del signo de ξ.

Por otro lado, una mayor cobertura fue, en general, acompañada por una mayor longitud de los estimadores de intervalo. Para cuantificar dicho incremento calculamos, para cada muestra en cada escenario, el cociente entre la longitud de los intervalos obtenidos usando: (a) la verosimilitud perfil (numerador) y MMV (denominador), y (b) MPP (numerador) y MMV (denominador). Los cocientes aparecen en la Figura 4, donde las tres curvas se refieren a los cuartiles Q_1, Q_2 y Q_3, respectivamente. No tienen etiquetas pues $Q_1 \leq Q_2 \leq Q_3$.

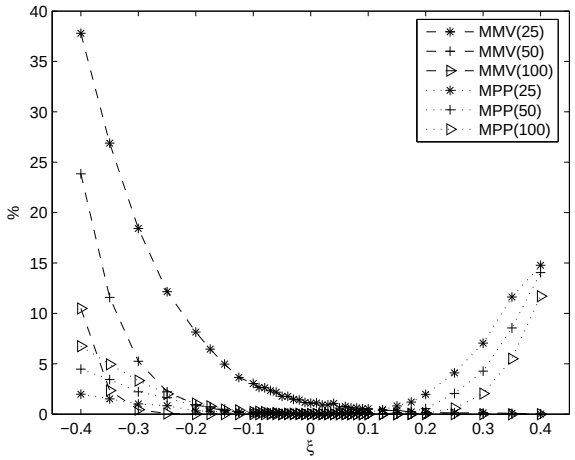
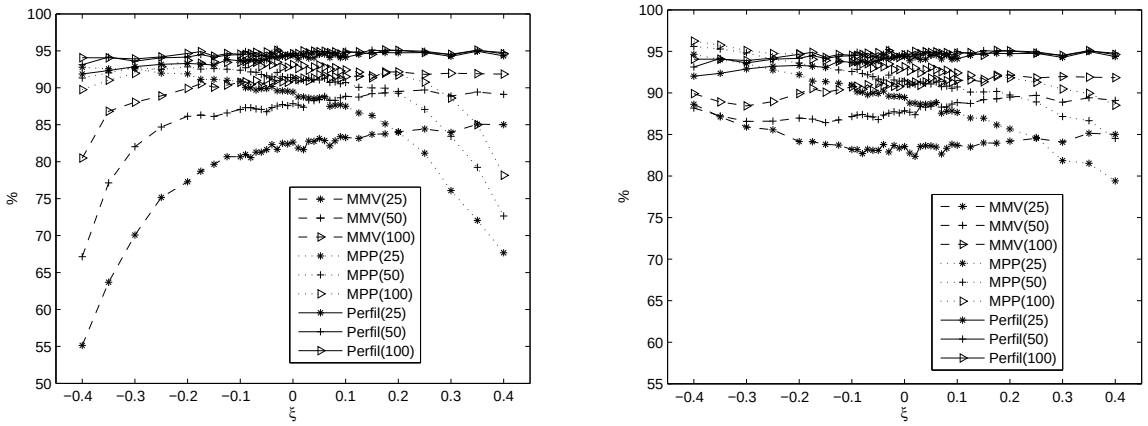


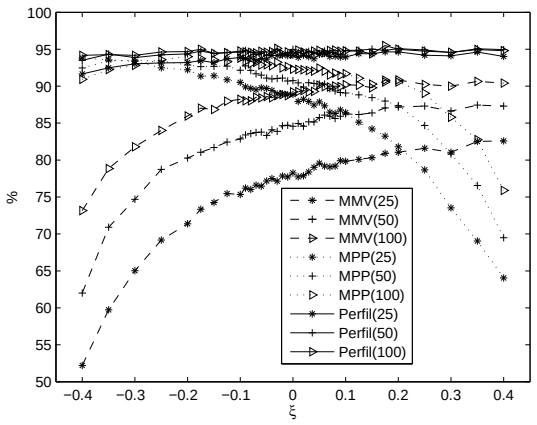
Figura 1: Porcentaje de muestras para las cuales no fue posible calcular los estimadores. El número entre paréntesis denota tamaño de muestra.



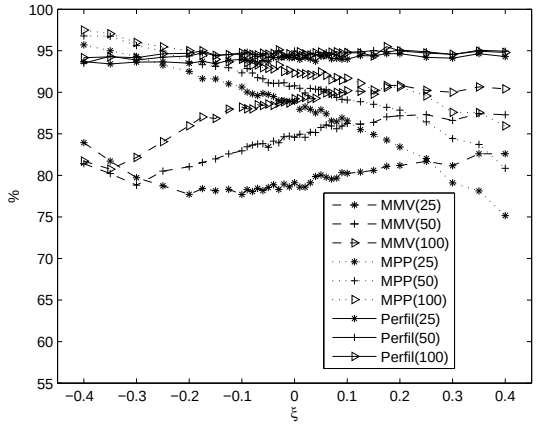
(a) Relativa al total de muestras

(b) Relativa al número de muestras con estimadores

Figura 2: Cobertura empírica por método para el cuantil del 95 %.

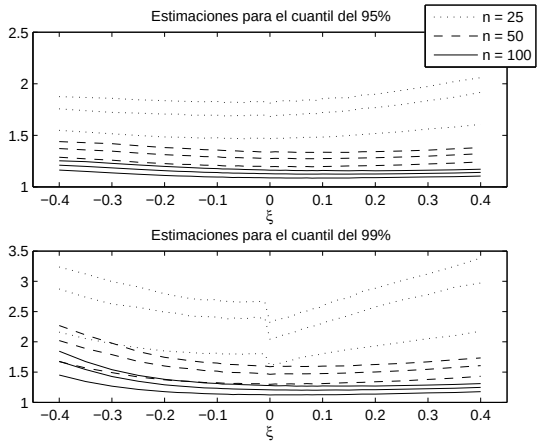


(a) Relativa al total de muestras

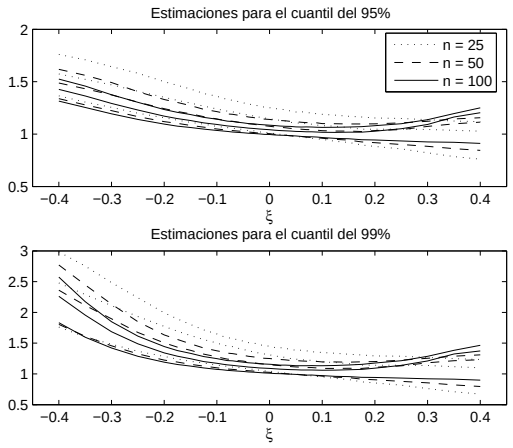


(b) Relativa al número de muestras sin errores

Figura 3: Cobertura empírica por método para el cuantil del 99 %.



(a) Perfil (numerador) y MMV (denominador)



(b) MPP (numerador) y MMV (denominador)

Figura 4: Cuartiles del cociente de la longitud entre los estimadores.

Concluimos que la diferencia relativa es mayor cuando los cálculos se refieren a cuantiles más altos y/o disminuye el tamaño de muestra.

4. Discusión

Con los resultados encontrados en este trabajo no pretendemos determinar cuál es el "mejor" método de estimación, en cambio, los resultados son información que puede usarse en la práctica para seleccionar alguno de los estimadores dependiendo del contexto del problema en cuestión. Aunque existen trabajos previos que comparan diferentes métodos de estimación para la $DGVE$ (ver por ejemplo, Hosking et al. (1985) o Coles y Dixon (1999)), la contribución de este trabajo a la literatura existente en el tema, consiste en haber comparado de manera sistemática y exhaustiva tres de los métodos más populares para obtener estimadores de intervalo.

Referencias

Coles, S. y Dixon, M. 1999. Likelihood-Based Inference for Extreme Values Models, *Extremes*, **2:1**, 5-23.

Coles, S. 2001. *An Introduction to Statistical Modelling of Extreme Values*. New York: Springer.

Cruz-Marcelo, A. 2008. *Estimaciones de Intervalo en el Análisis de Máximos por bloques. Un Estudio Basado en Simulaciones*. Tesis de Licenciatura en Matemáticas, UNAM, Enero 2008.

Dupuis, D. J. 1996. Estimating the Probability of Obtaining Nonfeasible Parameter Estimates of the Generalized Extreme-Value Distribution, *J. Statist. Comput. Sim.*, **56**, 23-38.

Hosking, J. R. M., Wallis, J.R., y Wood, E. F. 1985. Estimation of the Generalized Extreme Value Distribution by the Method of Probability Weighted Moments, *Technometrics*, **27**, 251-261.

Serfling, R. J. 1980. *Approximation Theorems of Mathematical Statistics*. New York: Wiley.

Smith, R. L. 1985. Maximum likelihood estimation in a class of non-regular cases, *Biometrika*, **42**, 67-90.

Memoria FONE XXIII (2009), Sección I, 9–15

Punto de cambio estructural en modelos de regresión lineal

María Guadalupe García Salazar[a], Blanca Rosa Pérez Salvador
Universidad Autónoma Metropolitana – Iztapalapa

1. Introducción

¿Te has preguntado si existe una cantidad en el ingreso personal de un individuo que le permite cambiar la tendencia en lo que ahorra? O ¿A qué edad, en qué nivel de la presión sanguínea, o cuántos cigarros fumados por un individuo, producen un cambio en la probabilidad de sufrir un infarto cardiaco? Para responder a estas preguntas suele utilizarse un modelo de regresión lineal que relaciona una variable respuesta Y con una o más variables explicativas $X_1, X_2, \ldots, X_p$, considerando que los parámetros del modelo son diferentes antes y después de un punto m, el cual se denota como **el punto de cambio estructural**, entendiéndose como cambio estructural a aquella alteración o modificación de los parámetros en un modelo de regresión.

Uno de los métodos que se ha utilizado para probar la ocurrencia de un cambio estructural es conocido como la prueba Chow (Gujarati, 1997). La característica principal de la prueba Chow es que el posible momento en que ocurre el cambio está bien determinado, es decir, se conoce el punto m para el cual los datos muestrales antes de m siguen un modelo de regresión lineal diferente al modelo de regresión lineal que siguen los datos después de m. La prueba Chow es inaplicable si se desconoce el punto donde pudo producirse el cambio.

Cuando se desconoce si se ha producido un cambio en la región de observación, lo adecuado es formular una prueba de hipótesis para determinar su existencia y, posteriormente estimar el momento en que el cambio se ha producido. Estudios sobre la prueba de hipótesis han sido efectuados por Beckman y Cook (1979), Horvath y Shao (1993), Antoch y

[a]mggasa@gmail.com

Hušcová (2001), quienes formularon como hipótesis nula el que no existe cambio en la región de observación y encontraron la región crítica asintótica. Por otro lado, la estimación del punto de cambio estructural fue abordado por Muggeo (2003), quien consideró que el método de máxima verosimilitud era inaplicable en este problema.

En este trabajo se analiza el punto de cambio estructural en el modelo de regresión lineal múltiple con p variables explicativas, se encuentra la región crítica mediante el cociente de verosimilitud y, a diferencia de otros trabajos, se deduce la distribución exacta de la estadística de prueba. El trabajo consta de 3 secciones, la primera es esta introducción, en la sección 2 se formula la prueba de hipótesis, se obtiene la estadística de prueba y la región crítica, en la sección 3 se presentan las conclusiones.

2. Prueba de hipótesis

2.1. Región crítica

Sea la variable de respuesta Y relacionada con las variables explicativas $X_1, X_2, \ldots, X_p$, mediante el modelo,

$$Y_i = \begin{cases} \beta_0 + \sum_{j=1}^{p} X_{ij}\beta_j + \varepsilon_i = X_i^T\beta + \varepsilon_i & \text{si} \quad i \leq m \\ \beta_0^* + \sum_{j=1}^{p} X_{ij}\beta_j^* + \varepsilon_i = X_i^T\beta^* + \varepsilon_i & \text{si} \quad i > m \end{cases}$$

con $\beta_j \neq \beta_j^*$ al menos para una j, $0 \leq j \leq p$ y $\varepsilon_i \sim N(0, sigma^2)$. Las hipótesis a probar son:

$$H_0 : m \geq n \quad \text{contra} \quad H_1 : m < n.$$

Para encontrar la región crítica se utiliza el cociente de verosimilitud

$$\frac{\max L_{H_0}}{\max L_{H_1}} = \frac{\max\limits_{\beta,\sigma} \frac{1}{(2\pi)^{n/2}\sigma^n}\ e^{-\frac{1}{2\sigma^2}(Y-X\beta)^T(Y-X\beta)}}{\max\limits_{\beta_1,\beta_2,\sigma,m} \frac{1}{(2\pi)^{n/2}\sigma^n}\ e^{-\frac{1}{2\sigma^2}[(Y_1-X_1\beta_1)^T(Y_1-X_1\beta_1)+(Y_2-X_2\beta_2)^T(Y_2-X_2\beta_2)]}} \leq \lambda \qquad (1)$$

donde $Y^T = (Y_1^T|Y_2^T)$ con $Y_1 \in \mathbf{R}^m$, $Y_2 \in \mathbf{R}^{n-m}$ y $X^T = (X_1^T|X_2^T)$ con $X_1 \in \mathbf{R}^{m\times(p+1)}$ y $X_2 \in \mathbf{R}^{(n-m)\times(p+1)}$. Haciendo los desarrollos pertinentes se tiene que la relación (1) es equivalente a,

$$\min_m \frac{(Y_1 - X_1\hat{\beta}_1)^T(Y_1 - X_1\hat{\beta}_1) + (Y_2 - X_2\hat{\beta}_2)^T(Y_2 - X_2\hat{\beta}_2)}{(Y - X\hat{\beta})^T(Y - X\hat{\beta})} = \min_m \left\{ \frac{SCE_{1m} + SCE_{2m}}{SCE} \right\} \leq \lambda^*$$

Esta expresión sólo tiene sentido cuando $p+1 \leq m \leq n-p-1$ debido a que no se puede realizar una regresión lineal con menos observaciones que parámetros.

2.2. Distribución de la estadística de prueba

La estadística de prueba que se dedujo en la sección anterior, es $\min_m \left\{ \frac{SCE_{1m}+SCE_{2m}}{SCE} \right\}$. Bajo el supuesto que H_0 es cierta se tiene que $Y_i = \beta_0+\beta_1 X_{1i}+\ldots+\beta_p X_{pi}+\varepsilon_i$; con $\varepsilon_i \sim N(0,\sigma^2)$, para $i = 1, 2, \ldots, p$; entonces la suma de cuadrados del error es $SCE = Y^T(I_n - X(X^TX)^T X^T)Y$; y debido a que $(I_n - X(X^TX)^{-1}X^T)$ es una matriz simétrica, idempotente y de rango igual a $n-p-1$, existe una matriz P de tamaño $n \times (n-p-1)$ tal que $SCE = Y^T P P^T Y = W^T W$, con $W = P^T Y \sim N(0, \sigma^2 I)$ y $P^T P = I_{n-p-1}$.

Teorema 2.1. *Sea $SCE = W^T W$ con $W \sim N(0, \sigma^2 I)$ y sean SCE_{1m} la suma de cuadrados del error de los m primeros datos de la muestra y SCE_{2m} la suma de cuadrados del error de los restantes $n-m$ datos de la muestra, entonces $SCE_{1m} + SCE_{2m} = W^T W - W^T Q_m W$ donde $Q_m = P^T \left(\begin{array}{c|c} X_1(X_1^T X_1)^{-1} X_1^T & 0 \\ hline 0 & X_2(X_2^T X_2)^{-1} X_2^T \end{array} \right) P$; con $p+1 \leq m \leq n-p-1$.*

Demostración. Obsérvese que $SCE_{1m} + SCE_{2m} = Y^T N_m Y$ con

$$N_m = \left(\begin{array}{c|c} I_m - X_1(X_1^T X_1)^{-1} X_1^T & 0 \\ \hline 0 & I_{n-m} - X_2(X_2^T X_2)^{-1} X_2^T \end{array} \right)$$

Sea Z_m una matriz tal que $SCE_{1m} + SCE_{2m} = Y^T N_m Y = W^T Z_m W$, entonces $W^T Z_m W = Y^T P Z_m P^T Y$ y por lo tanto $N_m = P Z_m P^T$, lo que nos lleva a que $Z_m = P^T N_m P = I_{n-p-1} - Q_m$ donde $Q_m = P^T \left(\begin{array}{c|c} X_1(X_1^T X_1)^{-1} X_1^T & 0 \\ \hline 0 & X_2(X_2^T X_2)^{-1} X_2^T \end{array} \right) P$ y $W^T Z_m W = W^T W - W^T Q_m W$. □

Como una consecuencia del teorema anterior, la región crítica de la prueba de hipótesis es

$$\min_m \frac{W^T Z_m W}{W^T W} = \min_m \left\{ 1 - \frac{W^T Q_m W}{W^T W} \right\} = 1 - \max_m \frac{W^T Q_m W}{W^T W} \leq \lambda^*.$$

Entonces, la región crítica es igual a $\max_m\{u^T Q_m u\} > \lambda^{**}$, con $u = w/||w||$. Para obtener el valor de λ^{**} con un nivel de significancia dado, es necesario encontrar la función de distribución de $u^T Q_m u$. La variable aleatoria $u^T Q_m u$ depende sólo de $n-p-2$ coordenadas

del vector u, pues $||u|| = 1$ y la función de densidad conjunta de las primeras $n-p-2$ coordenadas del vector u se presenta en el siguiente teorema.

Teorema 2.2. *Dado $W \sim N(0, \sigma^2 I)$, sea V el vector $V^T = (V_1, \ldots, V_{n-p-2})$ tal que $V_i = W_i/||W||$, entonces la función de densidad conjunta de V bajo H_0, es*

$$f_V(v_1, v_2, \ldots, v_{n-p-2}) = \begin{cases} \frac{\Gamma((n-p-1)/2)}{2(\pi)^{(n-p-1)/2}(1-v^Tv)^{1/2}}, & si\ v^Tv < 1 \\ 0, & en\ otro\ caso. \end{cases}$$

Demostración. La función de densidad conjunta del vector W es

$$f_W(w_1, w_2, \ldots, w_{n-p-1}) = \frac{e^{-w^Tw/2\sigma^2}}{(2\pi)^{(n-p-1)/2}\sigma^{n-p-1}}$$

y la función de densidad conjunta de $v_1, v_2, \ldots, v_{n-p-2}, w_{n-p-1}$ es igual a

$$f_{V,w_{n-p-1}}(v_1, \ldots, v_{n-p-2}, w_{n-p-1}) = f_W(w_1(V, w_{n-p-1}), \ldots, w_{n-p-2}(V, w_{n-p-1}), w_{n-p-1})|J|.$$

Entonces, se debe encontrar las variables w_i en función de V y de w_{n-p-1}. Por definición de V se tiene que $w_j^2 = v_j^2||W|| = v_j^2 \sum_{i=1}^{n-p-1} w_i^2$. En este sistema de ecuaciones se encuentra recursivamente el valor de w_i^2.

- Para $j = 1$ se tiene que

$$w_1^2 = v_1^2 \sum_{i=1}^{n-p-1} w_i^2 = v_1^2 w_1^2 + v_1^2 \sum_{i=2}^{n-p-1} w_i^2 \Rightarrow\ w_1^2 = \frac{v_1^2}{1-v_1^2} \sum_{i=2}^{n-p-1} w_i^2.$$

  De aquíse sigue que

$$\sum_{i=1}^{n-p-1} w_i^2 = w_1^2 + \sum_{i=2}^{n-p-1} w_i^2 = \frac{v_1^2}{1-v_1^2} \sum_{i=2}^{n-p-1} w_i^2 + \sum_{i=2}^{n-p-1} w_i^2 = \frac{1}{1-v_1^2} \sum_{i=2}^{n-p-1} w_i^2$$

- Para $j = 2$ se tiene que

$$w_2^2 = v_2^2 \sum_{i=1}^{n-p-1} w_i^2 = \frac{v_2^2}{1-v_1^2} \sum_{i=2}^{n-p-1} w_i^2 = \frac{v_2^2}{1-v_1^2} w_2^2 + \frac{v_2^2}{1-v_1^2} \sum_{i=3}^{n-p-1} w_i^2$$

$$\Rightarrow w_2^2 = \frac{v_2^2}{1-v_1^2-v_2^2} \sum_{i=3}^{n-p-1} w_i^2.$$

De aquíse sigue que

$$\sum_{i=1}^{n-p-1} w_i^2 = \frac{1}{1-v_1^2}\sum_{i=2}^{n-p-1} w_i^2 = \frac{1}{1-v_1^2}\left(w_2^2 + \sum_{i=3}^{n-p-1} w_i^2\right)$$
$$= \frac{1}{1-v_1^2}\left(\frac{v_2^2}{1-v_1^2-v_2^2}\sum_{i=3}^{n-p-1} w_i^2 + \sum_{i=3}^{n-p-1} w_i^2\right)$$
$$\Rightarrow \sum_{i=1}^{n-p-1} w_i^2 = \frac{1}{1-v_1^2-v_2^2}\sum_{i=3}^{n-p-1} w_i^2$$

siguiendo este procedimiento, recursivamente hasta $j = n-p-2$ se llega a:

$$\sum_{i=1}^{n-p-1} w_i^2 = \frac{w_{n-p-1}^2}{1-v_1^2-\ldots-v_{n-p-2}^2} = \frac{w_{n-p-1}^2}{1-v^Tv}.$$

Finalmente, se llega a que w_j^2 es igual a $w_j^2 = \frac{w_{n-p-1}^2}{1-v^Tv}\, v_j^2$; para $j = 1, 2, \ldots, n-p-2$.

Ahora se va a encontrar el jacobiano de esta transformación, para ello se deriva w_i respecto a v_j y se obtiene

- $\frac{\partial w_j}{\partial v_j} = \frac{|w_{n-p-1}|}{\sqrt{1-v^Tv}} + \frac{|w_{n-p-1}|v_j^2}{(1-v^Tv)^{3/2}} = \frac{|w_{n-p-1}|}{\sqrt{1-v^Tv}}\left(1 + \frac{v_j^2}{(1-v^Tv)}\right)$ para $1 \le j \le n-p-2$y
- $\frac{\partial w_j}{\partial v_i} = \frac{|w_{n-p-1}|v_jv_i}{(1-v^Tv)^{3/2}}$ para $1 \le i \ne j \le n-p-2$

entonces, el jacobiano de esta transformación es $|J| = \left(\frac{|w_{n-p-1}|}{\sqrt{1-v^Tv}}\right)^{n-p-2}\left|I + \frac{1}{1-v^Tv}vv^T\right|$. El determinante de la matrix $I + \frac{1}{1-v^Tv}vv^T$ se encuentra como el producto de sus valores propios. Los vectores propios de la matriz $I + \frac{1}{1-v^Tv}vv^T$ son v y cualquier vector z ortogonal a v. El valor propio asociado a v es $\frac{1}{1-v^Tv}$, ya que $(I + \frac{1}{1-v^Tv}vv^T)v = v + \frac{v^Tv}{1-v^Tv}v = \frac{1}{1-v^Tv}v$ y el valor propio asociado a z ortogonal a v es uno debido a que $(I + \frac{1}{1-v^Tv}vv^T)z = z$, por lo tanto $\left|I + \frac{1}{1-v^Tv}vv^T\right| = \frac{1}{1-v^Tv}y|J| = \frac{|w_{n-p-1}^{n-p-2}|}{(1-v^Tv)^{(n-p)/2}}$.

Ahora se sustituye $w_i^2 = \frac{w_{n-p-1}^2}{1-v^Tv}$ y $|J|$ en f_W y se obtiene

$$f_{V,W_{n-p-1}}(v_1, v_2, \ldots, v_{n-p-2}, w_{n-p-1}) = \frac{|w_{n-p-1}^{n-p-2}|e^{-\frac{w_{n-p-1}^2}{2\sigma^2(1-v^Tv)}}}{(2\pi)^{(n-p-1)/2}\sigma^{n-p-1}(1-v^Tv)^{(n-p)/2}}.$$

Finalmente, la función de densidad del vector V se encuentra integrando esta función de densidad respecto de la variable w_{n-p-1}

$$f_V(v_1, v_2, \ldots, v_{n-p-2}) = \frac{\displaystyle\int_{-\infty}^{\infty} |w_{n-p-1}^{n-p-2}| e^{-\frac{w_{n-p-1}^2}{2\sigma^2(1-v^Tv)}} dw_{n-p-1}}{(2\pi)^{(n-p-1)/2}\sigma^{n-p-1}(1-v^Tv)^{(n-p)/2}}$$

La integral en el numerador de esta expresión se resuelve usando el cambio de variable$v = \frac{w_{n-p-1}^2}{2\sigma^2(1-v^Tv)}$. Con este cambio de variable se tiene que $w_{n-p-1}^2 = 2\sigma^2(1-v^Tv)v$, lo que implica que

- $|w_{n-p-1}^{n-p-2}| = 2^{(n-p-2)/2}\sigma^{n-p-2}(1-v^Tv)^{(n-p-2)/2}|v|^{(n-p-2)/2}$
- $dw_{n-p-1} = 2^{1/2}\sigma(1-v^Tv)dv/2\sqrt{v}$

Al hacer el cambio de variable se tiene que

$$f_V(v_1, v_2, \ldots, v_{n-p-2}) = \frac{2^{(n-p-3)/2}\displaystyle\int_{-\infty}^{\infty} |v|^{((n-p-1)/2)-1} e^{-v} dv}{(2\pi)^{(n-p-1)/2}(1-v^Tv)^{1/2}} = \frac{\displaystyle\int_{0}^{\infty} |v|^{((n-p-1)/2)-1} e^{-v} dv}{2\pi^{(n-p-1)/2}(1-v^Tv)^{1/2}}$$

Por tanto, $f_V(v_1, v_2, \ldots, v_{n-p-2}) = \begin{cases} \frac{\Gamma((n-p-1)/2)}{2(\pi)^{(n-p-1)/2}(1-v^Tv)^{1/2}}, & \text{si } v^Tv < 1 \\ 0, & \text{en otro caso.} \end{cases}$ □

Finalmente, la región crítica $\max_m u^T Q_m u \geq \lambda^{**}$ con $u_i = v_i$ $i = 1, 2, \ldots, n-p-2$ y $u_{n-p-1} = \sqrt{1-v^Tv}$; se puede encontrar al considerar que

$$P\left(\max_m u^T Q_m u \geq \lambda^{**}\right) = 1 - P\left(\max_m u^T Q_m u < \lambda^{**}\right)$$

y

$$P\left(\max_m u^T Q_m u < \lambda^{**}\right) = \int \cdots \int_{A_{\lambda^{**}}} f_V(v_1, v_2, \ldots, v_{n-p-2}) dv_1 dv_2 \ldots dv_{n-p-2} = 1 - \alpha$$

con $A_{\lambda^{**}} = \{v \in \mathbb{R}^{n-p-2} | u^T Q_{p+1} u < 1 - \lambda^*, \cdots, u^T Q_{n-p-1} u < \lambda^{**}\}$.

Esta última integral es difícil de calcular debido a que la región de integración depende del vector u. Por ende, para calcular el valor de λ^{**} para un α dado, se utiliza para este trabajo el método de Monte Carlo y la integración numérica. Por falta de espacio no se presenta ningún ejemplo.

3. Conclusiones

En este trabajo se encontró la región crítica para la prueba cuya hipótesis nula dice que no existe un cambio en la región de observación. Se utilizó el método del cociente de máxima verosimilitud. La aportación del trabajo es haber encontrado la función de distribución exacta de la estadística de prueba. Sin embargo, el cálculo de las probabilidades de la distribución encontrada no se pueden obtener analíticamente, por lo que se debe recurrir a una integración estocástica y o a una integración numérica. Un resultado que no se presenta en este escrito es la obtención del estimador del punto de cambio, y el análisis de sus propiedades de sesgo y varianza.

Referencias

Antoch, Jaromir and Hušková, Marie 2001. Permutation tests in change point analysis. *Statistics and Probability Letters*, **53**, 37-46.

Beckman and Cook 1979. Testing for Two-Phase Regressions. *Thechnometrics*, **21**, 1.

Gujarati, Damodar N. 1997. *Econometría*. Ed. McGraw Hill, pp. 258-260.

Horvárth, Lajos and Shao, Qi-Man 1993. Limit theorems for the union-intersection test. *Journal of statistical planning and inference*, **44**, 133-148.

Mugeo V. R. 2003. Estimating regression models with unknown break-points, *Statistics in Medicine*, **22**, 3055-3071.

Memoria FONE XXIII (2009), Sección I, 17–23

Modelación espacio–temporal de la temperatura máxima anual en el estado de Veracruz

Luis Hernández Rivera, Sergio Francisco Juárez Cerrillo
Universidad Veracruzana

1. Introducción

Motivados por el problema de identificar evidencia de calentamiento en la temperatura en el estado de Veracruz, proponemos un modelo para máximos espacio-temporales basado en el suavizamiento local con polinomios propuesto por Hall y Tajvidi (2000). El modelo ajustado detecta tendencias de incremento en las temperaturas máximas anuales registradas en 22 estaciones climatólogicas del estado de Veracruz. Resultados similares se obtienen en localizaciones en las cuales no se han observado datos.

2. El Modelo

Consideremos al proceso de máximos espacio-temporales $\left\{X(\mathbf{s},t) : \mathbf{s} \in D \subset \mathbb{R}^d,\ t \in [0,\infty)\right\}$, donde $X(\mathbf{s},t)$ sigue la Distribución del Valor Extremo Generalizada (DVEG). Suponemos que los parámetros varian suavemente en función del tiempo t y la localización $\mathbf{s}$, es decir, $\mu(\mathbf{s},t)$, $\sigma(\mathbf{s},t)$ y $\xi(\mathbf{s},t)$ son funciones suaves y $X(\mathbf{s},t) \sim G(x;\mu(\mathbf{s},t),\sigma(\mathbf{s},t),\xi(\mathbf{s},t))$ donde

$$G(x;\mu,\sigma,\xi) = \exp\left\{-\left[1+\xi\left(\frac{x-\mu}{\sigma}\right)\right]_+^{-1/\xi}\right\},$$

con $\mu,\xi \in \mathbb{R}$, $\sigma \in \mathbb{R}^+$, y $[z]_+ = \text{máx}(z,0)$. Suponemos también que los máximos son independientes aunque no necesariamente idénticamente distribuidos. Consideremos que observamos al proceso en las localizaciones $S = \{\mathbf{s}_1,\ldots,\mathbf{s}_n\}$ y que en cada una de estas

localizaciones lo observamos en los mismos tiempos $\{t_1, \ldots, t_m\}$. Fijemos a la localización $\mathbf{s} \in S$, y sean $\{X(\mathbf{s}, t_j)\}_{j=1}^m$ las realizaciones en el tiempo del proceso $X(\mathbf{s}, t)$. Sea $g(x|\theta(\mathbf{s}, t)) = \log f(x|\theta(\mathbf{s}, t))$, donde f es la función de densidad de la DVEG y $\theta(\mathbf{s}, t) = (\mu(\mathbf{s}, t), \sigma(\mathbf{s}, t), \xi(\mathbf{s}, t))$. Dado un ancho de banda $h > 0$ y un núcleo K simétrico definimos $K_j(t) = K((t - t_j)/h)$. Sean v_0 y v_1 vectores candidatos de $\theta(\mathbf{s}, t)$ y $\dot{\theta}(\mathbf{s}, t) = d\theta(\mathbf{s}, t)/dt$, respectivamente. Sea $w_j = w_j(v_0, v_1) = \theta(\mathbf{s}, t_j) = v_0 + (t_j - t)v_1$, entonces la función de log-verosimilitud del modelo está dada por

$$\ell(v_0, v_1|t) = \sum_{j=1}^{m} g(X(\mathbf{s}, t_j)|w_j(v_0, v_1))K_j(t).$$

Los estimadores de $\theta(\mathbf{s}, t)$ se obtienen fijando la forma paramétrica local para θ en el punto t. Por ejemplo el estimador *local constante*, $\hat{\theta}(\mathbf{s}, t) = \hat{v}_0$, donde $\hat{v}_0$ maximiza a $\ell(v_0, 0|t)$ con respecto a v_0; o el estimador *local lineal*, $\tilde{\theta}(\mathbf{s}, t) = \tilde{v}_0$, donde $(\tilde{v}_0,\tilde{v}_1)$ maximiza $\ell(v_0, v_1|t)$ con respecto a (v_0, v_1). El lector que no esté familiarizado con este tipo de modelos lo invitamos a consultar Davison y Ramesh (2000), Hall y Tajvidi (2000) y Beirlant y Goegebeur (2004), así como los capítulos 7 y 10 de Beirlant et al. (2004).

Puesto que el modelo se ajusta en las localizaciones espaciales de manera marginal, éste no considera la posible asociación espacial entre distintas localidades. Sin embargo, con base en el supuesto de variación suave espacial de $X(\mathbf{s}, t)$, proponemos el siguiente enfoque de interpolación espacial de la DVEG en sitios en los cuales no se tienen datos. Sea $\mathbf{s}$ la localización de un sitio en el cual no se tienen datos, supongamos que $\mathbf{s}$ está en el casco convexo de S y sean $\hat{\mu}(\mathbf{s}_i, t)$, $\hat{\sigma}(\mathbf{s}_i, t)$ y $\hat{\xi}(\mathbf{s}_i, t)$ las estimaciones de los parámetros de la DVEG, donde $\mathbf{s}_i \in S$ y el tiempo t está fijo. Proponemos los siguientes estimadores de los parámetros de la DVEG en $\mathbf{s}$ para el tiempo t

$$\hat{\mu}(\mathbf{s}, t) = \sum_{i=1}^{n} w_{\mathbf{s}_i}(\mathbf{s})\hat{\mu}(\mathbf{s}_i, t), \quad \hat{\sigma}(\mathbf{s}, t) = \sum_{i=1}^{n} w_{\mathbf{s}_i}(\mathbf{s})\hat{\sigma}(\mathbf{s}_i, t) \quad \text{y} \quad \hat{\xi}(\mathbf{s}, t) = \sum_{i=1}^{n} w_{\mathbf{s}_i}(\mathbf{s})\hat{\xi}(\mathbf{s}_i, t), \quad (1)$$

donde $w_{\mathbf{s}_i}(\mathbf{s})$ son pesos que satisfacen

$$\textstyle\sum_{i=1}^{n} w_{\mathbf{s}_i}(\mathbf{s}) = 1, \quad \text{y} \quad w_{\mathbf{s}_j}(\mathbf{s}_i) = \delta_{ij} = \begin{cases} 1, & i = j \\ 0, & i \neq j \end{cases}, i = 1, \ldots, n.$$

Los pesos que utilizamos están dados por

$$w_{\mathbf{s}_i}(\mathbf{s}) = \begin{cases} 1, & \mathbf{s} = \mathbf{s}_i, \\ 0, & \mathbf{s} \in S \setminus \{\mathbf{s}_i\}, \\ \dfrac{1}{1 + \sum\limits_{r=1, r\neq i}^{m} \frac{d(\mathbf{s},\mathbf{s}_i)}{d(\mathbf{s},\mathbf{s}_r)}}, & \mathbf{s} \notin S, \end{cases}$$

donde d es una distancia. En nuestro caso usamos la longitud de arco formado por los dos puntos en la esfera terrestre.

3. Temperaturas Máximas Anuales en Veracruz

El modelo lo ajustamos, con S-Plus, a las temperaturas máximas anuales registradas de 1960 al 2002 en 22 estaciones climatológicas del estado de Veracruz.

Como ilustración, la Figura 1 presenta los resultados para la estación Ángel R. Cabada con los estimadores locales constantes. En la gráfica del parámetro de forma ξ notamos que este tiende a aumentar en el tiempo, lo que indica que el peso de la cola superior de la DVEG ajustada aumenta con el tiempo. Por lo que la probabilidad de observar temperaturas máximas cada vez más extremas aumenta durante el período de observación. Este fue el caso en la mayoría de las 22 estaciones analizadas, véase Hernández Rivera (2007). Para ilustrar la interpolación espacial elegimos arbitrariamente las localizaciones con latitud y longitud $(21.13°, 97.80°)$, $(20.80°, 97.60°)$, $(20.40°, 97.00°)$ y $(19.90°, 96.80°)$. La Figura 2 muestra los resultados de la interpolación espacial. Notamos que el parámetro de forma (tercera columna) tiende a aumentar en el tiempo, lo que da evidencia de un aumento en las temperaturas máximas.

Las Figuras 3 y 4 muestran la interpolación espacial, con kriging universal, de los parámetros de forma de los años de 1960 y 2002, respectivamente. En ambas gráficas se pueden apreciar las zonas del estado con mayores temperaturas (aquellas en rojo). Cuando comparamos las gráficas, apreciamos un incremento en el parámetro de forma, lo que da evidencia de un incremento en las temperaturas máximas de 1960 al 2002.

4. Investigación Adicional

En este artículo hemos presentado avances de un proyecto de investigación en modelación de máximos espacio-temporales con aplicaciones en climatología. Actualmente nos encontramos calculando expresiones para estimar los errores estándar de las estimaciones propuestas, así como también en el desarrollo de procedimientos para evaluar la bondad del ajuste del modelo.

Referencias

Beirlant, J. and Goegebeur, Y. 2004. Local Polynomial Maximum Likelihood Estimation for Pareto-type Distribution, *Journal of Multivariate Analysis*, **89**, 97-118.

Beirlant, J., Goegebeur, Y., Segers, J. and Teugels, J. 2004. *Statistics of Extremes. Theory and Applications* , Wiley: Chichester.

Davison, A.C., and Ramesh, N. I. 2000. Local Likelihood Smoothing of Sample Extremes, *J. R. Statistical Society Series B*, **62**, 191-208.

Hall, P., and Tajvidi, N. 2000. Nonparametric Analysis of Temporal Trend When Fitting Parametric Models to Extreme-Value Data, *Statistical Science*, **15**, 153-167.

Hernández Rivera, L. 2007. *Estimación no Paramétrica de Tendencias en Extremos Espacio-Temporales.* Tesis de Maestría en Matemáticas, Facultad de Ciencias Físico Matemáticas, Benemérita Universidad Autónoma de Puebla.

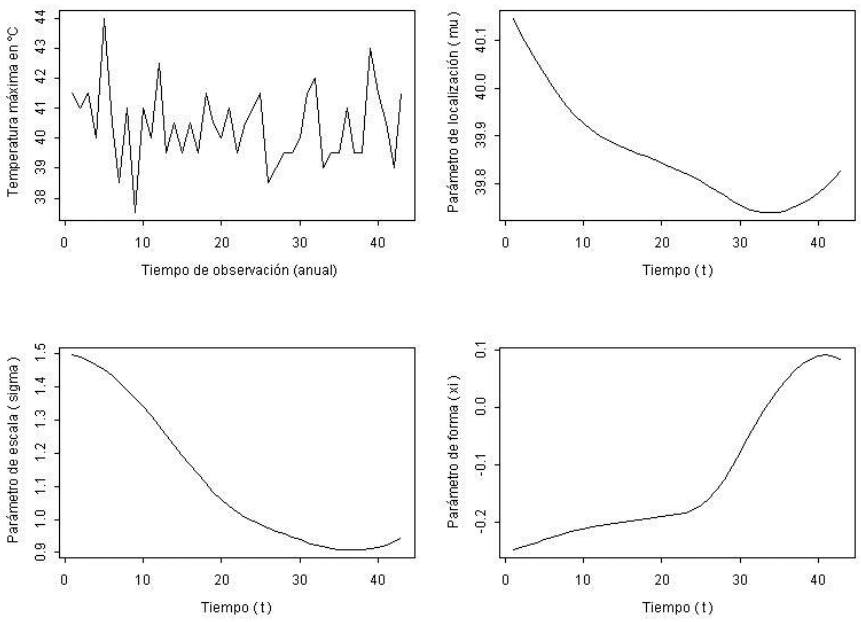


Figura 1: Temperaturas máximas anuales registradas en la estación Ángel R. Cabada y ajuste del modelo.

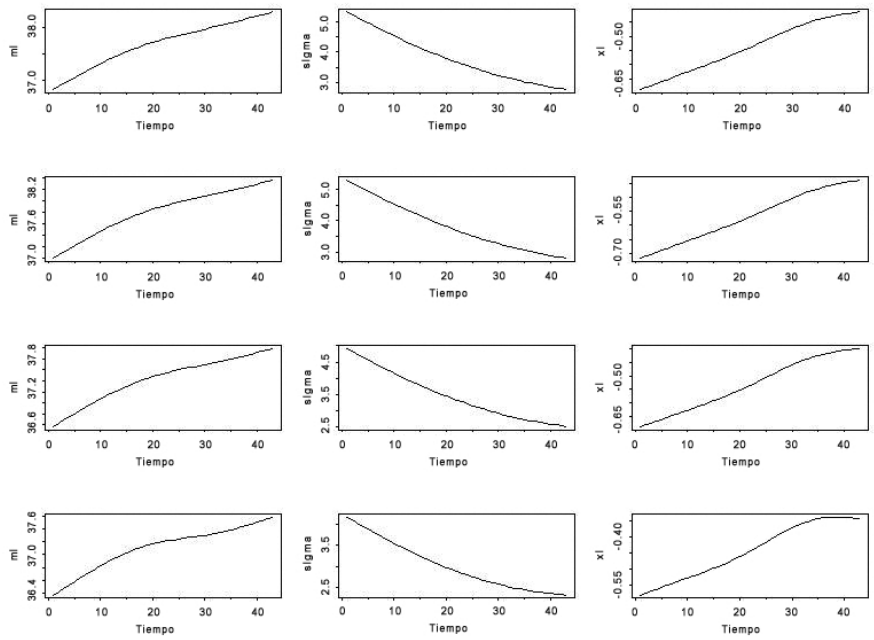


Figura 2: Interpolación espacial del modelo en cuatro puntos.

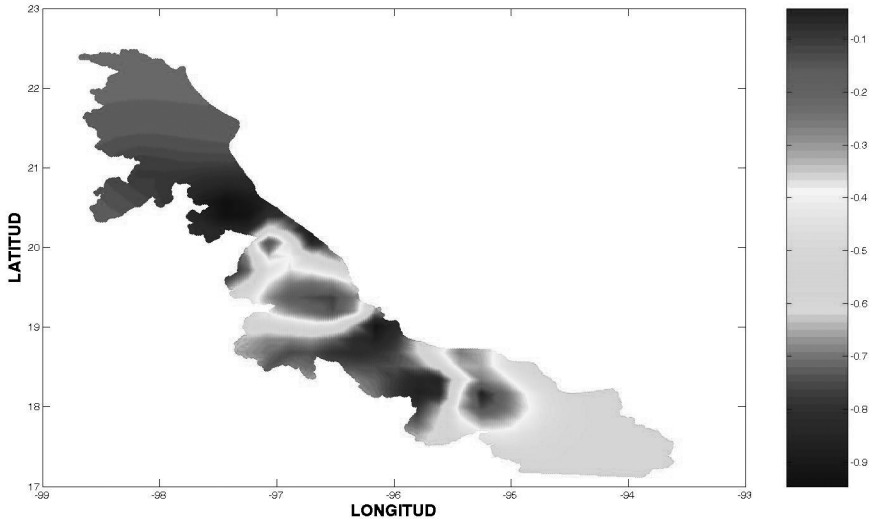


Figura 3: Interpolación espacial del parámetro de forma en 1960.

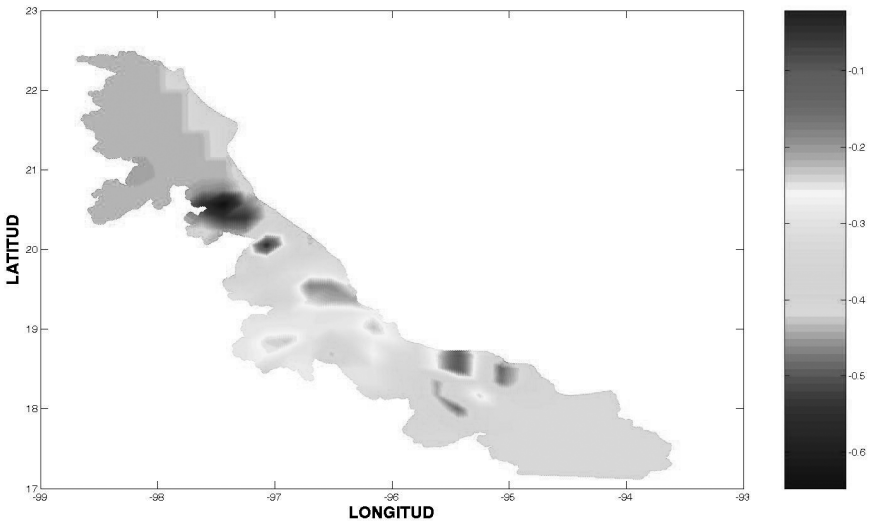


Figura 4: Interpolación espacial del parámetro de forma en 2002.

Memoria FONE XXIII (2009), Sección I, 25–29

Estudio de la validez del constructo inteligencia emocional, en estudiantes universitarios, mediante un modelo de análisis factorial confirmatorio: escala CASVI

Elena Vicente Galindo[a], J. Antonio Castro Posada
Universidad de Salamanca – España

Purificación Vicente Galindo, Purificación Galindo Villardón
Universidad de Salamanca – España

1. Introducción

La *Inteligencia Emocional* (IE) es entendida como una metahabilidad cuya base es el autodominio de las competencias afectivas (Goleman, 1996). A pesar de tratarse de una metahabilidad que incide en la totalidad del potencial del ser humano y que se encuentra implicada en una vida exitosa, ha recibido poca atención en la literatura científica.

El constructo IE presenta gran dificultad para su evaluación. Hay diferentes tests en la literatura que evalúan las diversas componentes de la inteligencia emocional y las diferencias individuales. Estos tests han sido desarrollados, la mayoría, en Estados Unidos (Mayer y Salovey, 1995; Davies, *et al*, 1998). Están dirigidos al medio laboral.

2. Objetivo

Este trabajo se encuadra dentro de la perspectiva psicométrica de evaluación de la IE y tiene como objetivo estudiar la validez factorial y la fiabilidad de la escala CASVI (Vicente, 2007), un instrumento diseñado para evaluar el perfil de IE de estudiantes universitarios.

[a]primer_canaryavg@hotmail.com

3. Material y métodos

3.1. Muestra

La muestra se recogió mediante muestreo no probabilístico y está formada por 642 estudiantes universitarios, de las dos universidades de Salamanca (España), la Universidad de Salamanca (pública) y la Universidad Pontificia de Salamanca (privada), con edades comprendidas entre 18 y 24 años. ($\bar{x}$= 21 años; s= 2). La distribución por género de la muestra es la siguiente: 64.9 % son mujeres y el 35.1 % varones. La distribución por carreras es la siguiente: CC SS 12.3 %, C. Salud 12.6 %; CC y Tecnología 16.8 %; Gestión y Leyes 14.2 %; Diplomaturas 15.3 %; Otras 4.1 %.

3.2. Instrumento

La escala fue elaborada a partir de afirmaciones extraídas de publicaciones sobre IE, fundamentalmente de Goleman, 1996; 1999 y Vallés y Vallés, 2000. Considerando los factores propuestos por la Asociación Internacional de *Inteligencia Emocional Aplicada* (ISAEI). La escala está pensada para evaluar las habilidades involucradas en la IE: a) evaluación y expresión de las emociones, tanto las propias como las de los demás; b) asimilación de la emoción y el pensamiento; c) entendimiento y análisis de las emociones; d) relación emocional para promover un crecimiento emocional e intelectual.

La primera versión constaba de 300 ítems que fueron sometidos al juicio de tres jueces, quienes eliminaron redundancias e ítems no claramente relacionados con el constructo y dejaron 213 cuestiones. Cada ítem se corresponde con un enunciado que representa un rasgo de comportamiento paradigmático de la IE, expresado en escala tipo Likert, con 6 posibles respuestas. El cuestionario completo puede verse en Vicente 2007.

4. Proceso de simplificación de la escala y propiedades psicométricas

Realizamos un análisis factorial exploratorio de las respuestas al cuestionario propuesto para evaluar IE, para intentar buscar una estructura factorial subyacente. Hemos utilizado como

método de extracción de ejes el ACP (Análisis de Componentes Principales) con rotación VARIMAX.

Se realizó un primer análisis con los 213 ítems tras el cual se eliminaron todos aquellos para los cuales el factor de carga era inferior a 0.50 y también aquellos que cargaban en varios ejes. Así se efectuó una primera simplificación, que dejó la escala en 156 ítems, que se agrupaban en 7 ejes; de ellos, se eligieron, para este trabajo, los 4 primeros componentes; una segunda simplificación, trabajando únicamente con los ítems de esos 4 componentes, dejó la escala en 43 ítems (escala CASVI).

Los cuatro primeros ejes absorben el 44.36 % de la inercia y la interpretación de los factores de carga nos permite afirmar que: el primer eje factorial está formado por ítems que se corresponden con aspectos de ***Autoestima*** (afectivo-cognitivo- comportamental); es decir sentimiento valorativo que los individuos tiene de si mismos, en diferentes niveles. El eje 2 se corresponde con ***Autoconcepto*** (Responsabilidad y Empatía), el eje 3 puede ser interpretado como ***Centración en sí mismo*** y el eje 4 como ***Dependencia Emocional***. Se ha analizado la consistencia interna para el instrumento completo y se ha alcanzado un valor del alfa de Cronbach de 0.94. Para las distintas subescalas, los coeficientes de consistencia interna son: Autoestima 0.93; Autoconcepto 0.86; Centración 0.82; Dependencia Emocional 0.71.

5. Modelo de análisis factorial confirmatorio

Contrastamos ahora la estructura de las subescalas obtenidas mediante Análisis Factorial Exploratorio. Los análisis estadísticos se han llevado a cabo utilizando el paquete AMOS 6.

5.1. Modelo para analizar autoestima

Teniendo en cuenta que la configuración del eje representaba distintas facetas de la autoestima, se contrastó un modelo con dos factores en el que 12 de los los ítems configuraban los aspectos afectivos de la autoestima, y 6 ítems conforman los aspectos cognitivo-comportamental. El modelo de dos factores establece la división del conjunto de la subescala en dos factores que hemos denominado *Autoestima-Afectivo* y *Autoestima Cognitivo Comportamental*. Téngase en cuenta que los modelos complejos, como es el caso, requieren el

reconocimiento de covarianzas entre los términos de error de las variables.

El diagrama se muestra en la Figura 1.

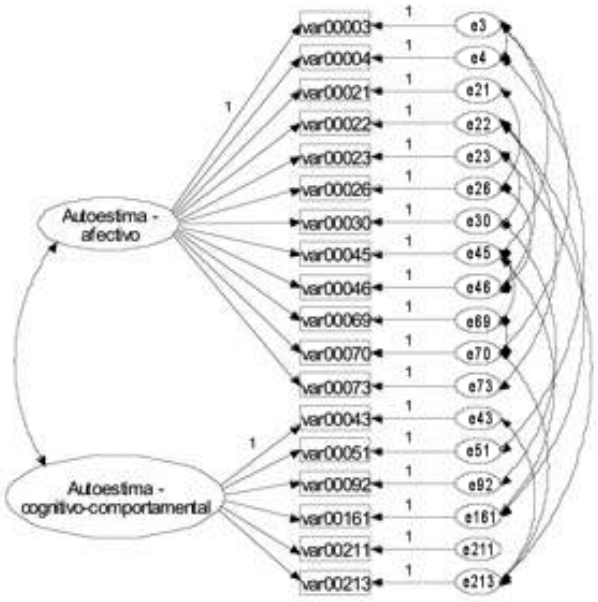


Figura 1: Diagrama del modelo de dos factores para la autoestima

Para el modelo bidimensional el valor de la χ^2 es 178.77 con 116 grados de libertad. Este modelo se ajusta bien a los datos ($p = 0.077$).

Los índices de ajuste incremental (NFI, RFI, IFI, TLI y CFI) fueron todos superiores a 0,90.

El cálculo de los estimadores de los parámetros del modelo bidimensional con sus errores estándar y sus significaciones nos permite afirmar que todos los coeficientes son significativamente distintos de cero. Los coeficientes de la subescala *Autoestima-Afectivo* parecen estar más cerca de los valores teóricos que los de la segunda subescala. El ítem con peor ajuste es el 26 (Soy feliz), tal vez por la ambigüedad de su formulación. Un análisis similar (Vicente, 2007) nos permitió afirmar que existen modelos unidimensionales para las subescalas de Autoconcepto, Centración y Dependencia Emocional que se ajustan a los datos

6. Conclusiones

Se presenta una escala construida con 213 ítems para evaluar IE en universitarios. El Análisis Factorial Exploratorio nos ha permitido simplificar la escala inicial y construir otra que contiene 43 ítems (ESCALA CASVI) que configuran 4 dimensiones de la IE: **Autoestima** (afectivo, cognitivo-comportamental), **Autoconcepto** (responsabilidad y empatía), **Centración en si mismo** y **Dependencia Emocional**. La escala propuesta tiene alta consistencia interna, no sólo en las subescalas, sino también en su conjunto. Hemos desarrollado modelos de Análisis Factorial Confirmatorio que han permitido contrastar la configuración de la escala de IE de una manera científicamente válida: corrobora la estructura en cuatro dimensiones obtenida en el análisis exploratorio. Se trata de un estudio piloto que requiere investigaciones posteriores para profundizar, al menos, en los siguientes aspectos: 1) buscar intervalos de valores para clasificar a los universitarios según categorías de IE, comparando los resultados de nuestra escala con alguno de los patrones que han probado su validez en colectivos semejantes y 2) estudiar su capacidad evaluativa en otras poblaciones y campos.

Referencias

Davies, Michaela, Lazar Stankov y Richard D. Roberts. 1998. Emotional Intellegence in search of an elusive construct. *Journal of Personality and Social Psychology*, **75**(4), 989-1015.

Goleman, Daniel. 1996. *Inteligencia Emocional*. Barcelona. Ed Kairos, S.A.

Goleman, Daniel. 1999. *La práctica de la Inteligencia Emocional*. Barcelona. Ed Kairos. S.A.

Mayer, John D. y Peter Salovey. 1995. Emotional intelligence and construction and regulation of feelings. *Applied Preventive Psychology*, **4**(3), 197-208.

Vallés, Antonio y Consol Vallés. 2000. *Inteligencia Emocional. Aplicaciones Educativas*. Madrid: Ed. EOS.

Vicente, María E. 2007. *Creación de una escala para medir inteligencia emocional. Estudio piloto*. Tesina Licenciatura. Universidad Pontificia. Salamanca. España.

Sección II

Tesis Doctorales y Trabajos de Investigación (Metodológicos y/o Aplicados)

Memoria FONE XXIII (2009), Sección II, 33–37

Estimación del modelo de espacio de estados lineal gaussiano con observaciones censuradas

Francisco J. Ariza-Hernández[a]
Colegio de Postgraduados

Gabriel A. Rodríguez-Yam[b]
Universidad Autónoma Chapingo

1. Introducción

Los modelos de espacio de estados han tenido una amplia aceptación para analizar las series de tiempo que provienen de áreas tales como la biología, la economía, la agronomía, las ciencias ambientales, etc. Frecuentemente los datos pueden estar censurados debido a muchos factores. Por ejemplo, en las ciencias ambientales es común que el equipo de monitoreo no detecte valores pequeños. En este trabajo, para estimar los parámetros de un modelo de espacio de estados cuando se tienen observaciones censuradas por la izquierda, se propone una aproximación a la verosimilitud a través de integración Monte Carlo. Con ligeros cambios, el método puede implementarse a otros tipos de censura.

Hopke *et al.* (2001) realizan una revisión sobre métodos de imputación múltiple para manejar datos perdidos y/o censurados, Andrieu y Doucet (2002) usan mezclas aleatorias de distribuciones normales para representar la distribución a posteriori de modelos de espacio de estados gaussianos parcialmente observados, Park et al. (2007) proponen un método de imputación para modelos de la clase ARMA censurados. Ariza-Hernandez y Rodriguez-Yam (2008) usan el algoritmo EM Estocástico en modelos de espacio de estados censurados por la izquierda. En la siguiente sección se formula el modelo de espacio de estados lineal gaussiano cuando se tienen observaciones censuradas y en la Sección 3 se presenta un ejemplo de aplicación a datos reales.

[a] arizahfj@colpos.mx
[b] grodrigu@correo.chapingo.mx

2. Modelo de espacio de estados censurado

Considere el modelo de espacio de estados lineal estándar

$$Y_t = G_t\alpha_t + w_t; \quad t = 1, 2, \ldots, n. \tag{1}$$

donde $\{Y_t\}$ es una serie observable, $\{G_t\}$ es una secuencia de constantes, $w_t \sim N(0, \sigma_w^2)$ y α_t, una variable no observable, es la variable de estado que satisface

$$\alpha_{t+1} = F_t\alpha_t + v_t; \quad t = 1, 2, \ldots, n. \tag{2}$$

donde $\{F_t\}$ es una secuencia de constantes, $v_t \sim N(0, \sigma_v^2)$, y $\{w_t\}$ no está correlacionada con $\{v_t\}$. En algunos casos, como el planteado en esta trabajo, las sucesiones $\{G_t\}$ y $\{F_t\}$ son constantes.

Sea $\mathbf{y} := (y_1, \ldots, y_n)'$ el vector de observaciones de tamaño n, $\boldsymbol{\alpha} := (\alpha_1, \ldots, \alpha_n)'$ el vector de variables de estados y $\boldsymbol{\theta}$ el vector de parámetros de este modelo. La función de verosimilitud para este modelo está dada por

$$L(\boldsymbol{\theta}; \mathbf{y}) = \int p(\mathbf{y}, \boldsymbol{\alpha}; \boldsymbol{\theta})\, d\boldsymbol{\alpha} = \int p(\mathbf{y}|\boldsymbol{\alpha}; \boldsymbol{\xi}) p(\boldsymbol{\alpha}; \boldsymbol{\psi})\, d\boldsymbol{\alpha}. \tag{3}$$

donde $\boldsymbol{\xi}$ y $\boldsymbol{\psi}$ son los vectores de parámetros de la densidad condicional de $\mathbf{y}$ dado $\boldsymbol{\alpha}$ y de la densidad de $\boldsymbol{\alpha}$ respectivamente. Note que $\boldsymbol{\theta} := (\boldsymbol{\xi}, \boldsymbol{\psi})$.

Considere ahora que algunas de las observaciones $y_1, y_2, \ldots, y_n$ del modelo en (1) están censuradas, y sea

$$\delta_t := \begin{cases} 1, & \text{si } y_t \text{ no está censurada,} \\ 0, & \text{si } y_t \text{ está censurada,} \end{cases} \qquad t = 1, 2, \ldots, n. .$$

Un caso importante que se considera en este trabajo es cuando se asume que el proceso de estados $\{\alpha_t\}$ es un proceso autoregresivo, i.e.,

$$\alpha_t = \phi_1\alpha_{t-1} + \cdots + \phi_p\alpha_{t-p} + \eta_t, \tag{4}$$

donde $\eta_t \sim \text{iid } N(0, \tau^2)$, $t = 1, 2, \ldots, n$, y p es un entero no negativo. Sea $\boldsymbol{\psi} := (\phi_1, \ldots, \phi_p, \tau^2)$ el vector de parámetros del proceso de estados. Si las observaciones están censuradas por la izquierda, entonces la verosimilitud "completa" es

$$\begin{aligned} L(\boldsymbol{\theta}; \mathbf{y}, \boldsymbol{\alpha}, \boldsymbol{\delta}) &= f(\mathbf{y}|\,\boldsymbol{\alpha}; \boldsymbol{\xi}) f(\boldsymbol{\alpha}; \boldsymbol{\psi}) \\ &= \left(\prod_{t=1}^{n} f(y_t\,|\alpha_t; \boldsymbol{\xi})^{\delta_t} F_{Y_t|\alpha_t}(y_t; \boldsymbol{\xi})^{1-\delta_t} \right) \\ &\qquad \times \ |\mathbf{V}|^{-1/2} e^{-\boldsymbol{\alpha}^T \mathbf{V}^{-1} \boldsymbol{\alpha}/2} / (2\pi)^{n/2}, \end{aligned} \tag{5}$$

donde f y F representan la función de densidad y la función acumulada de la distribución normal respectivamente, $\boldsymbol{\delta} := (\delta_1, \delta_2, \ldots, \delta_n)'$ y $\mathbf{V} := \text{cov}\{\boldsymbol{\alpha}\}$. De (5) se sigue que la función de verosimilitud se obtiene como la integral

$$L(\boldsymbol{\theta}; \mathbf{y}, \boldsymbol{\delta}) = \int L(\boldsymbol{\theta}; \mathbf{y}, \boldsymbol{\alpha}, \boldsymbol{\delta}) \, \mathbf{d}\boldsymbol{\alpha}. \tag{6}$$

Aún para este caso simple, la integral en (6) no puede ser calculada explícitamente, y así, los estimadores de máxima verosimilitud son difíciles de obtener. En este trabajo se aplica el muestreo de importancia para aproximar la integral en (6). La función de importancia que se propone, denotada por $g_{ic}(\boldsymbol{\alpha}|\,\mathbf{y})$, es la densidad condicional de $\boldsymbol{\alpha}$ dado $\mathbf{y}$ "ignorando" censura en las observaciones. Así, si $\boldsymbol{\alpha}^{(1)}, \boldsymbol{\alpha}^{(2)}, \ldots, \boldsymbol{\alpha}^{(m)}$ es una muestra aleatoria de $g_{ic}(\boldsymbol{\alpha}|\,\mathbf{y})$, por el muestreo de importancia se tiene que

$$\hat{L}_{MI}(\boldsymbol{\theta}; \mathbf{y}, \boldsymbol{\delta}) = \frac{1}{m} \sum_{j=1}^{m} \frac{L(\boldsymbol{\theta}; \mathbf{y}, \boldsymbol{\alpha}^{(j)}, \boldsymbol{\delta})}{g_{ic}(\boldsymbol{\alpha}^{(j)}|\mathbf{y})}, \tag{7}$$

el cual es un estimador consistente de (6) (Robert y Casella 2004). Para muestrear de $g_{ic}(\boldsymbol{\alpha}|\,\mathbf{y})$, sea $y_{1:t} := (y_1, \ldots, y_t)'$ y considere la factorización

$$g_{ic}(\boldsymbol{\alpha}|\mathbf{y}) = g_{ic}(\alpha_n|y_{1:n}) \prod_{t=1}^{n-1} g_{ic}(\alpha_t|\alpha_{t+1:n}, y_{1:n}), \tag{8}$$

donde

$$\begin{aligned} g_{ic}(\alpha_t|\alpha_{t+1:n}, y_{1:n}) &= g_{ic}(\alpha_t|\alpha_{t+1}, y_{1:t}) \\ &= g_{ic}(\alpha_t|y_{1:t}) f(\alpha_{t+1}|\alpha_t) / g_{ic}(\alpha_{t+1}|y_{1:t}) \\ &\propto g_{ic}(\alpha_t|y_{1:t}) f(\alpha_{t+1}|\alpha_t). \end{aligned}$$

Entonces, un estimador de máxima verosimilitud (EMV) aproximado de $\boldsymbol{\theta}$ es

$$\hat{\boldsymbol{\theta}} = \arg \max_{\theta \in \Theta} \; \hat{L}_{MI}(\theta; \mathbf{y}, \boldsymbol{\delta}).$$

3. Ejemplo de aplicación

En esta sección se analiza la serie $\{y_t, t = 1, 2, \ldots, 154\}$ de fósforo en solución reactiva medida en milagramos por litro, mg/L en un río monitoreado mensualmente de diciembre de 1994 a septiembre de 2007 por el Departamento de Ecología del estado de Washington, E.U. en la

estación 08C070 Cedar R Logan St Renton. Estos datos fueron tomados del sitio de internet `http://www.ecy.wa.gov/apps/watersheds/riv/regions/state.asp`.

En la Figura 1 se muestra la serie $\{y_t\}$ (línea sólida). Los valores no detectados (censurados) por el instrumento de medición se representan con un triángulo con pico hacia abajo. La serie $\{y_t\}$ presenta 26.62 % de censura.

Para modelar esta serie se propone el modelo de espacios de estados siguiente

$$Y_t = \mu + \alpha_t + \varepsilon_t, \tag{9}$$

donde μ es la media general y $\varepsilon_t \sim \text{iid}N(0, \sigma^2), t = 1, 2, \ldots, n$ representan los errores. Para el proceso de estados se considera un proceso $AR(1)$, i.e.,

$$\alpha_t = \phi\alpha_{t-1} + \eta_t, \tag{10}$$

donde $\eta_t \sim \text{iid}N(0, \tau^2), t = 1, 2, \ldots, n$, $|\phi| < 1$ y ε_t y η_t, $t = 1, 2, \ldots, n$ son independientes. El objetivo es estimar $\boldsymbol{\theta} := (\mu, \sigma^2, \phi, \tau^2)$. Como se puede verificar, las densidades $g_{ic}(\alpha_t|y_{1:t})$ y $g_{ic}(\alpha_{t+1}|y_{1:t})$ son ambas normales, *i.e.*, $\alpha_t|y_{1:t} \sim N(\alpha_{t|t}, \Omega_{t|t})$ y $\alpha_{t+1}|y_{1:t} \sim N(\hat{\alpha}_{t+1}, \Omega_{t+1})$ donde $\alpha_{t|t}$, $\Omega_{t|t}$, $\hat{\alpha}_{t+1}$ y Ω_{t+1} se obtienen de las recursiones de Kalman. También, la densidad $g_{ic}(\alpha_t|\alpha_{t+1:n}, y_{1:n})$ en (8), es normal $N(\mu_t^*, \nu_t^*)$, donde

$$\mu_t^* = \frac{\tau^2\alpha_{t|t} + \phi\alpha_{t+1}\Omega_{t|t}}{\tau^2 + \phi^2\Omega_{t|t}} \qquad \text{y} \qquad \nu_t^* = \frac{\tau^2\Omega_{t|t}}{\tau^2 + \phi^2\Omega_{t|t}}; \quad t = 1, \ldots, n-1.$$

Después de calcular a $\hat{L}_{MI}(\boldsymbol{\theta}; \mathbf{y}, \boldsymbol{\delta})$ mediante el muestreo de importancia y usando un algoritmo de optimización no lineal se obtiene que el estimador de $\boldsymbol{\theta}$, es $\hat{\boldsymbol{\theta}} = (6.51, 5.23, 0.75, 2.63)$ y sus desviaciones estándar estimadas respectivas (0.04, 0.10, 0.01, 0.07). En la Figura 1, se muestra el ajuste de la serie (linea punteada).

Referencias

Andrieu, C. y Doucet A. 2002. Particle filtering for partially observed Gaussian state space models. *J. R. Statist. Soc. B*, **64**, *Part* 4, *pp* 827–836.

Ariza-Hernandez, F. J. y Rodríguez-Yam, G. A. 2008. Estimación con el algoritmo EM estocástico de modelos de espacio de estados con observaciones censuradas. *INEGI, Memorias del XXII Foro Nacional de Estadistica*, pp. 1–6.

Hopke, P. K., Liu, C. y Rubin, D. B. 2001. Multiple imputation for multivariate data with missing and below-threshold measurement: Times-series concentrations of pollutants in the Arctic. *Biometrics* **57**, pp. 22–33.

Park, J. W., Genton, M. G. y Ghosh, S. K. 2007. Censored times series analysis with autoregressive moving average models. *Canadian Journal of Statistics*, **35**, 1, pp. 151–168.

Robert, C. P. y Casella, G. 2004. *Monte Carlo Statistical Methods*. Second Edition. Springer.

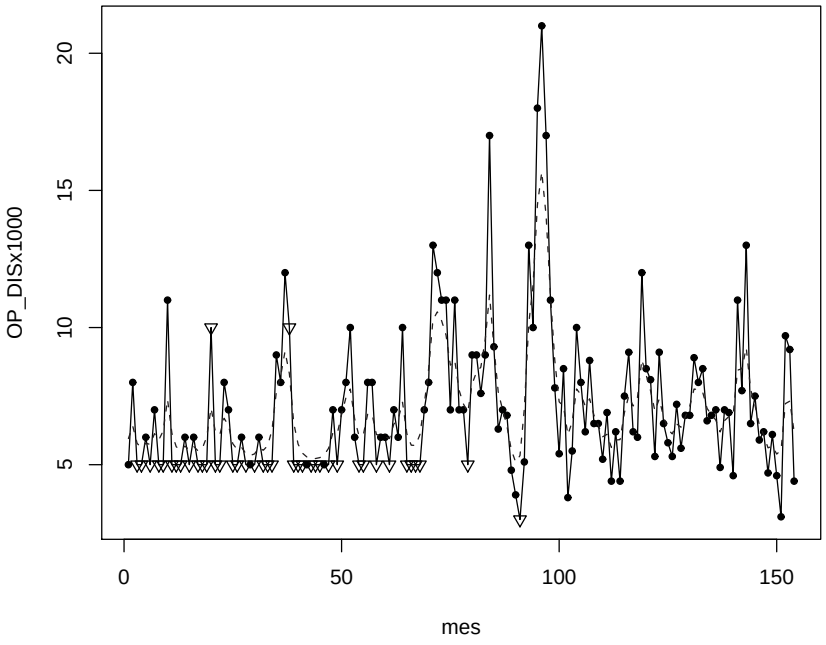


Figura 1: Serie de tiempo observada (linea sólida), serie de tiempo ajustada (línea punteada)

Memoria FONE XXIII (2009), Sección II, 39–48

Estudio computacional para encontrar el valor óptimo de la distancia esperada entre un punto y una variable aleatoria real

Luis Cruz-Kuri
Instituto de Ciencias Básicas – UV

Agustín Jaime García Banda, Ismael Sosa Galindo
Facultad de Ciencias Administrativas y Sociales – UV

1. Introducción

En algunos cursos elementales de estadística matemática, para establecer una propiedad de la varianza de una variable aleatoria X, se propone resolver el problema de optimización (mínimo) de la función $f(r)$ definida por el valor esperado del cuadrado de $X - r$, donde r es un número real arbitrario. Utilizando manipulaciones algebraicas sencillas, se encuentra que la cantidad que produce el valor óptimo de $f(r)$ ocurre cuando $r = E(X)$ y que el valor óptimo de $f(r)$ corresponde a la varianza de X. Aquí, la métrica involucrada es la usual L_2. Si ahora se utiliza, por ejemplo, la métrica $d(a, b)$, valor absoluto de $(a - b)$, es decir, la métrica L_1, entonces el problema de encontrar el óptimo de la función $g(r)$ definida por el valor esperado del valor absoluto de $X - r$, ocurre cuando $r = \text{Med}(X)$, una mediana de X (cualquiera del así llamado intervalo mediano). La justificación de este resultado requiere un análisis matemático más delicado, aunque elemental. Todo lo anterior da lugar a la pregunta natural acerca de lo que podría suceder cuando intervienen otras métricas, incluyendo por supuesto la del supremo. Más generalmente, sea X una variable aleatoria fija con una distribución dada, discreta o absolutamente continua, y sea d una métrica sobre el espacio de los números reales. Para cada número real r se define la función $f(r) = E(d(X, r))$. El objetivo es encontrar el valor óptimo (mínimo) de la función f. En el presente trabajo se describe como este problema se resuelve con el apoyo de un programa de

cómputo matemático, ejecutado en una computadora personal, para algunas métricas y para algunas variables aleatorias. En particular, se ilustra lo ya indicado en las dos situaciones específicas anteriores, a saber, que para la métrica euclidiana L_2 el valor óptimo se obtiene con $r = E(X)$ y, para la métrica L_1, el valor óptimo se alcanza con $r = \text{Med}(X)$. Asimismo, se busca encontrar resultados análogos en un contexto multivariado. Estas tareas pueden realizarse con el apoyo de un programa de cómputo matemático poderoso y versátil, como lo es *Mathematica*, que es el que se ha usado para el presente trabajo.

2. Ejemplos de utilización del programa *Mathematica*

En el presente trabajo, se dan las instrucciones pertinentes en Mathematica para lograr acercarse a los objetivos señalados. Esto se realiza en los tres tipos de procesamiento que este programa maneja, a saber, simbólico, numérico y gráfico. Las instrucciones aparecen en varios lugares, así como las salidas generadas por su ejecución. Si se tiene una colección de datos univariados reales, tal como A = {x1,x2,...,xn}, la instrucción en Mathematica Mean[A] produce al ejecutarla la media aritmética de la colección A. De igual manera, la instrucción Median[A] conduce a la mediana de la colección A. Lo anterior se ilustra a continuación.

A:=2,5,6,7,8,20

Median[A] *Resultado*: 13/2 Mean[A] *Resultado*: 8

Por supuesto, utilizando la capacidad de integración simbólica y/o numérica, también se pueden encontrar medias (valores esperados) y medianas de variables aleatorias con distribuciones absolutamente continuas. Para un contexto de estadística multivariada, también es posible aplicar *Mathematica* y definir funciones reales de varias variables.

Como primera ilustración de análisis gráfico, considérese la función real $f(r, s)$ de las variables reales s y r definida mediante las instrucciones en *Mathematica* que aparecen en los dos renglones que siguen.

f[r_,s_]:=Sqrt[(7-r)^2+(11-s)^2]+Sqrt[(9-r)^2+(8-s)^2]+

Sqrt[(13-r)^2+(17-s)^2]+Sqrt[(18-r)^2+(28-s)^2]

Adelante aparecen las instrucciones para construir la gráfica de $f(r, s)$. Dicha gráfica es la que aparece en la Figura 1 a continuación. El objetivo es encontrar los mínimos loca-

les de $f(r, s)$ y esto no por medios analíticos puros sino mas bien apoyado en el tipo de procesamiento que permite *Mathematica*.

```
Plot3D[f[r,s],{r,0,20},{s,0,30},ViewPoint
→ {-2.996, -1.493, 1.107}]
```

```
f2[r_,s_]:= (7-r)^2 + (11-s)^2 + (9-r)^2 +
(8-s)^2 + (13-r)^2 + (17-s)^2 + (18-r)^2
+ (28-s)^2. Plot3D[f2[r,s],r,7,18,s,11,28]
```

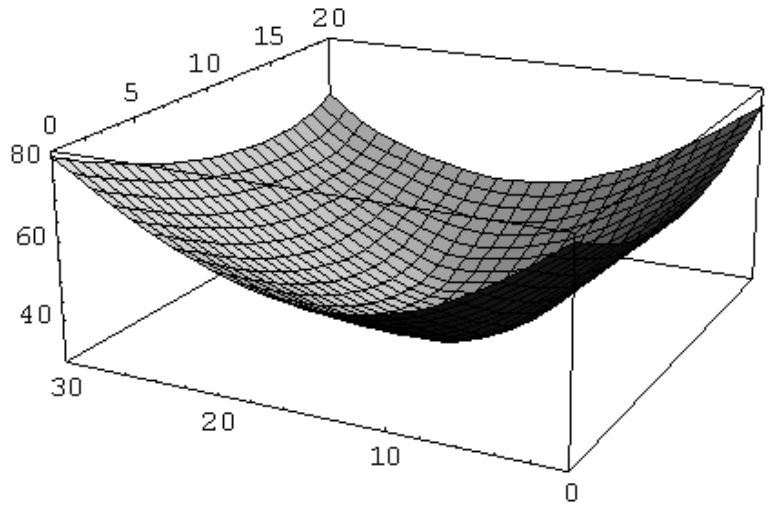


Figura 1: Gráfica de la función $f(r, s)$ sobre el rectángulo $(0, 20) \times (0, 30)$.

Como una segunda ilustración, la función f_2 que se define a continuación, usa la métrica L_2. Su gráfica tridimensional aparece abajo (Figura 2). Para determinar su óptimo (mínimo), ver las dos gráficas bidimensionales abajo de la superficie (Figura 3 y Figura 4). En términos estadísticos, $f_2(r, s)$ corresponde al valor esperado del cuadrado de la distancia del punto de coordenadas (r, s) a uno de los puntos (7,11), (9,8), (13,17), (18,28), cada uno de los cuales se selecciona con probabilidad de 1/5. Este ejemplo es bivariado discreto.

Con estas dos gráficas, es claro que el óptimo ocurre en (11.75, 16). Lo cual es consistente con la generalización del conocido resultado univariado, para ver ahora que el *centroide* de la distribución produce tal mínimo. En efecto, Media[X]=11.75, Media[Y] = 16. Los cálculos pueden realizarse cómodamente por medio de las instrucciones en *Mathematica* que siguen, y sus respectivas salidas al ejecutarlas.

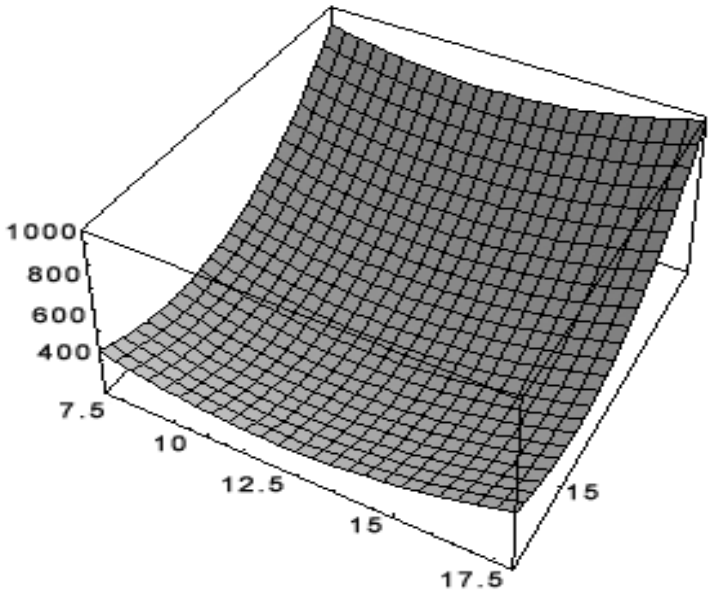


Figura 2: Gráfica de la función $f_2(r, s)$ sobre $(7, 18) \times (11, 28)$.

Mean[7,9,13,18.] 11.75 Mean[11,8,17,28.] 16.00

También puede corroborarse que el mínimo ocurre en (11.75, 16) mediante el análisis de las curvas de nivel de la superficie. Ver gráfica a continuación (Figura 5). Entre más obscuro es el tono de gris, más bajo es el valor de la función.

Plot[f2[r,16],{r,10.7,12.8}] Plot[f2[11.75,s],{s,14,18}]

ContourPlot[f2[r,s],{r,10,14},{s,14,18}]

La función $f_{20}(r, s)$ que se especifica enseguida mediante *Mathematica*, está utilizando la métrica L_{20} para su definición. Para su descripción gráfica, ver Figura 6 y Figura 7.

f20[r_,s_]:=((7-r)^20 +(11-s)^20)^(1/20)+((9-r)^20+(8-s)^20)^(1/20)+
((13-r)^20+(17-s)^20)^(1/20)+((18-r)^20+(28-s)^20)^(1/20)

Plot3D[f20[r,s],{r,0,20},{s,0,30},
PlotPoints → 60]

ContourPlot[f20[r,s],{r,0,20},{s,0,30},
PlotPoints → 100]

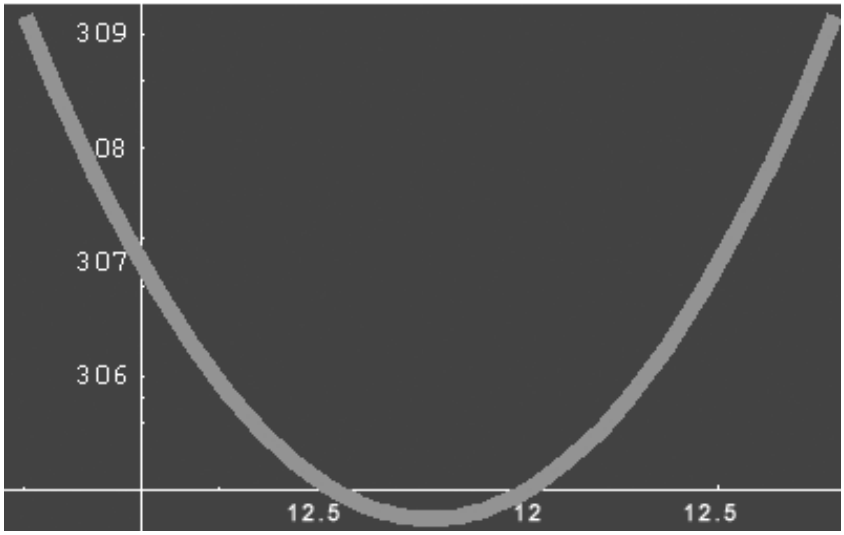


Figura 3: Gráfica de la función de una variable $f_2(r, 16)$ sobre (10.7,12.8).

A continuación se define otra función, g_{20}(r,s) en la que interviene la métrica L_{20}. Las gráficas correspondientes son las que aparecen en las figuras 8 y 9.

```
g20[r_,s_]:=((7-r)^20+(11-s)^20+(9-r)^20+(8-s)^20+
(13-r)^20+(17-s)^20+(18-r)^20+ (28-s)^20)^(1/20)
```

```
Plot3D[g20[r,s],{r,0,20},{s,0,30}]
```

```
ContourPlot[g20[r,s],{r,5,20},{s,15,22},
PlotPoints → 100]
```

A continuación se define la función $f_{abs}(r, s)$ mediante la utilización de la métrica L_1. El aspecto de la gráfica de $f_{abs}(r, s)$ aparece en la Figura 10 que sigue. Sus curvas de nivel tienen el aspecto que se presenta en la Figura 11, también a continuación. Siguen las instrucciones correspondientes en *Mathematica*.

```
fabs[r_,s_]:=Abs[7-r]+Abs[11-s]+Abs[9-r]+Abs[8-s]+
Abs[13-r]+Abs[17-s]+Abs[18-r]+Abs[28-s]
```

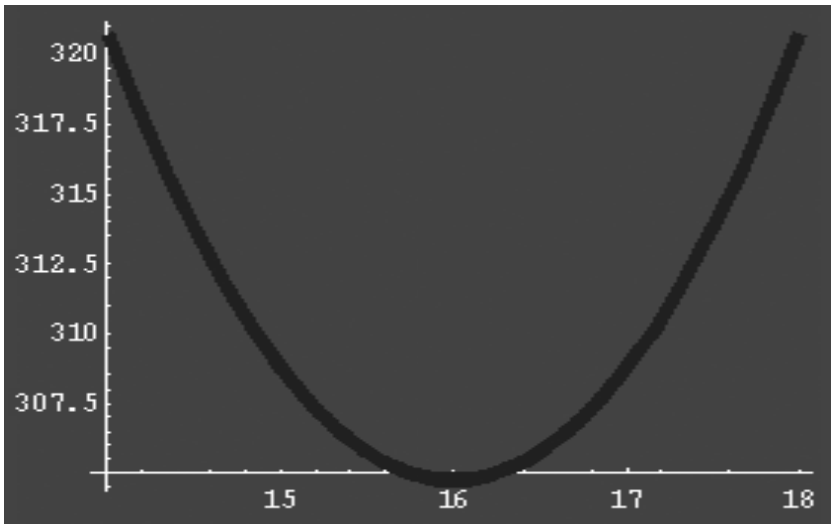


Figura 4: Gráfica de la función de una variable $f_2(11.75, s)$ sobre (14,18).

3. Comentarios finales

Aparte de la utilización de modelos estándar de probabilidad, se pueden definir modelos personalizados para su análisis. Un programa tal como *Mathematica* permite la exploración de modelos probabilistas univariados y multivariados, a niveles simbólico, gráfico y numérico. Para el presente trabajo, se consideraron distintas métricas en espacios de una, dos, y más dimensiones. La métrica del supremo puede aproximarse con L_{20}, por ejemplo. Algunos de nuestros hallazgos corresponden a resultados conocidos, en tanto que otros son nuevos.

Referencias

Blachman, N. 1992. *Mathematica: A Practical Approach.*. Prentice Hall, Inc.

Morrison, D.F. 1990. *Multivariate Statistical Methods*. Third Edition.McGraw-Hill.

Derman, C., Gleser, L. y Olkin I. 1973. *A Guide to Probability Theory and Application*. Holt, Rinehart and Winston, Inc.

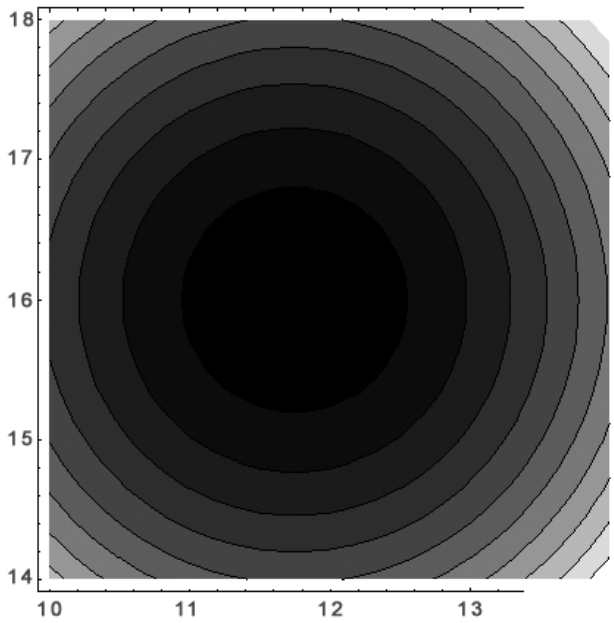


Figura 5: Curvas de nivel de la función $f_2(r, s)$ sobre el rectángulo $(10, 14) \times (14, 18)$

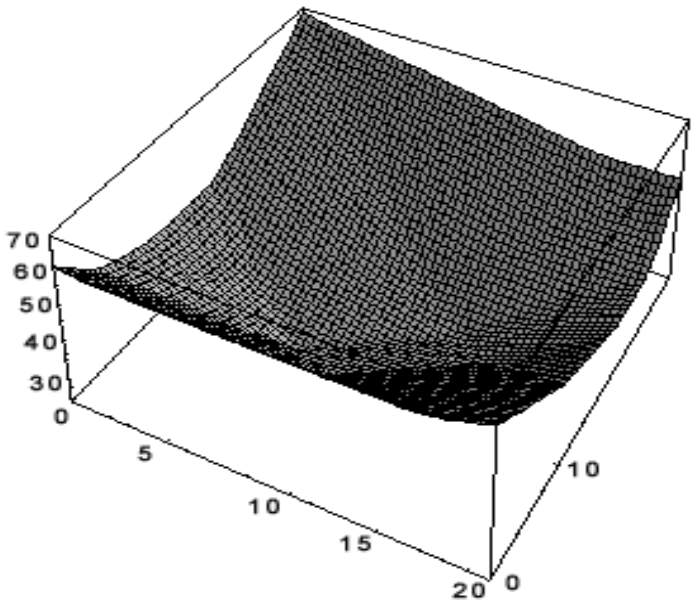


Figura 6: Superficie correspondiente a la función $f_{20}(r, s)$.

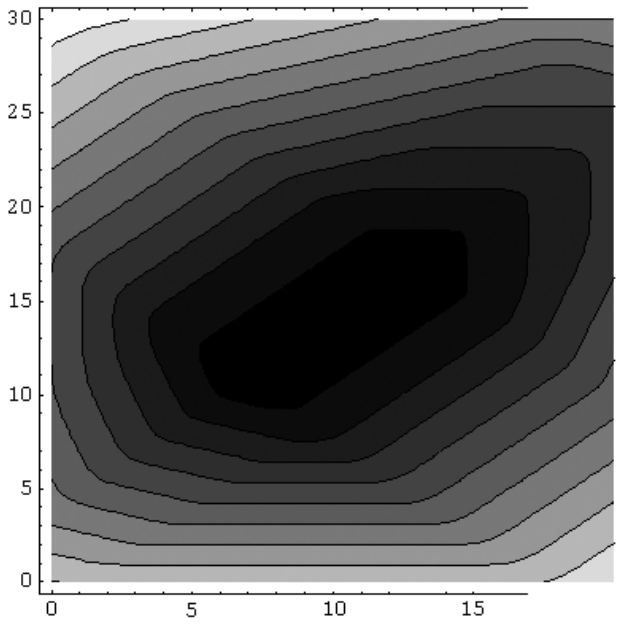


Figura 7: Curvas de nivel correspondientes a la función $f_{20}(r, s)$.

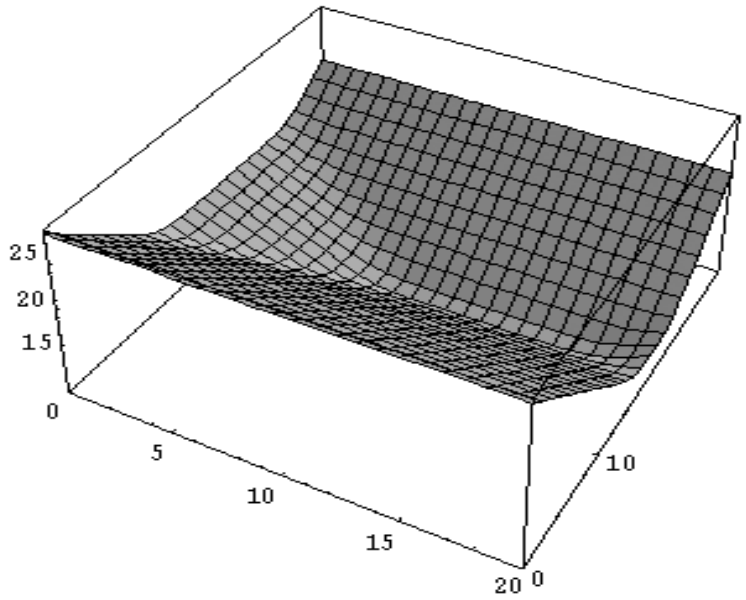


Figura 8: Superficie correspondiente a la función $f_{20}(r, s)$.

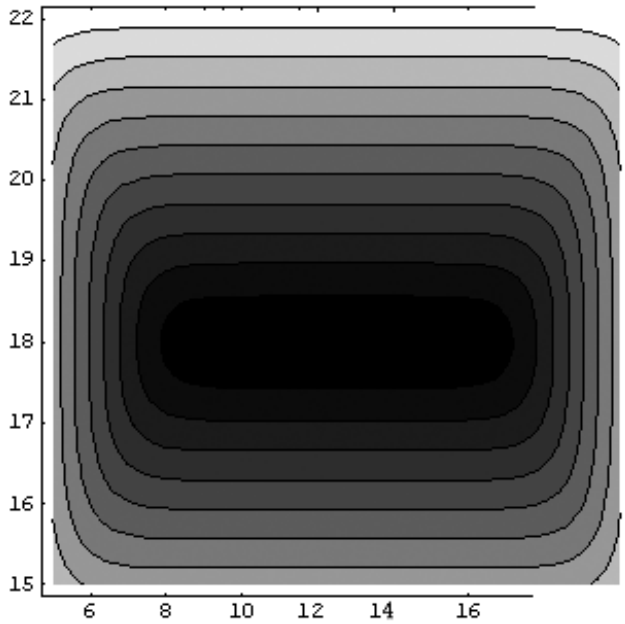


Figura 9: Curvas de nivel correspondientes a la función $f_{20}(r, s)$.

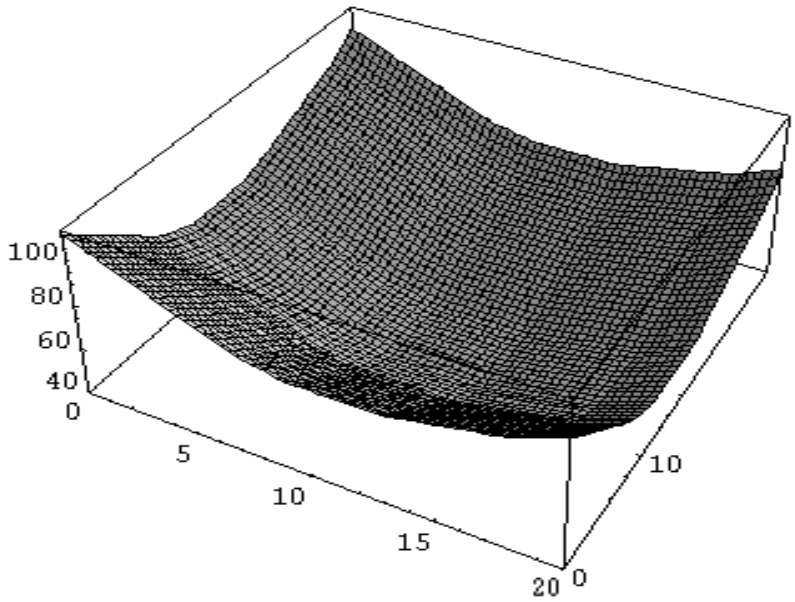


Figura 10: $f_{abs}(r, s)$ sobre $(0, 20) \times (0, 30)$.

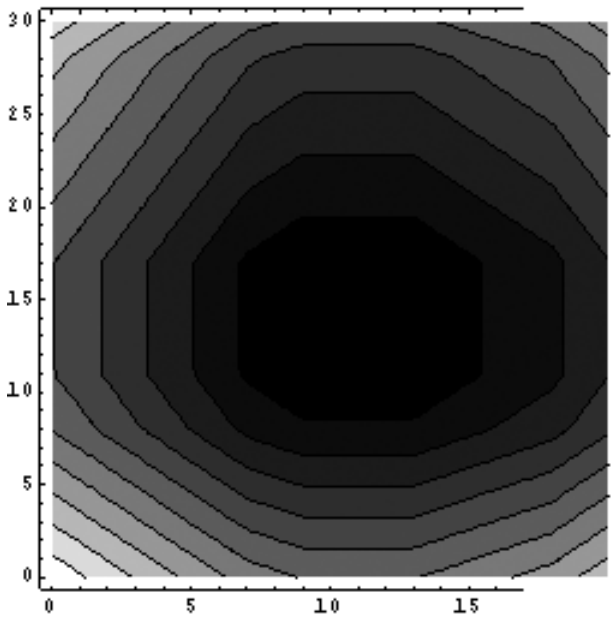


Figura 11: Curvas de nivel de $f_{abs}(r, s)$.

Memoria FONE XXIII (2009), Sección II, 49–53

Estimación del tamaño de una población de difícil detección en el muestreo por seguimiento de nominaciones y probabilidades de nominación heterogéneas*

Martín H. Félix Medina[a], Pedro E. Monjardin
Escuela de Ciencias Físico–Matemáticas – Universidad Autónoma de Sinaloa

1. Introducción

El Muestreo por Seguimiento de Nominaciones (denominado en Inglés como Link-tracing sampling o Snowball sampling) es un método que se ha propuesto para muestrear poblaciones de difícil detección, tales como poblaciones de drogadictos, niños de la calle, trabajadoras sexuales, etc. En este método se selecciona una muestra inicial de miembros de la población de interés, y se les pide a las personas que fueron seleccionadas que nominen a otros miembros de la población objetivo. A las personas que fueron nominadas se les puede pedir que nominen a otras personas, y el proceso de nominación puede continuar de esta manera hasta que se satisfaga alguna regla de terminación del muestreo.

Félix Medina y Thompson (2004) desarrollaron una variante de este tipo de muestreo y propusieron estimadores máximo verosímiles (EMV) del tamaño poblacional derivados bajo el supuesto de que las probabilidades de nominación no dependen de los individuos nominados, es decir, que son homogéneas. Sin embargo, en la práctica es muy común encontrar situaciones donde las probabilidades de nominación dependen de los individuos nominados, es decir, son heterogéneas, ya que individuos con mucha actividad social tienen mayores probabilidades de ser nominados que aquellos con poca actividad social. En este trabajo se

*Trabajo realizado con apoyos parciales de los proyectos PIFI-2005-25-06 de la SEP y PROFAPI 2008/054 de la UAS.

[a]mhfelix@uas.uasnet.mx

extiende el de Félix Medina y Thompson (2004) al caso de probabilidades de nominación heterogéneas y se proponen EMV's del tamaño poblacional. El procedimiento que se sigue es el usado por Coull y Agresti (1999) en el contexto de muestreo por captura-recaptura.

2. Diseño muestral, notación y modelos probabilísticos

El diseño muestral que consideraremos en este trabajo es el propuesto por Félix Medina y Thompson (2004). Así, supondremos que una parte U_1 de la población de interés U está cubierta por un marco muestral de N sitios $A_1, \ldots, A_N$, tales como parques, hospitales o cruceros de calles. De este marco se selecciona una muestra aleatoria simple sin reemplazo $S_0 = \{A_1, \ldots, A_n\}$ de n sitios, y a las personas de la población de interés que pertenecen al sitio seleccionado se les pide que nominen a otros miembros de la población. Como convención, diremos que una persona es nominada por un sitio si cualquiera de los miembros de ese sitio la nomina.

Denotaremos por τ el tamaño de U, por τ_1 el de U_1, por $\tau_2 = \tau - \tau_1$ el de $U_2 = U - U_1$, y por M_i el número de personas en A_i. Obsérvese que $\tau_1 = \sum_{i=1}^{N} M_i$ y que $M = \sum_{i=1}^{n} M_i$ es el número de individuos en S_0. Los conjuntos de variables $\{X_{ij}^{(1)}\}$ y $\{X_{ij}^{(2)}\}$ indicarán el proceso de nominación. Así, $X_{ij}^{(1)} = 1$ si la persona $u_j \in U_1 - A_i$ es nominada por el sitio A_i, y $X_{ij}^{(1)} = 0$ en otro caso. Similarmente, $X_{ij}^{(2)} = 1$ si la persona $u_j \in U_2$ es nominada por el sitio A_i, y $X_{ij}^{(2)} = 0$ en otro caso.

Al igual que en Félix-Medina y Thompson (2004) supondremos que las M_i's son variables aleatorias independientes con distribución Poisson con media λ_1. Cabe aclarar que aunque éste es un supuesto simplificador, experiencias previas con estimadores similares nos han mostrado que estos son robustos a desviaciones de este supuesto. Lo anterior implica que la distribución condicional conjunta de $(M_1, \ldots, M_n, \tau_1 - M)$, dado que $\sum_{i=1}^{N} M_i = \tau_1$, es multinomial con parámetro de tamaño τ_1 y vector de probabilidades $(1/N, \ldots, 1/N, 1-n/N)$. Asímismo, supondremos que dado M_i, la distribución condicional de $X_{ij}^{(k)}$ es Bernoulli con probabilidad $p_{ij}^{(k)} = \Pr[X_{ij}^{(k)} = 1 | M_i] = \exp(\alpha_i^{(k)} + \beta_j^{(k)})/[1 + \exp(\alpha_i^{(k)} + \beta_j^{(k)})]$, $i = 1, \ldots, n$; $j = 1, \ldots, \tau_k$, con $u_j \notin A_i$, y $k = 1, 2$. Este modelo se conoce como modelo Rash. El parámetro $\alpha_i^{(k)}$ es el efecto del potencial que tiene el sitio A_i de nominar individuos en U_k y $\beta_j^{(k)}$ es el efecto de suceptibilidad que tiene el individuo $u_j \in U_k - A_i$ de ser nominado.

3. Función de verosimilitud

Sea $\Omega = \{1, \ldots, n\}$, luego, la probabilidad de que el individuo $u_j \in U_k - S_0$ sea nominado solamente por cada uno de los sitios A_i con i en un subconjunto específico $\omega \subset \Omega$, $\omega \neq \emptyset$, es $\Pr_\omega(X_{1j}^{(k)} = x_{\omega 1}, \ldots, X_{nj}^{(k)} = x_{\omega n}) = \prod_{i=1}^{n} \exp[x_{\omega i}(\alpha_i^{(k)} + \beta_j^{(k)})]/[1 + \exp(\alpha_i^{(k)} + \beta_j^{(k)})]$, donde $x_{\omega i} = 1$ si $i \in \omega$ y $x_{\omega i} = 0$ en otro caso.

Al igual que en Coull y Agresti (1999) supondremos que los $\beta_j^{(k)}$'s son efectos aleatorios con distribución normal con media cero y varianza σ_k^2 desconocida. Así, la probabilidad de que un individuo en $U_k - S_0$ seleccionado al azar sea nominado sólo por los sitios A_i's con $i \in \omega$ es $\pi_\omega^{(k)} = \int \prod_{i=1}^{n} [\exp[x_{\omega i}(\alpha_i^{(k)} + \sigma_k z)]/[1 + \exp(\alpha_i^{(k)} + \sigma_k z)]\phi(z)dz$, donde $\phi(z)$ representa la función de densidad normal estándar.

La función de verosimilitud la podemos descomponer en varios factores. Uno es el factor $L_{Mult}(\tau_1)$ resultante del proceso de selección de la muestra inicial, el cual está dado por la distribución multinomial de $(M_1, \ldots, M_n, \tau_1 - M)$. Dos factores $L(\tau_1, \alpha^{(1)}, \sigma_1)$ y $L(\tau_2, \alpha^{(2)}, \sigma_2)$, donde $\alpha^{(k)} = (\alpha_1^{(k)}, \ldots, \alpha_n^{(k)})$, $k = 1, 2$, son resultado del proceso de nominación de los individuos en $U_1 - S_0$ y U_2 los cuales están dados por las distribuciones multinomiales de los vectores de variables $(\{R_\omega^{(1)}\}_{\omega \in \Omega}, \tau_1 - M - R_1)$ y $(\{R_\omega^{(2)}\}_{\omega \in \Omega}, \tau_2 - R_2)$, donde $\{R_\omega^{(1)}\}_{\omega \in \Omega}$ y $\{R_\omega^{(2)}\}_{\omega \in \Omega}$, $\omega \neq \emptyset$, son los conjuntos de números de individuos en $U_1 - S_0$ y U_2 nominados únicamente por los sitios A_i's con $i \in \omega \subset \Omega$, R_1 y R_2 son los números de distintos individuos en $U_1 - S_0$ y U_2 que fueron nominados y $\tau_1 - M - R_1$ y $\tau_2 - R_2$ son los números de individuos en $U_1 - S_0$ y U_2 que no fueron nominados. Estas distribuciones multinomiales tienen parámetros de tamaño $\tau_1 - M$ y τ_2, y vectores de probabilidades $(\{\pi_\omega^{(1)}\}_{\omega \in \Omega}, \pi_\emptyset^{(1)})$ y $(\{\pi_\omega^{(2)}\}_{\omega \in \Omega}, \pi_\emptyset^{(2)})$, $\omega \neq \emptyset$, donde $\pi_\emptyset^{(k)}$ es la probabilidad de que un individuo en $U_k - S_0$, $k = 1, 2$, no sea nominado y está dada por la expresión para $\pi_\omega^{(k)}$ con $x_{\omega i} = 0$, $i = 1, \ldots, n$. El último factor $L(\alpha^{(1)}, \sigma_1)$ es el resultado del proceso de nominación de los individuos en S_0. Este factor se descompone en n factores $L_{A_i}(\{\alpha_j^{(1)}\}_{j \neq i}, \sigma_1)$, $i = 1, \ldots, n$, que son resultado de los procesos de nominación de los individuos en los sitios $A_1, \ldots, A_n$ en S_0. El factor L_{A_i} está dado por la distribución multinomial del vector de variables $(\{R_\omega^{(A_i)}\}_{\omega \in \Omega - \{i\}}, M_i - R^{(A_i)})$, donde $\{R_\omega^{(A_i)}\}_{\omega \in \Omega - \{i\}}$ son los conjuntos de números de individuos en A_i nominados únicamente por los sitios A_j, $j \neq i$, con $j \in \omega \subset \Omega - \{i\}$, $R^{(A_i)}$ es el número de distintos individuos en A_i que fueron nominados y $M_i - R^{(A_i)}$ es el número de individuos en A_i que no fueron nominados. Esta distribución multinomial tiene parámetro de tamaño M_i y vector de probabilidades

$(\{\pi_\omega^{(A_i)}\}_{\omega\in\Omega-\{i\}}, \pi_\emptyset^{(A_i)})$, $\omega \neq \emptyset$, donde $\pi_\omega^{(A_i)}$ y $\pi_\emptyset^{(A_i)}$ están dadas por las respectivas expresiones para $\pi_\omega^{(1)}$ y $\pi_\emptyset^{(1)}$ pero sin el i-ésimo factor.

Cabe aclarar que, como en Coull y Agresti (1999), en el proceso de maximización de la función de verosimilitud las probabilidades $\pi_\omega^{(k)}$ y $\pi_\omega^{(A_i)}$ se calcularán mediante el método de cuadratura Gaussiana, esto es, serán aproximadas por $\tilde{\pi}_\omega^{(k)} = \sum_{t=1}^{q}\prod_{i=1}^{n}\exp[x_{\omega i}(\alpha_i^{(k)} + \sigma_k z_t)]/[1+\exp(\alpha_i^{(k)}+\sigma_k z_t)]\nu_t$, donde $\{z_t\}$ y $\{\nu_t\}$ se obtienen de tablas.

4. Estimadores máximo verosímiles

Al maximizar, numéricamente, la verosimilitud con respecto a τ_1, τ_2, $\alpha^{(1)}$, $\alpha^{(2)}$, σ_1 y σ_2 se obtienen las estimaciones máximo verosímiles incondicionales de estos parámetros. La estimación máximo verosímil incondicional de τ es la suma de las estimaciones de τ_1 y τ_2.

Un enfoque más simple para estimar los τ_k's es el siguiente enfoque de Sanathan (1972) basado en estimación máximo verosímil condicional. Se factorizan las distribuciones multinomiales de $(\{R_\omega^{(1)}\}_{\omega\in\Omega}, \tau_1 - M - R_1)$ y $(\{R_\omega^{(2)}\}_{\omega\in\Omega}, \tau_2 - R_2)$ como sigue: $f(\{R_\omega^{(1)}\}_{\omega\in\Omega}|M,\tau_1,\alpha^{(1)},\sigma_1) = f(\{R_\omega^{(1)}\}_{\omega\in\Omega}|M,R_1,\alpha^{(1)},\sigma_1)f(R_1|M,\tau_1,\alpha^{(1)},\sigma_1)$ y $f(\{R_\omega^{(2)}\}_{\omega\in\Omega}|M,\tau_2,\alpha^{(2)},\sigma_2) = f(\{R_\omega^{(2)}\}_{\omega\in\Omega}|M,R_2,\alpha^{(2)},\sigma_2)f(R_2|M,\tau_2,\alpha^{(2)},\sigma_2)$, donde en cada caso el primer factor es una distribución multinomial con parámetro de tamaño R_k y vector de probabilidades $(\{\pi_\omega^{(k)}/[1-\pi_\emptyset^{(k)}]\}_{\omega\in\Omega})$, $\omega\neq\emptyset$, que no depende de τ_k, $k=1,2$, y los segundos factores son distribuciones binomiales con parámetros de tamaño $\tau_1 - M$ y τ_2 y probabilidades $1-\pi_\emptyset^{(1)}$ y $1-\pi_\emptyset^{(2)}$. Luego, se maximiza numéricamente la parte de la función de verosimilitud compuesta por los factores que no contienen a los τ_k's, es decir que sólo contienen a $\alpha^{(k)}$ y σ_k. Se insertan las estimaciones de estos parámetros en la parte de la verosimilitud compuesta por los factores que si contienen a los τ_k's y se maximiza esta parte en términos de estos parámetros.

5. Estudio Monte Carlo

Se generó una población de $N = 100$ sitios A_i's con tamaños M_i's obtenidos de una distribución Poisson con media 7.2. Así, se obtuvo que $\tau_1 = 725$ y se fijó $\tau_2 = 200$ por lo que $\tau = 925$. Los valores de $\alpha_i^{(k)}$ se generaron de una distribución normal con media $-35/(M_i + 0.1)$ y varianza 9, $k = 1, 2$, y los de $\beta_j^{(1)}$ y $\beta_j^{(2)}$ de normales estándar. De la población de $N = 100$ sitios se obtuvo una muestra de $n = 10$ sitios y se estimaron τ_1, τ_2 y τ mediante los EMV

Esti-mador	Estimación promedio	Sesgo relativo	Raíz cuad. de ECM relativo	Esti-mador	Estimación promedio	Sesgo relativo	Raíz cuad. de ECM relativo
$\hat{\tau}_1$	744.8	0.03	0.11	$\tilde{\tau}_1$	451.1	-0.38	0.45
$\hat{\tau}_2$	159.9	-0.20	0.34	$\tilde{\tau}_2$	1289.8	5.45	16.3
$\hat{\tau}$	904.6	-0.02	0.10	$\tilde{\tau}$	1740.9	0.88	3.44

Notas: $\bar{m} = 73.0$, $\bar{r}_1 = 378.1$ y $\bar{r}_2 = 99.9$. Resultados basados en 200 muestras.

Tabla 1: Sesgos relativos y raíces cuadradas de errores cuadráticos medios relativos de los EMV condicionales $\hat{\tau}_1$, $\hat{\tau}_2$ y $\hat{\tau}$ y de los EMV $\tilde{\tau}_1$, $\tilde{\tau}_2$ y $\tilde{\tau}$ derivados bajo el supuesto de probabilidades heterogéneas y homogéneas, respectivamente.

condicionales $\hat{\tau}_1$, $\hat{\tau}_2$ y $\hat{\tau}$ propuestos en este trabajo y mediante los EMV $\tilde{\tau}_1$, $\tilde{\tau}_2$ y $\tilde{\tau}$ propuestos por Félix-Medina y Thompson (2004) y derivados bajo el supuesto de homogeneidad de las probabilidades de nominación. Los resultados de la simulación, basados en 200 muestras, son los siguientes:

6. Conclusiones

De acuerdo con los resultados del estudio Monte Carlo tenemos que los estimadores propuestos en este trabajo muestran desempeños aceptables. Por otro lado, si no se toma en cuenta la heterogeneidad de las probabilidades de nominación se presentan problemas de sesgo e inestabilidad en las estimaciones.

Referencias

Coull, B.A. and Agresti, A. 1999. The use of mixed logit models to reflect heterogeneity in capture-recapture studies. *Biometrics*, **55**, 294-301.

Félix-Medina, M.H., and Thompson, S.K. 2004. Combining cluster sampling and link-tracing sampling to estimate the size of hidden populations. *Journal of Official Statistics*, **20**, 19-38.

Sanathan, L. 1972. Estimating the size of a multinomial population. *Annals of Mathematical Statistics*, **43**, 142-152.

Memoria FONE XXIII (2009), Sección II, 55–59

Biplot versus coodernadas paralelas

Purificación Galindo Villardón[a], Purificación Vicente Galindo
Universidad de Salamanca – España

Carlomagno Araya Alpízar
Universidad de Costa Rica

1. Métodos biplot

Un **Biplot** es una representación gráfica de datos multivariantes. El objetivo del método Biplot es realizar una representación plana de la matriz $\mathbf{X}_{np}$ por medio de unos marcadores $g_1, \cdots, g_n$ para sus filas y $h_1, \cdots, h_p$ para sus columnas, elegidas de tal forma que el producto interno $g_i{}'h_j$ represente al elemento x_{ij} de la matriz $\mathbf{X}$ (Gabriel, 1971).

Si los g_i para $(i = 1, ..., n)$ son las filas de la matriz $\mathbf{G}$ y los h_j para $(j = 1, ..., p)$ las filas de una matriz $\mathbf{H}$, el producto de estas matrices representa a la matriz de partida, de la forma:

$$\mathbf{X} = \mathbf{G}\mathbf{H}' \tag{1}$$

$$\begin{bmatrix} x_{11} & x_{12} & x_{13} \\ x_{21} & x_{22} & x_{23} \\ x_{31} & x_{32} & x_{33} \\ x_{41} & x_{42} & x_{43} \end{bmatrix} = \begin{bmatrix} g_{11} & g_{12} \\ g_{21} & g_{22} \\ g_{31} & g_{32} \\ g_{41} & g_{42} \end{bmatrix} \begin{bmatrix} h_{11} & h_{12} & h_{13} \\ h_{21} & h_{22} & h_{23} \end{bmatrix}$$

El elemento x_{ij} de la matriz $\mathbf{X}$ se expresa como un producto de una fila de G por una columna de H. Por ejemplo: $x_{11} = g_{11}h_{11} + g_{12}h_{21}$ y $x_{41} = g_{41}h_{11} + g_{42}h_{21}$. Cada elemento de la matriz de partida puede expresarse como un producto de una fila de G por una columna de H. Se parte de una **descomposición en valores singulares** de la matriz $\mathrm{X_{nxp}}$ de rango **p**,

$$X = U\ \Sigma\ V^T \tag{2}$$

[a] pgalindo@usal.es

donde: **U** es una matriz de dimensión **(nxp)** cuyos vectores columna son ortonormales y vectores propios de XX^T donde XX^T tiene dimensión (nxn), V es una matriz ortogonal de dimensión (pxp) cuyos vectores columna son vectores propios de X^TX; donde X^TX tiene dimensión (pxp) y Σ es una matriz diagonal de dimensión (pxp) que contiene los valores singulares de X, ordenados de mayor a menor. Los valores singulares coinciden con los valores propios de X^TX y XX^T.

Debido a que estamos aproximando una matriz de $\mathbf{X}_{np}$, de rango **r**, por una matriz de rango menor, $\mathbf{X}_q$, estamos "perdiendo información", ya que la representación Biplot es aproximada. Una forma de medir esta pérdida es a través de la **Calidad de Representación** de los puntos fila y columna, cuanto más cercano este a cien, mayor cantidad de información está siendo recogida por la representación Biplot.

Entre los métodos Biplot, nos encontramos el GH-Biplot, JK-Biplot y HJ-Biplot. El **JK-Biplot**, es una representación simultánea de individuos y variables, donde los individuos tienen máxima calidad de representación, razón por la cual se conoce RMP-Biplot (Row Metric Preserving). A este Biplot Gabriel lo denominó JK-Biplot porque utilizó **J** para denotar la matriz de marcadores fila y **K** para la matriz de marcadores columna. El **GH-Biplot**, es una representación simultánea de individuos y variables, donde las variables tienen máxima calidad de representación. A este Biplot Gabriel lo denominó GH-Biplot porque utilizó **G** para denotar la matriz de marcadores fila y **H** para la matriz de marcadores columna. El producto escalar de las columnas de **X**, coincide con el producto escalar de los marcadores columna, de ahí el hecho de que este Biplot se denomine CMP-Biplot (Column Metric Preserving) ya que preserva la métrica euclídea usual entre las columnas de **X** obteniéndose una alta calidad de representación para éstas.

El **HJ-Biplot** a diferencia de los anteriores fue propuesto por Galindo, (1985, 1986) es una representación gráfica multivariante de marcadores fila y columna, elegidos de tal forma que puedan superponerse en el mismo sistema de referencia con máxima calidad de representación. Este Biplot es muy útil en la interpretación simultánea de relaciones entre filas y columnas, no siendo su objetivo principal la aproximación de los elementos de la matriz de datos, como es el caso de los Biplot definidos por Gabriel. Los elementos de la matriz **X** están centrados por filas y columnas, por lo que la métrica introducida en el espacio de las filas es equivalente a la inversa de la matriz de covarianzas entre variables, mientras

que en el espacio de las columnas la métrica es equivalente a la inversa de la matriz de dispersión entre individuos. Dado de que en el **HJ-Biplot** se puede hacer una representación simultánea de filas y columnas se lo denomina también **RCMP-Biplot (Row Column Metric Preserving)**.

El **HJ- Biplot** permite interpretar las posiciones de las filas, de las columnas y las relaciones fila-columna a través de los factores (ejes), como en el caso del Análisis Factorial de Correspondencias (Benzecri, 1973; Greenacre, 1984) teniendo además la ventaja de que un análisis Biplot puede llevarse a cabo sobre a cualquier tipo de datos.

2. Coordenadas paralelas

Las **Coordenadas Paralelas** (*Coords*‖) fueron propuestas por Alfred Inselberg (1992). Las *Coords*‖ es un sistema de visualización que permite representar n-dimensiones en un sistema bidimensional. En este sistema, cada eje vertical (ordenada) representa un atributo (dimensión) que puede ser continuo o categórico. Cada uno de los ejes verticales de un sistema de *Coords*‖ puede tener su propia escala o definirse todos con una sola escala, la primera forma nos permite la visualización de híper-superficies y el análisis del funcionamiento del conjunto de datos, con la segunda podemos hacer un análisis de las relaciones entre las variables.

Uniendo con líneas los ejes, podemos simbolizar los puntos en n-dimensiones. Asimismo, un punto en un espacio n-dimensional es transformado en una línea poligonal a través de n ejes paralelos como $n-1$ segmentos de línea. De tal forma, el vector $\mathbf{x} = [x_1, x_2, \ldots, x_n]$ es representado por medio de x_1 en la coordenada 1, x_2 en la coordenada 2 y así sucesivamente, hasta la x_n en al coordenada n. A partir de la representación resultante, podemos sacar conclusiones al respecto, por ejemplo sobre la relación entre las variables.

El orden de las *Coords*‖ es una condición que puede afectar significativamente la expresividad del gráfico, variando el orden es posible abreviar el problema sin la reducción del contenido o de la modificación de los datos de alguna manera. También las correlaciones entre las variables (o dimensiones) pueden ser descubiertas concentrándose en las intersecciones de las polilíneas, al detectar grupos de observaciones con pendientes comunes en las líneas de conexión inter-variables, poniendo de relieve un determinado tipo de correlación entre dichas variables (positiva, negativa o nula).

Las *Coords*‖ resultan útiles para captar agrupamientos ("clústeres"). Las polilíneas que tienden a estar cercanas constituirán un grupo a diferencia de aquellas que se separan y cuando hay líneas que no pertenecen a ningún grupo (fuera de los patrones) pueden considerarse como valores extremos. El descubrimiento de grupos o racimos de polilíneas diferenciadas del resto se consigue cambiando los órdenes de las dimensiones, para procurar que las relaciones de los datos puedan ser visualizadas. Es recomendable estandarizar las variables para poder compáralas y permitir un mejor descubrimiento del patrón anormal.

3. Biplot versus coodernadas paralelas

Se pretende con este apartado, estudiar las semejanzas y diferencias entre ambos métodos de análisis de datos multivariantes. De algún modo, se intenta establecer que ambas técnicas multivariantes integradas maximizan el éxito en la interpretación de los resultados.

En los métodos Biplot la variabilidad de las variables está determinada por longitud de los vectores, mientras que en *Coords*‖ se visualiza con la dispersión de las polilíneas en los ejes.

Se observa en la Figura 1 que las variables **V4** y **V5** tienen una alta correlación positiva, tal que el ángulo entre los dos vectores es muy pequeño (en los Biplots), esto en *Coords*‖ se visualiza como líneas entre los ejes asociados a **V4** y **V5** que tienden a ser horizontales. Una correlación negativa en *Coords*‖ es por ejemplo, entre las variables **V1** y **V2**, se observa como líneas que se intersecan.

En la Figura 2 puede verse el perfil del grupo formado por los individuos **A**, **F** y **K**. Se observa que poseen valores altos en la variable **V2**, y por lo contrario valores muy pequeños en comparación a los demás individuos en las restantes variables. Las *Coords*‖ nos proporcionan un instrumento para hacer un diagnóstico de las contribuciones de las individuos y variables a los ejes factoriales de la representación Biplot.

Referencias

Benzécri, Jean Paul. 2004. *L'Analyse des Données: L'analyse des correspondances*. SParis: Dunod.

Gabriel, Karl Ruben. 1971. *The Biplot Graphic Display of Matrices with Application to Principal Component Analysis*. *Biometrika*, **58**, 453-467.

Galindo, María Purificación. 1985. *Contribuciones a la Representación Simultánea de datos Multidimensionales.*Tesis doctoral. Universidad de Salamanca.

Galindo, María Purificación. 1986. *Una Alternativa de Representación Simultánea: HJ-biplot*. *Questió*, **10**, 13-23.

Greenacre, Michael John. 1984. *Theory and Applications of Correspondence Analysis*. London: Academic Press.

Inselberg, Alfred. 1992. *The Plane R2 with Coordinate Parallel*. Tel Aviv: Computer Science and Applied Mathematics Departments.

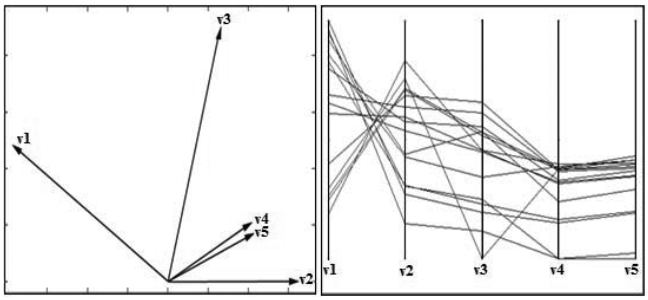


Figura 1: Diagnóstico de la correlación en los Biplots y *Coords*‖

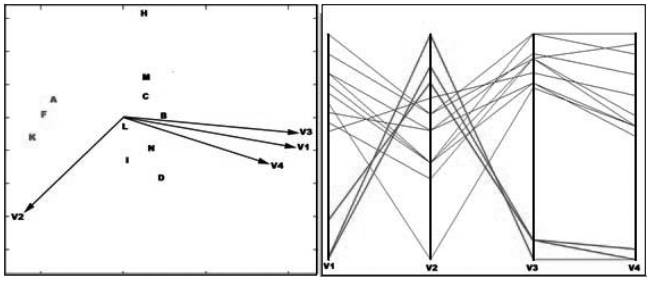


Figura 2: Diagnóstico de grupos en los Biplots y *Coords*‖

Memoria FONE XXIII (2009), Sección II, 61–67

Estudio computacional sobre aleatoriedad para sucesiones grandes en desarrollo decimal de algunos números

Agustín Jaime García Banda[a], Ismael Sosa Galindo
Facultad de Ciencias Administrativas y Sociales – UV

Luis Cruz-Kuri
Instituto de Ciencias Básicas – UV

1. Introducción

Desde la antigüedad números no racionales, tales como raíz cuadrada de dos y π, han ejercido una fascinación en los estudiosos de las matemáticas. Recientemente, con el advenimiento de facilidades de computo de alta velocidad y gran capacidad de almacenamiento, así como el desarrollo de algoritmos de gran eficiencia, se han podido obtener expansiones decimales de tales números a una cantidad gigantesca de cifras. Es conocido que el número π, cociente del perímetro de un círculo y su diámetro, es *trascendente*, en el sentido de no ser raíz de algún polinomio con coeficientes enteros. Otro número que también es trascendente es e, la base de los logaritmos naturales; la raíz cuadrada de 2, por otra parte, es irracional pero no es trascendente. Lo anterior ha dado lugar para los estudiosos de la probabilidad a investigaciones de frecuencias relativas con las que ocurren cada uno de los dígitos en el desarrollo decimal, o las frecuencias relativas con que aparecen pares y ternas de dígitos, con el propósito de explorar si se tienen los resultados esperados tanto de frecuencias como de independencia estocástica. Con la introducción de programas de cómputo matemático, tales como *Mathematica*, es relativamente fácil obtener en computadoras personales expansiones decimales de tales números a millones de cifras. En el presente trabajo, se introduce un algoritmo computacional, desarrollado por nosotros, el cual permite hacer aplicación de algunas

[a] jaimegarciabanda@yahoo.com

pruebas de aleatoriedad para las sucesiones correspondientes de los dígitos así generados; asimismo se discuten los hallazgos para algunos número selectos. Lo anterior nos ha remitido a la comprobación por medios propios acerca de la validez o no de resultados probabilistas como los arriba indicados. Esto fue posible para nosotros con el apoyo de un programa de cómputo matemático muy poderoso en conjunción con el desarrollo de programas propios, obteniéndose resultados interesantes, algunos de los cuales se presentan en este trabajo.

2. Utilización del programa *Mathematica*

El programa de cómputo *Mathematica* tiene tres niveles de procesamiento: el gráfico, el simbólico y el numérico. En el nivel numérico, las operaciones aritméticas pueden realizarse con una precisión arbitraria, lo cual es muy conveniente para la realización de experimentos en teoría de números y su correspondiente análisis estadístico. En nuestro caso, se ejecutaron los procesamientos con este programa de cómputo matemático utilizando una computadora personal de escritorio. Con una instrucción sencilla de *Mathematica*, se pudieron generar expansiones decimales a 10 millones de cifras para números algebraicos (tales como raíz cuadrada de dos) así como números trascendentes (tales como π, e base de logaritmos naturales y raíz cuadrada de dos elevado a la potencia raíz cuadrado de dos). Por ejemplo, para generar la expansión decimal de π a 10 millones de cifras, la instrucción en Mathematica es N[Pi, 10000000]; de manera análoga la instrucción para generar la expansión del número raíz cuadrada de dos elevado a la potencia raíz cuadrada de dos, a 10 millones de cifras también, la instrucción en Mathematica es N[Sqrt[2]^Sqrt[2],10000000]. Una vez que se obtuvieron números como los anteriores se construyeron los respectivos archivos en extensión txt, los cuales son susceptibles de ser analizados mediante programas propios desarrollados por nosotros. Los resultados de tales análisis se presentan en distintas partes del presente trabajo. Ver por ejemplo, la tabla 1 adelante para procesamientos de tipo estadístico. Los programas propios se desarrollaron en Java y algunos detalles también aparecen en el presente trabajo (ver sección 3).

3. Algoritmo implementado en Java

En esta sección se presenta la secuencia del algoritmo que tiene como objetivo leer números de un archivo con extensión txt (datos generados del programa *Mathematica*), y obtener una tabla con el promedio, variancia y desviación estándar de 100 números pares de 10,000,000 de datos.

1. Coloque separadores despues de cada dos dígitos en la parte decimal del número, así obtendra un vector de 5,000,000 de números enteros entre 00 y 99.

2. Subdivida el vector de números enteros de tamaño 5,000,000 en 50,000 vectores de tamaño cien, y obtenga las estadísticas básicas de estos 50,000 vectores.

4. Aspectos estadísticos

La tabla 1 que aparece adelante fue generada con un programa propio en el que se recorrió la sucesión de 10 millones de dígitos en la expansión del número π. Más específicamente, se obtuvo la distribución de frecuencias de pares consecutivos. Bajo la suposición de independencia en probabilidad y de distribución uniforme, las frecuencias esperadas deben ser de $5{,}000{,}000 \times 0.01 = 50{,}000$, lo cual coincide notablemente con las cantidades que aparecen en cada celda de la tabla 1 (e.g. la frecuencia observada para el par (9,7) es 49,999; para el par (4,0) es 50036). Resultados análogos a los de la tabla 1 se obtuvieron para el desarrollo, también a 10 millones de dígitos, de otros números trascendentes, tales como $\sqrt{2}^{\sqrt{2}}$, e = base de logaritmos naturales, etc.; sin embargo, por razones de espacio, para el presente reporte, se omiten los detalles.

También con programas propios se tomaron 100 enteros entre 0 y 99, formados cada uno de ellos con dos dígitos consecutivos en el desarrollo de π a 10 millones de cifras; esto permite obtener 50,000 números enteros entre 0 y 99; para cada uno de ellos se obtuvieron sus estadísticas de media aritmética, variancia y desviación estándar. Para cada uno de los 50,000 valores de dichas estadísticas se realizaron estudios de distribución de frecuencias. Inicialmente, se tomaron únicamente los primeros 200 valores de cada estadística, se procedió enseguida a realizar el mismo procesamiento para los primeros 1,000 valores, luego para los primeros 10,000 valores y finalmente para los 50,000. Todo lo anterior con la intención de ver el posible funcionamiento del Teorema Central del Límite (TCL). Por consideraciones

	0	1	2	3	4	5	6	7	8	9
0	49883	50263	49952	50030	50123	49699	49431	49855	50484	49993
1	50145	49760	50187	50170	50182	49660	49756	49706	50276	50192
2	49805	49970	50022	50092	50108	50166	49918	50214	49913	50166
3	50230	49461	49561	50220	50079	49769	49974	49892	49824	50415
4	50036	50056	50200	49890	50068	50097	50117	50044	49918	50198
5	49995	50178	49827	50144	49848	50351	50085	50400	50191	50099
6	49960	49842	50362	50057	50078	49758	49901	50019	50094	49931
7	50062	50089	50057	50170	50184	49779	50049	50012	49762	49785
8	49986	49647	49878	50006	49721	50042	50410	50116	49797	49647
9	49622	50033	49886	49760	50075	50025	49694	49999	50304	50108

Tabla 1: Distribución de frecuencias de pares de dígitos en el desarrollo decimal de π a 10 millones de cifras.

de espacio, solo se presentan tres gráficas, de las cuales las dos primeras corresponden a 200 y a 50,000 casos, respectivamente, para ilustración del TCL (ver figuras 1 y 2).

La Figura 3 ilustra la Ley de los Grandes Números únicamente para los primeros 200 casos. Por otra parte, para la serie de tiempo correspondiente, pero con todos los 50,000 casos, se calcularon las autocorrelaciones desde 0 hasta 16 unidades de desplazamiento entre números de 0 a 99 en la expansión decimal de π; de cumplirse con la hipótesis de independencia, cada autocorrelación entre dos de tales números, a cualquier distancia, debe ser nula, lo cual precisamente corresponde a la información que se presenta en la gráfica correspondiente (ver figura 4). Las gráficas correspondientes a diagramas de dispersión para pares de tales números también apoyan la hipótesis de independencia, aunque con menor detalle que el análisis de autocorrelaciones. Por consideraciones de espacio no se incluyen dichas gráficas en el presente trabajo, aunque es pertinente hacer notar el aspecto circular de cada enjambre de puntos, el cual esta más acentuado en el que consiste de 50,000 puntos en comparación con el que solamente consiste de 100 puntos, digamos.

5. Conclusiones

La utilización de un programa de cómputo matemático, tal como Mathematica, facilita notablemente la generación de sucesiones grandes en expansión decimal (o en otras bases) de números trascendentes, tales como π, e, raíz cuadrada de dos elevada a la potencia de

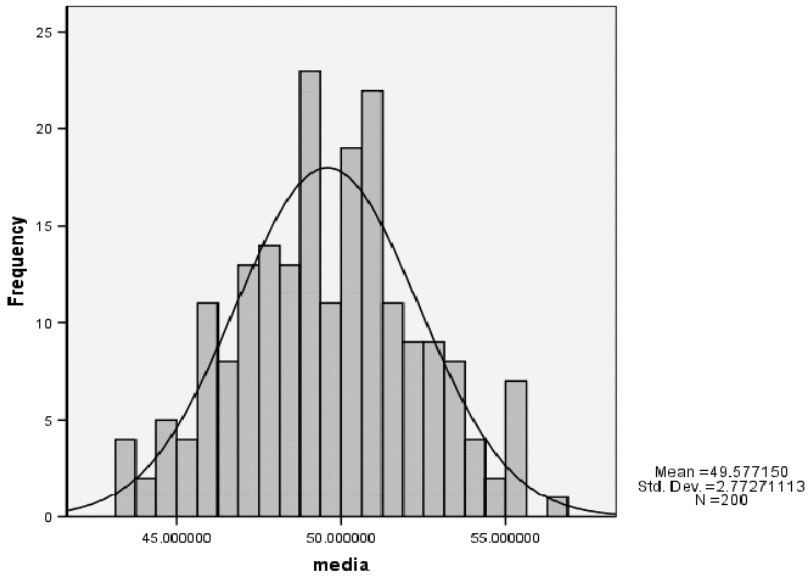


Figura 1: Distribución de frecuencias y estadísticas descriptivas para 200 promedios de números de dos cifras en la expansión decimal de π

raíz cuadrada de dos, etc. En nuestro caso, se generaron tales sucesiones a diez millones de dígitos para cada número; aunque fácilmente se puede hacer a expansiones mayores. Para los distintos análisis de tipo estadístico, se utilizaron programas tales como el *SPSS* y el *STATISTICA*, los cuales permiten procesamientos gráficos y numéricos. Finalmente, también se desarrollaron programas propios para explorar las sucesiones generadas. Lo obtenido hace ver que dichas sucesiones se comportan como si fueran generadas por mecanismos aleatorios.

Referencias

Blachman, N. 1992. *Mathematica: A Practical Approach.*. Prentice Hall, Inc.

Derman, C., Gleser, L. y Olkin I. 1973. *A Guide to Probability Theory and Application*. Holt, Rinehart and Winston, Inc.

Morrison, D.F. 1990. *Multivariate Statistical Methods*. Third Edition.McGraw-Hill.

Shawn, W. y Tigg, J. 1994. *Applied Mathematica - Getting Started, Getting It Done*. Addison Wesley. New York, USA.

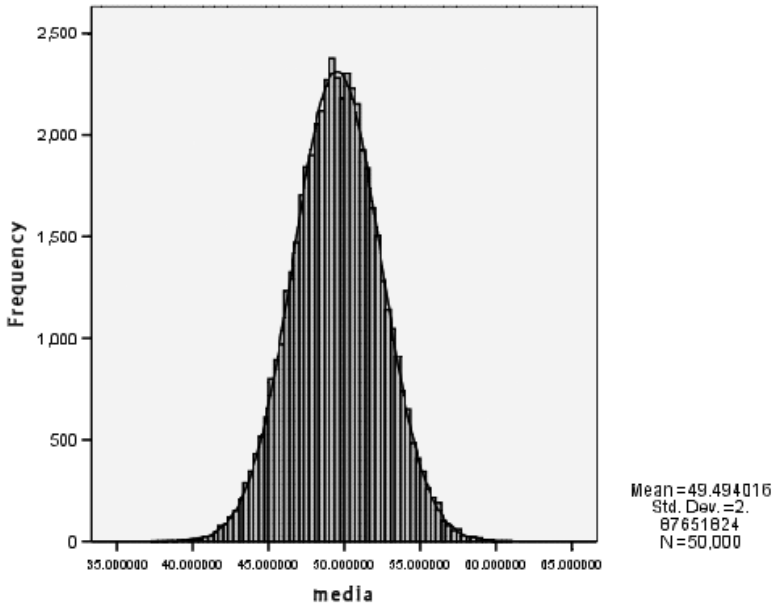


Figura 2: Distribución de frecuencias y estadísticas descriptivas para 50,000 promedios de números de dos cifras en la expansión decimal de π

Wolfram, S. 1999. *MATHEMATICA -A System for Doing Mathematics by Computer*. Fourth Edition. Addison-Wesley. U.S.A.

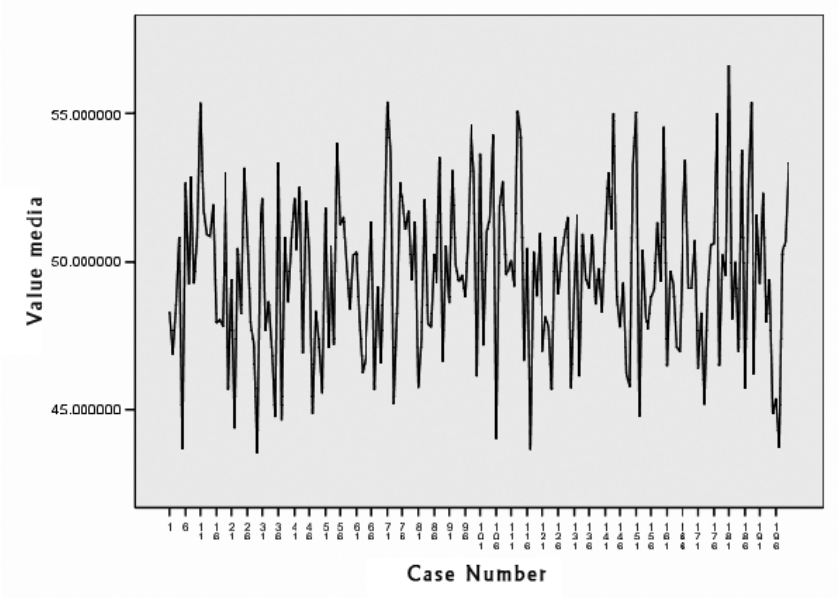


Figura 3: Serie de tiempo de los primeros 200 promedios de números de dos cifras en la expansión decimal de π. Nótese la tendencia alrededor de 49.5

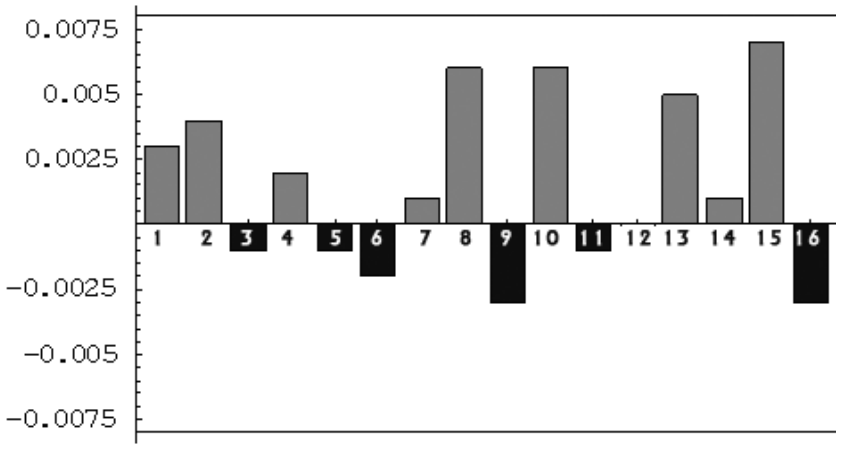


Figura 4: Autocorrelaciones con desplazamientos desde 0 hasta 16 unidades para 50,000 promedios de pares de cifras de π.

Memoria FONE XXIII (2009), Sección II, 69–74

Prueba de asociatividad para cópulas*

José M. González-Barrios[a]
IIMAS – Universidad Nacional Autónoma de México

1. Introducción

En muchas aplicaciones se ha tratado de ajustar cópulas Arquimedeanas a datos bivariados, ver definiciones en Nelsen (2006), para modelar la dependencia que existe entres sus coordenadas. Una cópula Arquimedeana es necesariamente asociativa, por lo que antes de tratar de ajustar cualquier cópula de esta clase se debería ver si los datos pasan una prueba estadística de asociatividad. Hasta donde sabemos no existen pruebas de asociatividad conocidas en la literatura. En este trabajo se propone una prueba de este tipo.

2. Preliminares

Recordar que una cópula C es simplemente una función de distribución conjunta en I^2, donde $I = [0, 1]$ con marginales $U(0, 1)$. Un resultado básico de cópulas Arquimedeanas es:

Definición 2.1. *Una cópula C es Arquimedeana si y sólo si satisface las siguientes dos condiciones:*

1. $C(C(u, v), w) = C(u, C(v, w))$ *para todas* $u, v, w \in I$, *es decir,* C *es* **asociativa**.
2. $\delta_C(u) = C(u, u) < u$ *para toda* $u \in (0, 1)$, *que es una condición sobre la* **diagonal** *de* C.

Sea $L = \{0, 1, \ldots, n\}$

Definición 2.2. *Una cópula discreta C definida en L es una operación binaria sobre L, i.e., $C : L \times L \to L$ que satisface las siguientes propiedades:*

*Este trabajo fue apoyado por el Proyecvto de Conacyt 50152
[a]gonzaba@sigma.iimas.unam.mx

1. $C(i,0) = C(0,j) = 0$ *para toda* $i, j \in L$.

2. $C(i,n) = C(n,i) = i$ *para toda* $i \in L$.

3. *Si* $0 \leq i \leq i' \leq n$ *y* $0 \leq j \leq j' \leq n$, *entonces*

$$C(i',j') - C(i',j) - C(i,j') + C(i,j) \geq 0,$$

es decir, C *es 2-creciente.*

Una cópula discreta se puede rescalar a una subcópula, ver Nelsen (2006), si en lugar de tomar L tomamos $L' = \{0, 1/n, 2/n, \ldots, 1\}$. Recordemos ahora la definición de cópula empírica.

Definición 2.3. *Sea* $S = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ *una muestra aleatoria de tamaño* n, *de un vector aleatorio con coordenadas continuas* (X, Y). *Definimos la* **cópula empírica** C_n *asociada a estos datos mediante la fórmula:*

$$C_n\left(\frac{i}{n}, \frac{j}{n}\right) = \frac{1}{n} \sum_{(x,y)\in S} \mathbf{1}_{(-\infty, x_{(i)}]\times(-\infty, y_{(j)}]}(x, y),$$

donde $x_{(i)}$, $i = 1, \ldots, n$ *y* $y_{(j)}$, $j = 1, \ldots, n$ *denotan las estadísticas de orden de las coordenadas* X *y* Y *respectivamente, y* $\mathbf{1}_A$ *denota la función indicadora del conjunto* A. *También definimos*

$$C_n\left(\frac{i}{n}, 0\right) = 0 \quad y \quad C_n\left(0, \frac{j}{n}\right) = 0 \quad \text{para cada} \quad i, j = 1, \ldots, n.$$

Observemos que el dominio de una cópula empírica C_n es el conjunto $\{0, 1/n, \ldots (n-1)/n, 1\} \times \{0, 1/n, \ldots (n-1)/n, 1\}$, por lo que C_n es una subcópula en lugar de ser una cópula.

Una cópula empírica es invariante bajo transformaciones estrictamente crecientes de los datos, por lo que supondremos que la muestra está dada por $(1, \sigma(1)), \ldots (n, \sigma(n))$, con σ una permutación de $\{1, 2, \ldots, n\}$.

Por lo tanto, en lo que sigue estudiaremos una forma totalmente equivalente de la cópula empírica dada por

$$C'_n(i,j) = \sum_{(x,y)\in S} \mathbf{1}_{(-\infty, i]\times(-\infty, y_{(j)}]}(x, y),$$

donde la muestra modificada está dada por $S = \{(1, \sigma(1) = Y_1), (2, \sigma(2) = Y_2), \ldots, (n, \sigma(n) = Y_n)\}$, $(\sigma(1), \sigma(2), \ldots, \sigma(n))$ es una permutación σ de $\{1, 2, \ldots, n\}$ y $C'_n(i, 0) = 0 = C'_n(0, j)$.

Observemos que C'_n es simplemente una cópula discreta con la definición anterior.

Sea C una copula discreta en $L = \{0, 1, \ldots, n\}$, definida en la sección anterior. Entonces C es **asociativa** si y sólo si $C(C(i, j), k) = C(i, C(j, k))$ para todas $i, j, k \in L$. Un elemento $i \in l$ es un **idempotente** de C si y sólo si $C(i, i) = i$.

Definamos la **matriz de permutación de Łukasiewicz** de orden $n \times n$ mediante $A = (a_{ij})_{i,j\in\{1,\ldots,n\}}$, donde $a_{ij} = 1$ si $i + j = n + 1$, y $a_{ij} = 0$ en otro caso. Entonces una cópula discreta C es asociativa si y sólo si C es una suma ordinal de matrices de Łukasiewicz como lo prueban en la Proposición 9, Mayor et al. (2005).

Supongamos que $1 \leq i_1 < i_2 < \cdots < i_k = n$ son los **elementos idempotentes** de la muestra modificada $(1, \sigma(1)), \ldots, (n, \sigma(n))$, es decir $C'_n(i_j, i_j) = i_j$ para $j = 1, 2, \ldots, k$. Observemos que n es siempre un elemento idempotente de cualquier muestra ya que $C'_n(n, n) = n$.

Como una cópula discreta C es asociativa si y sólo si C es una suma ordinal de matrices de Łukasiewicz, que queda determinada por sus elementos idempotentes. Definimos una estadística que mide la asociatividad de una muestra.

Definición 2.4. *Sea* $(1, \sigma(1)), \ldots, (n, \sigma(n))$ *una muestra modificada con elementos idempotentes* $1 \leq i_1 < i_2 < \cdots < i_{k-1} < i_k = n$. *Definimos la* **estadística de asociatividad** *de la muestra mediante:*

$$A^n_\pi = \sum_{l=1}^{k} \sum_{j=i_{l-1}+1}^{i_l} (i_l + i_{l-1} - j + 1 - \sigma(j))^2, \tag{1}$$

donde $i_0 = 0$. *El simbolo* π *simplemente denota el hecho de que la estadística se basa en permutaciones.*

Entonces A^n_π mide la distancia cuadrátrica entre la muestra modificada y la suma ordinal de matrices de permutación de Łukasiewicz con elementos idempotentes $1 \leq i_n < i_2 < \cdots < i_{k-1} < i_k = n$. Por lo tanto A^n_π es una medida de la asociatividad de una muestra modificada.

3. Propiedades de A_π^n y nueva prueba de asociatividad

Teorema 3.1. *Sea* $\underline{X_n} = (X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$ *una muestra aleatoria continua de tamaño* n *donde* X *y* Y *son variables aleatorias continuas e independientes, es decir, la pareja* (X, Y) *tiene cópula* Π. *Sea* $(1, \sigma(1)), (2, \sigma(2)), \ldots, (n, \sigma(n))$ *la muestra modificada y sea* K *el número de elementos idempotentes de la cópula discreta* C_n *inducida por la muestra modificada. Entonces*

$$P(K = n) = \frac{1}{n!} \quad y \quad P(K = 1) = 1 - \frac{\sum_{j=1}^{n-1} K_1(j)(n-j)!}{n!},$$

donde $K_1(j)$ *es el número de permutaciones de* $\{1, 2, \ldots, j\}$ *con sólo un idempotente.*

Este Teorema nos dice que para n grande el número de idempotentes tiende a uno, por ejemplo si $n > 500$, por lo que podemos estudiar una simplificación de A_π^n dada por:

$$\hat{A}_\pi^n = \sum_{j=1}^{n} (n - j + 1 - \sigma(j))^2. \quad \text{Entonces} \quad A_\pi^n = \hat{A}_\pi^n \qquad \text{si} \qquad K = 1.$$

Consideremos

$$A_n = \frac{\hat{A}_\pi^n - E(\hat{A}_\pi^n)}{\sqrt{\text{Var}(\hat{A}_\pi^n)}}.$$

Entonces A_n es una variable aleatoria discreta simétrica, con $E(A_n) = 0$ y $\text{Var}(A_n) = 1$.

Daremos la distribución asintótica de A_n. Para esto recordemos la rho de Spearman, ver por ejemplo Nelsen (2006). Sea $\{(X_i, Y_i)\}_{i=1}^n$ una muestra aleatoria continua de una distribución conjunta F. Supongamos que $X_1 < X_2 < \cdots < X_n$, y sea

$$\rho = \frac{12}{n(n^2-1)} \left\{ \sum_{i=1}^{n} i R_i - \frac{n(n+1)^2}{4} \right\},$$

donde $R_1, R_2, \ldots, R_n$ son los rangos de $Y_1, Y_2, \ldots, Y_n$. En nuestra notación R_i corresponde a $\sigma(i)$ para $i = 1, 2, \ldots, n$. Entonces

$$\begin{aligned} A_n &= \frac{\hat{A}_\pi^n - E(\hat{A}_\pi^n)}{\sqrt{\text{Var}(\hat{A}_\pi^n)}} \\ &= \sqrt{n-1} \cdot \rho \end{aligned}$$

Es bien sabido que bajo independencia $\sqrt{n-1}\rho$ es asintóticamente $N(0,1)$, y la aproximación a la normal es muy rápida.

Para probar la hipótesis

$$H_0 : C \text{ es una cópula asociativa} \quad \text{vs} \quad H_1 : C \text{ no es una cópula asociativa},$$

al nivel $0 < \alpha < 1$. Proponemos la siguiente metodología para $n \leq 500$:

- Obtener el valor de A_π^n para la muestra modificada.
- Simular un número grande m de muestras de tamaño n provenientes de la cópula independiente Π.
- Obtener para cada muestra simulada el valor de A_π^n.
- Estimar el cuantil $(1-\alpha)$ de A_π^n usando las m simulaciones.
- Rechazar H_0 si el valor de A_π^n de la muestra original excede al cuantil estimado.

Si $n > 500$ usar la normalidad asintótica de $\hat{A}_\pi^n$ para obtener los cuantiles, ya que A_π^n es muy cercana en distribución a $\hat{A}_\pi^n$.

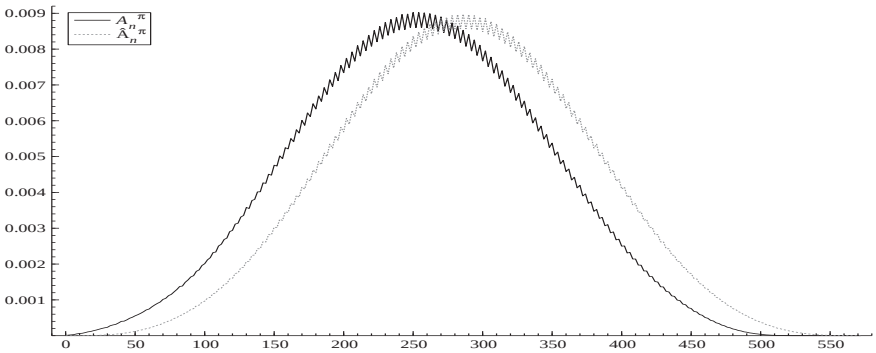


Figura 1: Densidades exactas de A_π^{12} y $\hat{A}_\pi^{12}$

Se utilizan muestras independientes ya que bajo independencia se tiene un buen rango de variación para los valores de A_π^n y de $\hat{A}_\pi^n$. No incluimos ejemplos por falta de espacio, pero la prueba funciona bien cuando hay evidencia de no asociatividad.

Referencias

Mayor, G., Suñer, J. and Torrens, J. 2005. Copula-like operations on finite settings. *IEEE Transactions on Fuzzy Systems*. **13**, 468-477.

Nelsen, R.B. 2006. *An introduction to copulas*, 2nd edition. Springer, New York.

Memoria FONE XXIII (2009), Sección II, 75–81

Un modelo para el máximo de un conjunto de observaciones dependientes

Elizabeth González Estrada[a], José A. Villaseñor Alva
Colegio de Postgraduados

1. Introducción

El estudio del comportamiento probabilístico de las concentraciones de contaminantes en la atmósfera es un tema de gran interés con el propósito de tomar acciones en favor de la protección de la salud de la población en grandes ciudades. En este trabajo se propone un modelo probabilístico paramétrico para los máximos de grupos de concentraciones diarias que exceden un umbral dado, el cual toma en consideración la posible dependencia que existe entre las observaciones dentro de los grupos. Este modelo proporciona una familia de distribuciones conjuntas para el máximo del grupo (Y) y el número de observaciones dentro del grupo (N), el cual es una variable aleatoria, por lo que es posible realizar predicciones sobre el número esperado de observaciones del grupo dado que el máximo del grupo no excede un nivel dado y, mediante el uso de la distribución condicionada de N dado Y. En este contexto la literatura consultada no reporta ningún antecedente. Con referencia a estudios relacionados a la variable Y véase Coles (2001).

El contenido de este trabajo es el siguiente. En la Sección 2 se describe el modelo de grupos de observaciones en el tiempo y el modelo conjunto para las variables (Y, N). En la Sección 3 se discuten los métodos de estimación para los parámetros del modelo conjunto. La Sección 4 contiene una aplicación a un conjunto de datos de ozono de la ciudad de México. Por último se presenta una sección de conclusiones.

[a]egonzalez@colpos.mx

2. El modelo

Sea $X_1, X_2, , X_3, ...$ una serie estacionaria de observaciones con función de distribución común F_X. Se dice que ocurre una excedencia en X_j si $X_j > u$, donde u denota un umbral dado. Un *grupo* de excedencias es un conjunto de observaciones dependientes en el que ha ocurrido al menos una excedencia. Dos grupos adyacentes en el tiempo están separados por un intervalo de tiempo fijo en donde no ocurrieron excedencias, (Coles, 2001).

Sea $C_1, C_2, C_3, ...$ una secuencia en el tiempo de grupos de excedencias. Para $X_i \in C_j$ se define un valor exceso como $X_i - u$ dado que $X_i > u$ con función de distribución $F_u(x) = \dfrac{F_X(u+x) - F_X(u)}{1 - F_X(u)}$, $x \geq 0$.

El máximo del j–ésimo grupo de excesos se define como $Y_j = \text{máx}\{X_i - u : X_i \in C_j\}$, $j = 1, 2, ...$

Sea $N_1, N_2, N_3, ...$ una secuencia de v.a. valuadas en los enteros no negativos las cuales denotan los tamaños de grupos. En la Figura 1 se ilustra la relación que existe entre las variables Y y N dado u.

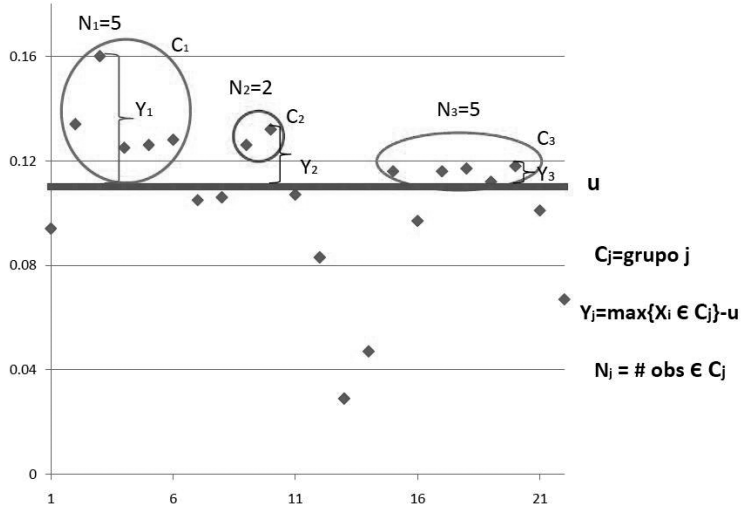


Figura 1: Relación entre Y_j, N_j y u.

2.1. Distribución conjunta de Y y N

Supóngase que los máximos de grupos de excesos, $Y_1, Y_2, ...$, son copias de una v.a. Y con función de distribución F_Y. Suponga que los tamaños de grupos $N_1, N_2, ...$ son copias iid de una v.a. N con función de densidad P_N y media $1/\theta$.

Para proponer un modelo para la distribución conjunta de M y N, primero definimos las distribuciones condicionales de la v.a. M para cada valor dado n de N y posteriormente modelamos la distribución marginal de N.

Si suponemos que dentro de cada grupo se tiene una realización de tamaño n de una serie estacionaria de observaciones dependientes con distribución marginal común F_u, entonces el índice extremo de la serie (θ) existe para observaciones débilmente dependientes. De acuerdo con la definicón del índice extremo (Leadbetter, 1983), se sigue que en términos generales, la distribución de Y se puede aproximar con $F_u^{n\theta}$ (Hsing, 1993). Por lo tanto, para un valor dado n de N proponemos modelar la distribución condicional de Y con

$$P\left(Y \leq y \mid N = n\right) = \left[F_u\left(y\right)\right]^{n\theta}, y > 0,$$

donde F_u es la distribución de los excesos. Cuando $\theta = 1$ las observaciones del grupo son independientes.

Por lo tanto, la función de distribución marginal de Y es:

$$F_Y\left(y\right) = \sum_{n=0}^{\infty} F_u^{n\theta}\left(y\right) P_N\left(n\right) = \varphi_N\left(F_u^{\theta}\left(y\right)\right), \tag{1}$$

donde φ_N denota la función generatriz de probabilidades de N. Entonces, con el propósito de hacer predicción sobre N con base en Y, considérese la siguiente expresión $E\left\{N \mid Y \leq y\right\} = \dfrac{F_u^{\theta}\left(y\right)\varphi_N'\left(F_u^{\theta}\left(y\right)\right)}{\varphi_N\left(F_u^{\theta}\left(y\right)\right)}$, donde φ_N' es la derivada de φ_N.

2.2. Distribución del tamaño de grupo

Suponga que N tiene distribución Binomial-Negativa(r, p) donde $r = \dfrac{p}{\theta\left(1-p\right)}$, $0 < p < 1$ y $\theta > 0$. Entonces $F_Y\left(y\right) = \varphi_N\left(F_u^{\theta}\left(y\right)\right) = \left\{p/\left(1 - qF_u^{\theta}(y)\right)\right\}^{p/\theta q}$, donde $q = 1 - p$. Así, la distribución de $N|Y \leq y$ es Binomial-Negativa$\left(r, 1 - qF_u^{\theta}\left(y\right)\right)$. Por lo tanto,

$$E\left\{N \mid Y \leq y\right\} = \frac{pF_u^{\theta}\left(y\right)}{\theta\left(1 - qF_u^{\theta}\left(y\right)\right)} \rightarrow \frac{1}{\theta}, y \rightarrow \infty. \tag{2}$$

2.3. Distribución de los excesos

Supóngase que $F_u(x)$ es la distribución Pareto generalizada $PG(\sigma, \gamma)$:

$$G(x; \gamma, \sigma) = 1 - \left(1 + \frac{\gamma}{\sigma} x\right)^{-1/\gamma},$$

con $\sigma > 0$ y $\gamma \in R$ tales que $x > 0$ para $\gamma \geq 0$ y $0 < x < -\sigma/\gamma$ cuando $\gamma < 0$.

3. Estimación de los parámetros

Se tiene el problema de estimar el parámetro θ, los parámetros γ y σ de la distribución Pareto generalizada y el parámetro p de la distribución binomial negativa con base en los máximos de grupos y tamaños de grupos observados (Y_i, N_i) $i = 1, 2, ..., n$.

El estimador de máxima verosimilitud de θ, el tamaño promedio de grupo, está dado por $\hat{\theta} = 1/\bar{N}$, donde $\bar{N} = \frac{1}{n} \sum_{i=1}^{n} N_i$.

El estimador obtenido por el método de momentos del parámetro p de la distribución Binomial-Negativa es $\hat{p} = \left(\sum_{i=1}^{n} N_i^2 - n\bar{N}^2\right) / (n-1)\bar{N}$.

3.1. Estimación de los parámetros γ y σ

Es bien conocido que los estimadores de máxima verosimilitud no siempre existen para la familia Pareto generalizada (Coles, 2001), por lo que a continuación se propone una metodología de estimación alternativa que asegura la obtención de estimadores para los parámetros γ y σ.

3.1.1. Caso $\gamma \geq 0$

Usando el método de estimación de máxima verosimilitud asintótica propuesto por Klüppelberg y Villaseñor (1993) se obtienen los siguientes estimadores de γ y σ basados en las k estadísticas extremas superiores:

$$\widehat{\gamma} = -\left(W_{n-k+1} - \frac{1}{k} \sum_{j=1}^{k} W_{n-j+1}\right)$$

y

$$\widehat{\sigma} = \widehat{\gamma} \exp\left\{W_{n-k+1} - \widehat{\gamma} \log(k/n)\right\},$$

donde $W_j = \log X_{(j)}, j = n-k+1, n-k+2, ..., n$ y $X_{(1)} < X_{(2)} < ... < X_{(n)}$ son las estadísticas de orden de una muestra aleatoria $X_1, X_2, ..., X_n$ de la distribución PG(σ, γ).

3.1.2. Caso $\gamma < 0$

Sea $U = \left(\bar{F}(X)\right)^{-\gamma}$, esto es, $U = 1 + \frac{\gamma}{\sigma}X$. Note que U tiene distribución $Beta(-1/\gamma, 1)$. Proponemos el siguiente procedimiento de dos etapas para estimar el parámetro γ.

3.1.3. Etapa 1: Método de Momentos

Sean $X_1, X_2, ..., X_n$ una muestra aleatoria de tamaño n de la distribución PG(σ, γ). El momento muestral de primer orden de U es $m = \frac{1}{n}\sum_{i=1}^{n}\left(1 + \frac{\gamma}{\sigma}X_i\right) = 1 + \frac{\gamma}{\sigma}\bar{X}$ donde $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$. Por otro lado, el valor esperado de U es $E\{U\} = 1/(1-\gamma)$. Entonces, por el método de momentos, un estimador $\tilde{\gamma}$ debe satisfacer la ecuación: $\frac{1}{1-\gamma} = 1 + \frac{\gamma}{\sigma}\bar{X}$. Resolviendo para γ se obtiene:

$$\gamma = 1 - \frac{\sigma}{\bar{X}}. \tag{3}$$

3.1.4. Etapa 2: Máxima Verosimilitud

De la definición de la distribución PG(σ, γ), se tiene que $0 < x < \frac{\sigma}{-\gamma}$, cuando $\gamma < 0$. Entonces, el EMV de $\frac{\sigma}{-\gamma}$ es $X_{(n)} = \text{máx}\{X_1, X_2, ..., X_n\}$.

Un estimador $\hat{\sigma}$ de σ es dado por:

$$\hat{\sigma} = -\gamma X_{(n)}. \tag{4}$$

Por lo tanto, sustituyendo $\hat{\sigma}$ de (4) por σ en (3) se tiene:

$$\tilde{\gamma} = \frac{\bar{X}}{\bar{X} - X_{(n)}} \text{ y } \tilde{\sigma} = \tilde{\gamma}X_{(n)}.$$

4. Aplicación: Datos de ozono de la Ciudad de México

Los datos considerados son los niveles máximos diarios de ozono (en partes por millón ppm) de los veranos de los años 2001 a 2007 registrados en la estación Pedregal de la Red Atmosférica de Monitoreo Ambiental de la Cd. México.

Para usar el modelo propuesto se fijó $u = .11$ ppm. Con base en el conocimiento de la meteorología en el Valle de México, se sabe que dos eventos son independientes cuando transcurren al menos dos días entre ellos. Por lo tanto, los grupos de excesos fueron conformados con una separación entre ellos de dos días. El número de grupos obtenidos es 69.

En el Cuadro 1 se presentan algunos tamaños promedios de grupos estimados dados diferentes niveles de excesos máximos. Estos valores fueron calculados usando la expresión (2), en donde F_u es la distribución Pareto generalizada y los parámetros fueron reemplazados por sus estimaciones correspondientes: $\hat{\theta} = 0.27684$, $\hat{p} = 0.43387$, $\tilde{\gamma} = -0.34899$ y $\tilde{\sigma} = 0.0607$.

Note que la estimación del tamaño promedio de grupo condicionado a que el máximo de grupo no excede un nivel dado y tiende a $1/\hat{\theta} = 3.6$ a medida que y se incrementa. Este aspecto es congruente con la idea intuitiva de que si en un punto en el tiempo se observa un exceso grande entonces se espera tener un tamaño de grupo grande. Esta idea es apoyada en la expresión (2) y en la Figura 2, en la cual la línea continua representa la estimación del tamaño promedio de grupo condicionado correspondiente a que el máximo de grupo no excede un nivel y dado y la línea punteada representa el tamaño promedio de grupo condicional observado.

y	$E\{N \mid Y \leq y\}$	y	$E\{N \mid Y \leq y\}$
0.02	1.88	0.09	3.33
0.04	2.50	0.10	3.41
0.06	2.92	0.12	3.53
0.07	3.08	0.15	3.60
0.08	3.22	0.20	3.60

Tabla 1: Tamaño promedio de grupo condicionado a que el máximo no excede un nivel dado y.

5. Conclusiones

De los resultados obtenidos en este trabajo se puede concluir que el modelo propuesto ajusta razonablemente el conjunto de datos estudiados. Asimismo, del modelo propuesto es posible obtener la distribución marginal para describir el comportamiento probabilístico del máximo de grupo. En la literatura revisada no se encontró ningún antecedente para modelar la distribución conjunta de Y y N.

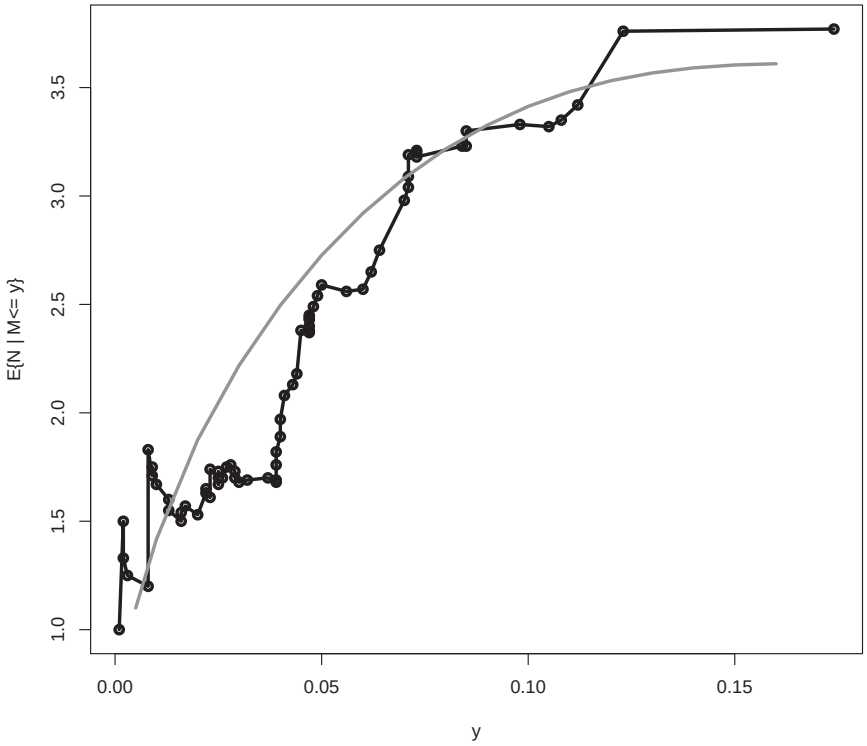


Figura 2: Tamaño promedio de grupo estimado y observado condicionado a que el máximo no excede un nivel dado y.

Referencias

Coles, S. 2001. An introduction to statistical modeling of extreme values. Springer. 208 p.

Hsing T. 1993. Extremal index estimation for a weakly dependent stationary sequence. Annals of Statistics 21: 2043–2071.

Klüppelberg, C., Villaseñor, J. A. 1993. Estimation of distribution tails: a semiparametric approach. DGVM Blätter, Band XXI, Heft 2, 213–235.

Leadbetter MR. 1983. Extremes and local dependence in a stationary sequence. Z. Wahrsch. Verw. Gebiete 65: 291–306.

Memoria FONE XXIII (2009), Sección II, 83–88

Modelos autorregresivos para series de tiempo ambientales

Lorelie Hernández[a], Gabriel Escarela[b], Angélica Hernández[c]
Universidad Autónoma Metropolitana – Iztapalapa

1. Introducción

En la ciencia ambiental uno de los propósitos principales es analizar series de tiempo cuyas respuestas pueden ser discretas o no-Gaussianas. Esto ocurre, por ejemplo, cuando se modela el número de hospitalizaciones o tasas de mortalidad. Otro ejemplo claro, se encuentra cuando se desea determinar si existe una tendencia en los niveles de contaminación de una metrópolis en presencia de otras variables atmosféricas y periódicas. Muchas de estas series de tiempo pueden ser convenientemente estudiadas a través de la construcción de modelos autorregresivos de orden uno (AR(1)). El propósito de este trabajo es mostrar el uso de la técnica de la cópula para la construcción de modelos AR(1) en el estudio de series de tiempo no-Gaussianas.

2. Definición del modelo

Un proceso de Markov estacionario de primer orden a tiempo discreto cuyo espacio de estados es continuo, puede construirse a partir de una distribución bivariada dada mediante $F(y_1, y_2) = \Pr\{Y_1 \leq y_1, Y_2 \leq y_2\}$, la cual corresponde a un vector aleatorio conjuntamente continuo (Y_1, Y_2) con ambas distribuciones marginales univariadas iguales a la *distribución estacionaria*. De esta forma, la *distribución de transición*, definida como $F_{2|1}(y_2 \mid y_1) = \Pr\{Y_2 \leq y_2 \,|\, Y_1 = y_1\}$, representa a la serie de tiempo AR(1).

[a] heilerol@yahoo.com.mx
[b] gabriel@escarela.com
[c] angyka302@gmail.com

No existen muchas distribuciones bivariadas con las propiedades que se acaban de mencionar. Sin embargo, una forma relativamente fácil de construir las distribuciones de transición es usando modelos de cópula. Una *cópula* bidimensional es una función de distribución bivariada de un vector aleatorio $\mathbf{V} = (V_1, V_2)$ cuyas marginales V_1 y V_2 son uniformes en el intervalo $(0, 1)$; aquí, el grado de asociación entre las dos marginales es controlado por un parámetro o un vector de parámetros denotado por θ (Joe, (1997)).

Si $F_1(y_1)$, $F_2(y_2)$ son continuas y la cópula $C(v_1, v_2)$ son diferenciables, la densidad conjunta de (Y_1, Y_2) correspondiente a la función de distribución conjunta, puede expresarse como $f(y_1, y_2) = f_1(y_1)f_2(y_2) \times c[F_1(y_1), F_2(y_2)]$, donde $f_1(y_1)$ y $f_2(y_2)$ son las funciones de densidad marginales correspondientes y $c(v_1, v_2) = \partial^2 C(v_1, v_2)/\partial v_1 \partial v_2$ es la *función de densidad de la cópula*; en este contexto, $(v_1, v_2)^T \in (0, 1)^2$. Como consecuencia, se tiene que la función de densidad condicional de Y_2 dada Y_1 puede expresarse como $f_{2|1}(y_2 \mid y_1) = f_2(y_2) \times c[F_1(y_1), F_2(y_2)]$. Usando esta densidad condicional en términos de una función de densidad de cópula y una marginal dada, $Y_t \sim f$, donde Y_t denota a la respuesta en el tiempo t con $t = 1, ..., n$, se puede construir un modelo de transición para respuestas continuas; para ajustar los modelos resultantes es posible usar la técnica de máxima verosimilitud; para el modelo AR(1) ésta función es $L = f(y_1) \prod_{t=2}^{n} f_{2|1}(y_t \mid y_{t-1})$, donde y_t representa el valor observado de Y_t, y $f(y)$ es la función de densidad correspondiente a la distribución marginal.

3. Ilustración: Series de tiempo de extremos

3.1. Los datos de Guadalajara

La meta principal de este estudio es evaluar si las diversas políticas ambientales implementadas en el Área Metropolitana de la Ciudad de Guadalajara en un periodo de diez años han reflejado una tendencia a la baja en los niveles de contaminación. Para ello, se analizan los niveles máximos diarios de ozono Y_t medidos en partes por millon (ppm), los cuales fueron registrados por siete estaciones de monitoreo localizados en el Área Metropolitana de dicha ciudad a partir del 6 de enero de 1997 hasta el 31 de Diciembre de 2006. Debido a que se comprobó que la concentración de ozono es mayor a media tarde, se tomó el valor máximo entre las 12 y 17 horas de toda la red de monitoreo de cada día; de manera análoga, las variables atmosféricas empleadas fueron restringidas a éste horario.

3.2. Un modelo AR(1) para máximos

Debido a que cada observación representa el máximo de un bloque, en este estudio se emplea la distribución marginal de Valor Extremo Generalizado para la respuesta Y_t, donde t denota el índice del día, dado el vector de variables explicativas en el tiempo t denotado aquí por $\mathbf{x}_t$, de la siguiente forma:

$$F(y_t;\mathbf{x}_t) = \exp\left[-\left\{1+\gamma\left(\frac{y_t-\mu(\mathbf{x}_t)}{\sigma}\right)\right\}_+^{-1/\gamma}\right], \qquad \text{para} \quad y_t > \mu_t, \tag{1}$$

donde μ, σ y γ son los parámetros de localizacion, escala y forma respectivamente, con $-\infty < \mu < \infty$, $\sigma > 0$, $-\infty < \gamma < \infty$ y $h_+ = \text{máx}(h, 0)$; aquí, μ_t está relacionado con el vector de variables explicativas $\mathbf{x}_t$ a través de una componente lineal de manera tal que $\mu(\mathbf{x}_t) = \boldsymbol{\beta}^T\mathbf{x}_t$, donde $\boldsymbol{\beta}$ es el vector de coeficientes correspondiente. Para definir el modelo de transición para máximos, se empleará una cópula de valor extremo para modelar la estructura de dependencia entre las observaciones adyacentes como se especificó en la sección anterior. Específicamente, se usará la Cópula Positiva Estable, conocida también como Cópula Logística, la cual está dada por $C_\theta(v_1, v_2) = \exp\left\{-\left[(-\log v_1)^{1/\theta} + (-\log v_2)^{1/\theta}\right]^\theta\right\}$, donde $\theta \in (0, 1)$. Nótese que al incluir variables explicativas que corresponden al tiempo t en $\mathbf{x}_t$, de las cuales la tendencia t es una de ellas, la formulación resultante define un modelo AR(1) no estacionario.

La especificación del modelo de transición que resulta de copular a las marginales adyacentes dadas por la ecuación 1 usando la cópula de valor extremo, es una generalización de los modelos del valor extremo para estudiar excedentes de ozono de alto nivel para series de máximos no estacionarias propuesto por Smith (1989), el cual sólo considera a la tendencia t en $\mathbf{x}_t$ en un modelo que supone independencia entre las Y_t's; esto es, Smith (1989) usa la cópula de independencia definida por $C(v_1, v_2) = v_1\, v_2$. En la presente formulación se modela simultáneamente tanto a la dependencia entre las respuestas adyacentes y la dependencia de la respuesta con variables explicativas presentes y pasadas.

Para especificar el modelo condicional se usó la biblioteca `evd`, (Stephenson (2002)), del paquete estadístico **R**; también se usó la función `optim` para optimizar la función de log verosimilitud así como para obtener la matriz de covarianzas aproximada.

3.3. Implementación

Son diversas las variables atmosféricas que se registran en todas y cada una de las estaciones de monitoreo de la ciudad de Guadalajara; sin embargo, sólo se consideraron las variables atmosféricas que se enlistan a continuación junto con la notación y unidades de medida correspondientes:

ozono	o2	$\mu g/m^3$
dirección del viento	dv	grados al Norte
humedad relativa	hum	porcentaje
temperatura	tem	$^0\,C$
velocidad del viento	vv	m/s

La dirección del viento es una variable importante y es un elemento climatológico definido como "el aire en movimiento", el cual principalmente se describe por la velocidad y la dirección. En este estudio se considera un vector de viento que para poder incluirlo en el componente lineal es necesario hacer una parametrización pues su unidad de medida está en grados con respecto al norte. Para ello Huang y Smith, (1999) proponen crear las dos siguientes variables

$$\texttt{wu} = -\texttt{vv} * \frac{\sin(2\pi * \texttt{dv})}{360} \qquad \texttt{wv} = -\texttt{vv} * \frac{\cos(2\pi * \texttt{dv})}{360}.$$

La variable `wu` es la componente este-oeste del viento, la cual es positiva cuando el viento viene del oeste; de manera análoga, `wv` es la componente norte-sur, la cual es positiva cuando el viento viene del sur.

Para ajustar efectos no lineales del tiempo t se usan polinomios ortogonales de t de un grado predeterminado; además, para incluir efectos `anuales` se usan los términos $\cos(2\pi t)/365$ y $\sin(2\pi t)/365$ y para efectos `semestrales` se agregan los términos $\cos(2\pi t)/182.5$ y $\sin(2\pi t)/182.5$ donde 365 y 182.5 equivale a los dias que hay en un año y en un semestre respectivamente.

Se sabe que las temperaturas altas favorecen generalmente la difusión de contaminantes. Al contrario del viento, la humedad juega un papel negativo en el aumento de los contaminantes, por lo que, temperaturas altas junto con bajas velocidades de viento están asociadas con observaciones altas de ozono, es por esto que se considera una interacción entre éstas dos variables.

Cuando se realizó el ajuste de los datos, se encontró que en presencia de todas las variables explicativas, el parámetro de forma no tenía efectos significativos, por lo que se decidió emplear la distribución marginal Gumbel, la cual es un caso especial de la ecuación (1) y está representada por la siguiente función de distribución marginal:

$$F(y_t; \mathbf{x}_t) = \exp\left[-\exp\left\{-\left(\frac{y_t - \mu(\mathbf{x}_t)}{\sigma}\right)\right\}\right]$$

donde $-\infty < \mu < \infty$ y $\sigma > 0$. Para determinar con un alto grado de precisión el modelo más parsimonioso penalizando tanto al número de parámetros como al tamaño de la muestra, se utilizó el criterio **BIC** *(Bayesian Information Criterion)* definido mediante $2\ln L(\hat{\theta}) + n_p \ln n$, donde n_p es el número de parámetros. Este criterio consiste en elegir el modelo cuyo valor resulta ser el más pequeño. Utilizando también eliminación en retroceso se encontró que el mejor grado para el polinomio de t es cuatro. Las variables que resultaron significativas son las que aparecen en el modelo cuya formulación para el componente lineal en R es:

$$t^4 + \mathsf{semestrales} + \mathtt{dvel}_t * \mathtt{tem}_t + \mathtt{hum}_t + \mathtt{wv}_t.$$

En este contexto el superíndice denota el orden del poliniomio.

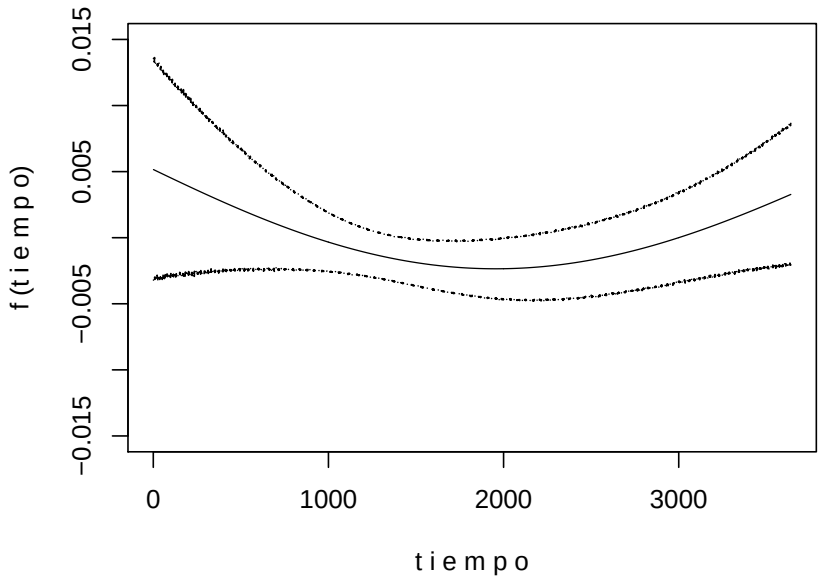


Figura 1: Ajuste del polinomio de tiempo del modelo más parsimonioso junto con bandas de confianza de 95 %.

La Figura 1 muestra gráficamente el ajuste del polinomio de los efectos del tiempo en el modelo más parsimonioso . Es posible observar que hay un decremento los primeros cinco años, que es cuando se implementó el plan de mejoramiento de la calidad del aire en Guadalajara, y después de este periodo hay un incremento sostenido. Es claro que hay un crecimiento en los máximos en ausencia de un programa de mejoramiento ambiental.

Referencias

Huang, L.; Smith, R. 1999. *Meteorologically-dependent Trends in Urban Ozone.* Environmetrics, **10**, 103-108.

Joe, H. 1997. *Multivariate Models and Dependence Concepts*. Chapman & Hall, New York.

Smith, R.L. 1989. *Extreme Value Analysis of Environmental Time Series: An Application to Trend Detection in Ground-Level Ozone*. Statistical Science, **4**, 367-377.

Song, P.X.K. 2000. *Multivariate Dispersion Models Generated from Gaussian Copula*. Scandinavian Journal of Statistics, **27**, 305-320.

Stephenson, A. 2002. *EVD: Extreme Value Distributions*. R News, **2**, 31-32.

Memoria FONE XXIII (2009), Sección II, 89–94

Métodos multivariados en la búsqueda de indicadores de degradación ambiental en la Sierra Norte de Puebla

Gladys Linares Fleites, Miguel Angel Valera Pérez
Instituto de Ciencias – BUAP

María Guadalupe Tenorio Arvide
Posgrado en Ciencias Ambientales. Instituto de Ciencias – BUAP

1. Introducción

Debido a la inquietud existente con respecto a la degradación del suelo y a la necesidad de un manejo sostenible de los agro- ecosistemas, ha resurgido el estudio de las propiedades del suelo, enfatizándose hacia una función específica del uso del suelo. Este enfoque ecológico del suelo reconoce las interacciones suelo - ser humano.

El objetivo del trabajo es identificar las propiedades del suelo que permitan establecer indicadores para diagnosticar la degradación de estos suelos y, por ende, la degradación ambiental.

2. Marco teórico

La región estudiada corresponde a Teziutlán, estado de Puebla. Esta se encuentra en la porción nor-oriental del estado, entre los paralelos 19°43´30´´ y 20°14´54´´ de latitud norte y los meridianos 97°07´42´´ y 97°43´30´´ de longitud occidental y a una altura que va de los 800 a los 3000 msnm. (Valera et al., 2001)

Las muestras de suelo corresponden a los horizontes A de 25 perfiles representativos y geo-referenciados, que fueron tomados como los individuos de la tabla de datos. Las variables consideradas fueron 12, a saber: % Carbono Orgánico ($\%CO$); % Nitrógeno Total

($\%N$);% Volumen Raíces ($\%VR$);% Residuos ($\%R$; porcentaje en volumen); Acidez Hidrolítica ($pH - H_2O$, relación 1:2); Acidez Intercambiable ($pH - KCl$, en relación 1:2); Capacidad de Intercambio Catiónico (CIC); Saturación en Bases ($\%V$); Calcio, Magnesio, Sodio y Potasio intercambiables (Ca, Mg, Na y K).

El análisis estadístico de los resultados obtenidos se efectuó con el programa MINITAB (Minitab, 2003).

2.1. Análisis exploratorio

En el análisis exploratorio de los datos se aplicó primero, la técnica de componentes principales para reducir la dimensión y obtener nuevas variables incorrelacionadas, y segundo, la técnica de conglomerado jerárquico aglomerativo con distancias euclidianas y enlace de Ward, para obtener agrupaciones de estos suelos.(Linares, 2006)

En el cuadro 1 se muestra el resultado de aplicar el ACP a la base de datos de la región de Teziutlán (se usó la matriz de correlaciones). Puede observarse que con las primeras cinco componentes se logra explicar el 83 % de la variabilidad total. En la primera componente (CP1) se destacan las variables $pH - H_20$, $pH - KCI$, Ca y Mg, mientras que en la segunda componente (CP2) las variables $\%N$,%Volrai y %Residuos muestran los coeficientes más altos. Las dos primeras componentes explican casi el 47 % de la variabilidad total y las tomaremos como indicadores de la degradación del suelo en la región. La primera componenete puede interpretarse como *grado de acidez de los suelos* y la segunda como *nivel de fertilidad*.

A partir de una nueva base de datos obtenida con las puntuaciones (*factor scores*) de las cinco primeras componentes en los 25 perfiles de suelo, se llevó a cabo el análisis de conglomerado que se muestra en la Figura 1. El análisis de conglomerados, conocido también con otros nombres como taxonomía numérica, clasificación automática, etc., son técnicas para la exploración de datos. (Johnson, 2000). Entre los diferentes usos de este tipo de análisis está descubrir si los datos reflejan alguna tipología correcta en un campo de estudio particular. Es precisamente el uso que se persigue al tratar de agrupar los suelos de la región de Teziutlán. Para ello hemos utilizado los algoritmos jerárquicos, en particular, el método aglomerativo donde cada objeto en el conjunto de información (perfiles de suelo, en nuestro caso) forma su propio *cluster* , y entonces, sucesivamente se van uniendo hasta quedar con un *cluster* que contiene el total de la información.

Valores Propios	3.6884	1.9048	1.6973	1.4722	1.2132
Prop. Expli	**0.307**	0.159	0.141	0.123	0.101
Prop. acum	0.307	0.466	0.608	0.730	0.831

Variable	PC1	PC2	PC3	PC4	PC5
%C.Org	0.079	0.248	0.355	**0.389**	-0.242
%N	0.060	**0.536**	0.115	-0.026	**0.489**
%Volrai	0.266	**0.444**	0.244	0.261	-0.090
%Residuos	-0.023	**0.462**	**-0.410**	0.084	0.271
pH-H20	**-0.413**	-0.130	-0.223	0.240	0.267
pH-KCI	**-0.415**	0.016	-0.152	**0.324**	0.219
CIC	-0.180	-0.154	0.061	**0.659**	-0.152
%V	-0.372	0.093	**0.399**	-0.299	-0.043
Ca	**-0.402**	0.078	**0.433**	-0.012	-0.021
Mg	**-0.402**	0.197	0.078	-0.221	-0.117
Na	-0.045	0.272	-0.339	0.050	**-0.591**
K	-0.284	0.268	-0.298	-0.188	-0.336

Tabla 1: Análisis de Componentes Principales de los suelos de la región de Teziutlán, Puebla.

En estos algoritmos se transforma la base de datos en una matriz de distancia entre los objetos y, posteriormente, por un proceso iterativo se agrupan los objetos para hacer un gráfico denominado *diagrama de árbol o dendrograma* . En el proceso iterativo se utilizan pseudo-distancias llamadas *enlaces* . En este trabajo utilizamos el *enlace de Ward* , en donde la distancia entre dos agrupamientos se define como el cuadrado de la distancia entre las medias de esos agrupamientos dividida entre la suma de los recíprocos de la cantidad de puntos que se encuentran dentro de cada uno de estos.

Como puede apreciarse en la Figura 1, los suelos de la región de Teziutlán pueden agruparse en tres conglomerados: en el extremo izquierdo del gráfico se muestran los perfiles de suelo con alta degradación, que llamaremos grupo 1 y que agrupa 7 perfiles; en el extremo derecho se muestran los suelos mejor conservados, que designaremos como grupo 2 y que agrupa 4 perfiles de suelo; en el centro se agrupan 14 perfiles con degradación moderada.

Estas deniominaciones se asignaron bajo un riguroso estudio de los paisajes de los sitios de donde se obtuvieron los perfiles analizados y fue realizado por especialistas en edafología.

Una pregunta de interés en este contexto es: ¿podrían los indicadores elaborados previamente a través de las dos primeras componentes principales caracterizar los tres grupos de perfiles de suelos obtenidos anteriormente?

Un estudio confirmatorio de datos a través del análisis discriminante lineal podría ayudar a esclarecer esta cuestión.

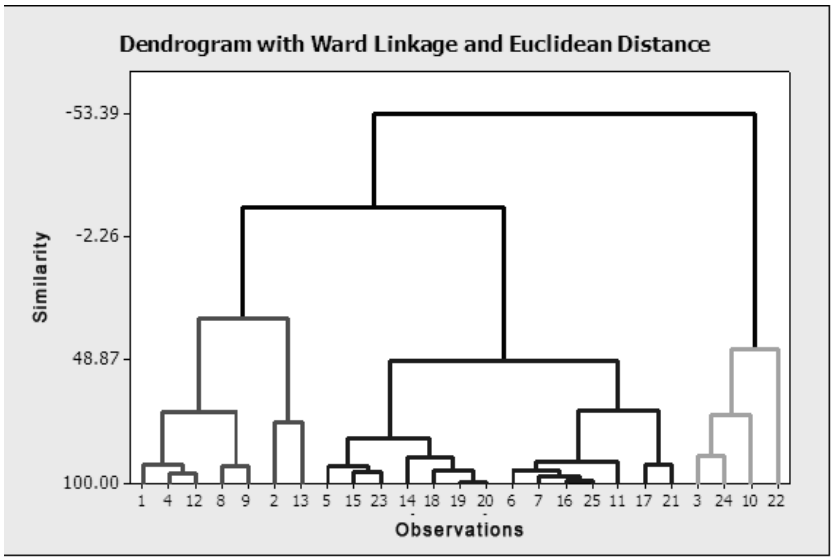


Figura 1: Análisis de Conglomerado Jerárquico de los suelos de la región de Teziutlán, Puebla

2.2. Análisis confirmatorio

Dado que existen diferentes enfoques en el problema de la discriminación, decidimos utilizar el análisis discriminante clásico (Peña, 2002). En este enfoque se construye una función discriminante para cada una de las poblaciones consideradas (*suelos altamente degradados, suelos conservados y suelos moderadamente degradados*) y se establece la regla de clasificación que coloca el individuo a clasificar en la población con valor máximo de la función discriminante. Dado que la utilidad de la regla de clasificación depende de los errores esperados, estos se calcularon aplicando la función discriminante a las 25 observaciones y clasificándolas. Este método tiende a subestimar las probabilidades de error de mala clasificación ya que los

mismos datos se utilizan para estimar los parámetros y para evaluar la regla resultante. Un procedimiento mejor es clasificar cada elemento (muestra de suelo) con una regla que no se ha construido usándolo. Para ello, se construyeron n=25 funciones discriminantes con las muestras que resultan al eliminar uno a uno cada elemento de la población y clasificar después cada dato en la regla construida sin él. Este método se conoce como validación cruzada y conduce a una mejor estimación del error de clasificación. En el estudio se utilizan ambos procedimientos de estimación de ese error y en ambos la proporción de clasificación correcta es de 0.92.

El Cuadro 2 muestra las funciones discriminantes de cada grupo. Sólo dos suelos fueron mal clasificados, el 4 y el 24. El perfil 4 considerado dentro del grupo de los altamente degradados y el 24 dentro de los conservados fueron clasificados como moderadamente degradado.

	1	2	3
	Alta degra	Conserv.	Moderada degra
Constante	-1.4145	-4.7220	-0.7470
CP1	0.3005	-2.5040	0.5652
CP2	1.7432	1.5760	-1.3219

Tabla 2: Función Discriminante para los grupos de suelos de la región de Teziutlán , Puebla.

3. Conclusiones

Los indicadores obtenidos a través del Análisis de Componentes Principales permiten diagnosticar el grado de degradación ambiental de los suelos de la región de Teziutlán en la Sierra Norte de Puebla, con error de mala clasificación bastante pequeño. La primera componente (CP1) expresa el grado de acidez de estos suelos mientras la segunda componente (CP2) señala el nivel de fertilidad . Contar con indicadores de degradación ambiental en la zona bajo estudio permitirá hacer recomendaciones sobre una agricultura sostenible y establecer políticas ambientales que redunden en beneficio de la sociedad.

Referencias

Johnson D.E. 2000. *Métodos Multivariados Aplicados al Análisis de Datos* International Thomson Editores. México

Linares, G. 2006. *Análisis de Datos Multivariados*. México: Editorial Benemérita Universidad Autónoma de Puebla.

Peña D. 2002. *Análisis de Datos Multivariantes* McGraw-Hill Interamericana de España, España

MINITAB 2003. SRelease 14.1 Minitab Inc

Valera, M.A., Tenorio. M.G., Linares, G., Ruiz, J. y Tamariz, J.V. 2001. *Aplicación de indicadores químicos de degradación para suelos ácidos de la Sierra Negra de Puebla*. Memorias COLOQUIOS Cuba-México sobre manejo sostenible de los suelos. Benemérita Universidad Autónoma de Puebla. pp 57-64. (ISBN 968 8635 10 3).

Memoria FONE XXIII (2009), Sección II, 95–99

Distribución de estimadores en modelos de regresión con parámetros sujetos a restricciones lineales de desigualdad

Leticia Gracia Medrano Valdelamar[a], Federico O'Reilly Togno[b]
Departamento de Probabilidad y Estadística IIMAS-UNAM

1. Introducción

Considere un problema de regresión del tipo $Y = X\boldsymbol{\beta} + \epsilon$ con Y el vector de respuestas, X matriz de variables explicativas y ϵ un vector aleatorio de media cero, componentes no correlacionadas y varianzas constantes. En ocasiones se requieren imponer restricciones lineales de desigualdad en el valor del vector $\boldsymbol{\beta}$. Esto resulta en un problema de mínimos cuadrados con restricciones, esto es: mín $\|Y - X\boldsymbol{\beta}\|^2$ sujeto a $A\boldsymbol{\beta} \geq \boldsymbol{a}$ con Y vector $n \times 1$, X matriz (rango completo) de $n \times p$, A matriz $m \times p$ y $\boldsymbol{a}$ vector $m \times 1$. Aquí se considera el caso $m \leq p$. Este problema es un caso particular de un problema de programación cuadrática, para el cual el paquete QUADPROG de R permite encontrar la solución; obteniéndose así el estimador puntual de $\boldsymbol{\beta}$. La distribución de las componentes del estimador de $\boldsymbol{\beta}$ puede tomar formas muy complicadas, que dependen de las restricciones. Un esfuerzo por encontrar la distribución final de las componentes de $\boldsymbol{\beta}$ aparece en el trabajo Rodríguez-Yam et. al (2004), que con base en un simulador tipo Gibbs permiten generar valores de dichas distribuciones finales. Ilustran su procedimiento con algunos ejemplos y aquí usaremos uno de ellos para comparar la distribución final obtenida a través de su simulador con la obtenida a través de un simulador que nosotros proponemos, basado en remuestreo paramétrico. Antes de presentar este remuestreo paramétrico para la regresión con restricciones, se ilustra la idea en un caso sencillísimo de estimación, con una sola restricción.

[a] lety@sigma.iimas.unam.mx
[b] federico@sigma.iimas.unam.mx

2. Ejemplo de estimación con una restricción

Considere el problema de hacer inferencias sobre el parámetro desconocido μ restringido a $\mu \geq 0$, con base en una sóla observación x, de la $\mathcal{N}(\mu, 1)$. Si se construye la distribución *a posteriori* para μ bajo la restricción de no negatividad, utilizando la *a priori* no informativa $\pi(\mu)d\mu \propto d\mu$ sobre el eje positivo, la distribución *a posteriori* resulta ser una normal con media x sobre los valores positivos de μ, esto es: $\phi(\mu|x) = \phi(\mu - x)/\Phi(x)$ restringida a $\mu > 0$.

Por otro lado, si se considera de manera clásica la inducción de una distribución en $[0, \infty)$ para μ basada en la "distribución de significancia", esto es, si se calcula el *p-value* para la hipótesis $H_0 : \tilde{\mu} < \mu$ con $\tilde{\mu}$ la verdadera media, y μ fija en $[0, \infty)$. Se obtiene que el *p-value* es $1 - \Phi(x - \mu)$, una función monótona creciente en μ, con los siguientes límites:

$$\lim_{\mu \to \infty} 1 - \Phi(x - \mu) = 1$$

y

$$\lim_{\mu \to 0} 1 - \Phi(x - \mu) = 1 - \Phi(x).$$

Así que esta "distribución" tiene una masa en $\mu = 0$ de tamaño $1 - \Phi(x)$, y para $\mu > 0$ se comporta como una normal centrada en x con densidad $\phi(\mu - x)$, (ver [2]). Debe observarse que si de esta "distribución" (fiducial) se calcula la distribución condicional para μ dado el evento $[\mu > 0]$, ésta coincide plenamente con la distribución *a posteriori*. Además el *p-value* asociado a la hipótesis nula $H_o : \tilde{\mu} = 0$ es popr construcción $1 - \Phi(x)$.

Se propone un simulador basado en remuestreo paramétrico para generar observaciones de la fiducial como sigue:

1. Obtenga el estimador máximo verosímil $\hat{\mu}$ para μ, $\hat{\mu} = x$ si $x > 0$ y $\hat{\mu} = 0$ si $x < 0$.

2. Colocando $\hat{\mu}$ en lugar del parámetro en la distribución $\mathcal{N}(\mu, 1)$ se generan $\tilde{x}_i$, con $i = 1, \ldots, N$. Para cada $\tilde{x}_i$ calculamos el estimador de μ, sujeto a la no negatividad. Estas N realizaciones del estimador de μ darán una buena aproximación de la distribución muestral del estimador que a su vez está relacionada con la fiducial.

Si $x > 0$ estas simulaciones difieren de aquellas hechas con la *a posteriori* en que aproximadamente una proporción $1 - \Phi(x)$ son 0, pero el resto es como si se hubieran muestreado de la distribución *a posteriori*.

Si $x < 0$ al seguir los pasos anteriores, aproximadamente el 50 % de los estimadores generados serán 0, y el resto como si provinieran de una $\mathcal{N}(0, 1)$.

Por lo anterior, puede sugerirse un cambio en el paso 1 y usar el estimador máximo verosímil no restringido, es decir $\hat{\mu}^* = x$, y curiosamente, se tendrían resultados iguales a los de la distribución *a posteriori*, excepto por la referida masa en cero. Esto, que aquí parece justificado, veremos que no proporciona un buen procedimiento en el ejemplo de la regresión.

El procedimiento de generación de muestras del estimador de μ usando este simulador paramétrico se ilustra a continuación dentro del contexto de **regresión restringida**, aquí el cálculo de la distribución *a posteriori* no es directo y se requiere de un proceso de simulación, como el propuesto por Rodríguez-Yam *et al.* (2004). Por supuesto que no existe una generalización de la distribución de significancia para regresión restringida; por ello el interés de usar un simulador basado en remuestreo paramétrico para explorar la distribución muestral y relacionar ésta con una posible "distribución fiducial".

3. Ejemplo de regresión restringida (datos de renta)

Los datos consisten en la renta y_i pagada por estudiantes en diferentes casas, las variables explicativas son: s_i el sexo de los ocupantes (0 para femenino, 1 para masculino), r_i el número de cuartos en la casa y d_i la distancia en cuadras de cada casa hacia la universidad. El modelo explorado es: $y_i = \beta_1 + \beta_2 s_i r_i + \beta_3 (1 - s_i) r_i + \beta_4 s_i d_i + \beta_5 (1 - s_i) d_i + e_i$ con las $e_i's \sim \mathcal{N}(0, \sigma^2)$, β_2 y $\beta_3 \geq 0$, ya que a mayor número de cuartos el precio de la renta debe ser más alto y β_4 y $\beta_5 \leq 0$, ya que a mayor distancia a la universidad menor debe ser el precio de la renta. El tamaño de muestra es $n = 32$ y el número de parámetros es $p = 5$.

Se hicieron dos corridas usando el simulador de remuestreo paramétrico como está descrito inicialmente (insertando el estimador restringido). Se generaron 10,000 remuestreos, insertando el estimador restringido: $\hat{\boldsymbol{\beta}}$=(37.63, 130.14, 123.04, 0, −1.15) y $\hat{\sigma}^2 = 1602.48$, obtenidos con los datos originales, aquí sólo las restricciones para $\hat{\beta}_4$ y $\hat{\beta}_5$ resultaron activas. También se exploró el remuestreo paramétrico pero habiendo insertado inicialmente los estimadores sin restricciones, $\hat{\boldsymbol{\beta}}$=(38.56, 103.56, 122.04, 3.32, −1.15) y $\hat{\sigma}^2 = 1395.46$. Se obtuvieron 10,000 remuestreos.

Para cada uno de los remuestreos, las $\hat{\beta}$ fueron guardadas, al igual que el número de veces que las restricciones resultaron activas, y se hicieron las gráficas de las distribuciones

de éstas y se compararon con las distribuciones simuladas por Rodríguez-Yam *et al.*

Se observó que cuando las restricciones se activan (sólo los casos de β_4 y β_5) en estos casos las distribuciones que producen el simulador paramétrico y el Gibbs de Rodríguez-Yam *et al.* no coinciden, para las otras β's las gráficas resultan casi idénticas.

En el caso de los estimadores de β_4, aproximadamente el 45 % de las veces $\hat{\beta}_4 = 0$, si uno quisiera asignarle un tipo de distribución fiducial a β_4, el valor de 0.45 puede tomarse como el *p-value* asociado a la hipótesis nula $H_0 : \beta_4 = 0$ y la distribución, ahora estandarizada en el intervalo $(0, \infty)$ puede compararse entonces con la de la distribución *a posteriori*, lo mismo puede hacerse para la β_5. Las gráficas de la figura 1 muestran como las distribuciones de las observaciones para β_4 y β_5 del simulador Gibbs y del simulador paramétrico, en el caso de la β_5 coinciden bastante, pero para la β_4 se alcanza a ver que la gráfica es ligeramente curva.

Para los remuestreos usando el estimador de $\hat{\boldsymbol{\beta}}$ no restingido al inicio, se observó por ejemplo que el número de veces que $\beta_4 = 0$, fue de 95 % aproximadamente, así que la distribución de estas observaciones no se parece a ninguna de las otras dos distribuciones anteriores.

4. Conclusiones

La relación entre las distribuciones muestrales de los parámetros y las "fiduciales" es a través de pivotales, tipo: $\sqrt{n}(x - \theta)$, que se distribuye asintóticamente con una distribución Q (conocida, típicamente de media cero) y que desde el punto de vista fiducial ha permitido decir que θ tiene como localización a x y debidamente estandarizada (por la $\sqrt{n}$) tiene distribución Q, o alternativamente que x tiene como localización θ y estandarizada (por $\sqrt{n}$), tiene distribución Q.

Por otro lado la *a posteriori* asintóticamente es, bajo condiciones generales también del tipo $\sqrt{n}(\theta - x)$ aproximadamente con distribución Q. En adición, hay resultados sobre las similitudes, no necesariamente asintóticas. En el caso complicado de regresión restingida acabamos de explorar una posible relación entre las finales (las de Rodríguez-Yam *et.al*, obtenidas con Gibbs) y las nuestras usando el remuestreo paramétrico; y parece que existe una relación excepto por las masas que aparecen en la fiducial. Pero condicionando en el intervalo abierto correspondiente (sin considerar los ceros) si hay resultados muy parecidos.

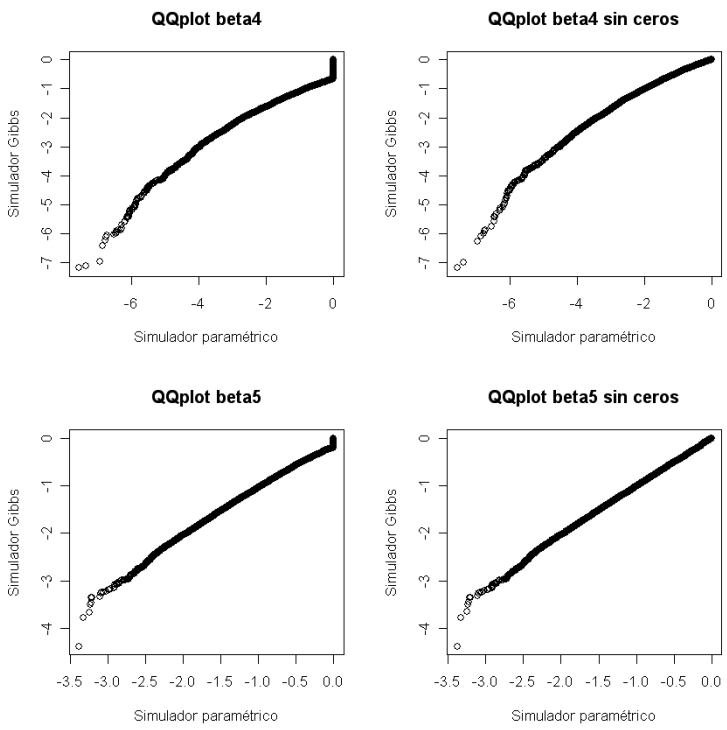


Figura 1: QQplots para las simulaciones de β_4 y β_5

Referencias

Rodríguez-Yam, G. A., Davis R. A. & Scharf, L L. 2004. *Efficient Gibbs sampling of truncates multivariate normal with application to constrained linear regresion.* `http://www.stat.columbia.edu/~ rdavis/technical_reports.html`

O'Reilly, F. 2003. *Significance Distributions* Preimpreso No. 117, IIMAS, UNAM.

Memoria FONE XXIII (2009), Sección II, 101–106

Comparación de estimadores de regresión del total bajo tres especificaciones de la matriz de varianzas y covarianzas

Ignacio Méndez Ramírez[a]
IIMAS – Universidad Nacional Autónoma de México

Flaviano Godínez Jaimes[b]
Unidad Académica de Matemáticas – Universidad Autónoma de Guerrero

Ma. Natividad Nava Hernández
Maestría en Ciencias Área Estadística Aplicada – Universidad Autónoma de Guerrero

1. Introducción

Las poblaciones grandes con frecuencia tienen una estructura compleja, es decir poblaciones estratificadas con varias etapas de formación de conglomerados. Además, la variable de interés puede estar asociada con variables auxiliares para las cuales se conocen sus valores en la población. En este trabajo se consideran ambas situaciones a través de estimadores de regresión en muestras complejas y se amplia el trabajo de Godínez y Méndez (2008) en donde la estimacion se hace considerando que la matriz de varianzas y covarianzas es la identidad. En este trabajo se considera tambien una matriz con las probabilidades de inclusión y otra que se obtiene usando las ecuaciones de estimación generalizadas.

Considere una población formada por L estratos y cada estrato esta formado por N^h unidades primarias de muestreo (UPM) con N_i^h elementos o unidades secundarias de muestreo (USM), $h = 1, \ldots, L$ e $i = 1, \ldots, N^h$. Raj (1968) propuso los esquemas A y B para muestrear este tipo de poblaciones. En el Esquema A en cada estrato las UPM se eligen por muestreo aleatorio simple sin reemplazo (MAS) y en las UPM seleccionadas se usa cualquier

[a] imendez@servidor.unam.mx
[b] fgodinezj@gmail.com

forma de muestreo incluso diferente. Una implementación del esquema es elegir con MAS n^h de las N^h UPM y con MAS n_i^h de las N_i^h USM. El vector de probabilidades de inclusión (PI) es $\boldsymbol{\pi} = (\boldsymbol{\pi}^1, ..., \boldsymbol{\pi}^L)^T$ donde $\boldsymbol{\pi}^h = (\pi_i)^h = (\pi_{Ii}\pi_{k|i})^h = (\frac{n^h}{N^h}\frac{n_i^h}{N_i^h}, \ldots, \frac{n^h}{N^h}\frac{n_i^h}{N_i^h})^T$ de $n = n^h n_i^h$, $h = 1, \cdots, L$. En el Esquema B las UPM en cada estrato se seleccionan con probabilidad proporcional al tamaño de la UPM con reemplazo (PPT) y cada vez que se extrae una UPM se realiza el muestreo dentro de ella. Una implementación del esquema es seleccionar n^h de las N^h UPM con PPT y en cada UPM en la muestra se seleccionan n_i^h de las N_i^h USM con MAS ($\pi_{Ii} = n^h\frac{N_i^h}{M^h}$ y $\pi_{Iij} = \frac{n^h(n^h-1)}{2}\frac{N_i^h(N_j^h)}{M^h(M^h)}$ con $M^h = \sum_{i=1}^{N^h} N_i^h$). Así, en el estrato h el vector $\boldsymbol{\pi}^h$ es $(n^h\frac{n_i^h}{M^h}, \ldots, n^h\frac{n_i^h}{M^h})^T$.

Cuando la variable de interés, Y, depende de las características de los individuos medidas en las variables auxiliares $X_1, X_2, \ldots, X_p$ es natural usar el modelo de regresión lineal para explicar la relación entre ellas. Este modelo, M, se expresa matricialmente:

$$E_M(\mathbf{Y}) = \mathbf{X}\boldsymbol{\beta}, \quad Var_M(\mathbf{Y}) = \mathbf{V}$$

donde $\mathbf{X}$ y $\mathbf{V}$ son matrices de Nxp y NxN de los valores poblacionales de las variables explicatorias y varianzas, $\mathbf{Y}$ y $\boldsymbol{\beta}$, son vectores de Nx1 y px1 de los valores poblacionales de las variables de interés y parámetros de regresión. Los Estimadores de Regresión del Total (ERT) aprovechan la relación lineal entre Y y $X_1, X_2, \ldots, X_p$ y que se conoce $\mathbf{X}$ o al menos los totales poblacionales, $\mathbf{1}^T\mathbf{X}$. Para obtener los ERT se pueden usar los enfoques Basado en Modelo (BM) o Asistido por Modelo (AM).

Sea $T = \sum_{i=1}^{L}\sum_{j=1}^{N^i}\sum_{k=1}^{N_j^h} Y_{ijk}$ el total poblacional. En el enfoque BM el ERT es $\hat{T}_{BM} = \mathbf{g}_s^T\mathbf{Y}_s$, donde $\mathbf{g}_s^T = \mathbf{1}_s^T + \mathbf{a}_s^T$ es el vector de pesos con $\mathbf{a}_s = \mathbf{V}_s^{-1}\mathbf{X}_s(\mathbf{X}_s^T\mathbf{V}_s^{-1}\mathbf{X}_s)^{-1}\mathbf{X}_r^T\mathbf{1}_r^T$ y s (r) indica las n ($N-n$) observaciones de $\mathbf{X}$, $\mathbf{V}$ y $\mathbf{Y}$ que están (no están) en la muestra. En la expresión de $\hat{T}_{BM}$ está implícito el calculo de $\hat{\boldsymbol{\beta}} = (\mathbf{X}_s^T\mathbf{V}_s^{-1}\mathbf{X}_s)^{-1}\mathbf{X}_s^T\mathbf{V}_s^{-1}\mathbf{Y}_s$. Valliant *et al.* (2000) muestra que $V(\hat{T}_{BM}) \approx \sum_{i=1}^{n^h}\mathbf{g}_i^T(Var_M(\mathbf{Y}_{si}))\mathbf{g}_i$ pues otros dos términos pueden ignorarse. Dos estimadores de $V(\hat{T}_{BM})$ son: $\hat{V}_1(\hat{T}_{BM}) = \sum_{i=1}^{n^h}\mathbf{g}_i^T(\mathbf{r}_i\mathbf{r}_i^T)\mathbf{g}_i$ y $\hat{V}_2(\hat{T}_{BM}) = \sum_{i=1}^{n^h}\mathbf{a}_i^T(\mathbf{r}_i\mathbf{r}_i^T)\mathbf{a}_i$, donde $\mathbf{g}_i$, $\mathbf{a}_i$ y $\mathbf{r}_i$ corresponden al cluster i en la muestra de $\mathbf{g}_s$, $\mathbf{a}_s$ y $\mathbf{r}_s = \mathbf{Y}_s - \mathbf{X}_s\hat{\boldsymbol{\beta}}$.

En el enfoque AM y el modelo de trabajo $E_M(\mathbf{Y}) = \mathbf{X}\boldsymbol{\beta}$, $Var_M(\mathbf{Y}) = diag(\sigma_1^2, \cdots, \sigma_N^2)$, el ERT es $\hat{T}_{AM} = \mathbf{g}_{sB}^T\mathbf{Y}_s$ conocido también como estimador de regresión generalizado (Särn-

dal *et al.* 1992) donde $g_{isB} = 1 + \left(\mathbf{1}^T\mathbf{X} - \mathbf{1}_s^T\mathbf{\Pi}_s^{-1}\mathbf{X}_s\right)^T \mathbf{A}_{\pi s}^{-1}\mathbf{x}_i \Big/ v_{ii}$ con $\mathbf{A}_{\pi s} = \mathbf{X}_s^T\mathbf{V}_s^{-1}\mathbf{\Pi}_s^{-1}\mathbf{X}_s$, $\mathbf{V}_s = diag\,(v_{ii})$, $\mathbf{\Pi}_s = diag\,(\pi_i)$ implicitamente aparece $\hat{\mathbf{B}} = \mathbf{A}_{\pi s}^{-1}\mathbf{X}_s^T\mathbf{V}_s^{-1}\mathbf{\Pi}_s^{-1}\mathbf{Y}_s$.

Una aproximacion de la varianza es

$$\hat{V}(\hat{T}_{AM}) = \sum_{i=1}^{n^h}\sum_{j=1}^{n^h}\left(\frac{\pi_{Iij} - \pi_{Ii}\pi_{Ij}}{\pi_{Iij}}\right)\frac{\hat{t}_{Ei}}{\pi_{Ii}}\frac{\hat{t}_{Ej}}{\pi_{Ij}} - \sum_{i=1}^{n^h}\frac{1}{\pi_{Ii}}\left(\frac{1}{\pi_{Ii}} - 1\right)\hat{V}_{BEi} + \sum_{i=1}^{n^h}\frac{\hat{V}_{BEi}}{\pi_{Ii}^2}$$

con $\hat{V}_{BEi} = \sum_{k=1}^{n_i^h}\sum_{l=1}^{n_i^h}\left(\frac{\pi_{kl|i} - \pi_{k|i}\,\pi_{l|i}}{\pi_{kl|i}}\right)\frac{g_{ksB}e_{ks}}{\pi_{k|i}}\frac{g_{lsB}e_{ls}}{\pi_{l|i}}$, $\hat{t}_{Ei} = \sum_{k=1}^{n_i^h} g_{ksB}e_{ks}/\pi_{k|i}$ y $e_{ks} = y_k - \mathbf{x}_k'\hat{\mathbf{B}}$.

Los estimadores cosméticamente calibrados poseen características de los estimadores AM y BM y aseguran la consistencia con totales conocidos de variables auxiliares. El ERT cosméticamente calibrado es

$$\begin{aligned}\hat{T}_{CC} &= \mathbf{1}_s^T\mathbf{\Pi}_s^{-1}\mathbf{Y}_s + \left(\mathbf{1}^T\mathbf{X} - \mathbf{1}_s^T\mathbf{\Pi}_s^{-1}\mathbf{X}_s\right)\left[\mathbf{X}_s^T\mathbf{Z}_s^{-1}\left(\mathbf{\Pi}_s^{-1} - \mathbf{I}_s\right)\mathbf{X}_s\right]^{-1}\mathbf{X}_s^T\mathbf{Z}_s^{-1}\left(\mathbf{\Pi}_s^{-1} - \mathbf{I}_s\right)\mathbf{Y}_s \\ &= \mathbf{1}_s^T\mathbf{Y}_s + \left(\mathbf{1}^T\mathbf{X} - \mathbf{1}_s^T\mathbf{X}_s\right)\left[\mathbf{X}_s^T\mathbf{Z}_s^{-1}\left(\mathbf{\Pi}_s^{-1} - \mathbf{I}_s\right)\mathbf{X}_s\right]^{-1}\mathbf{X}_s^T\mathbf{Z}_s^{-1}\left(\mathbf{\Pi}_s^{-1} - \mathbf{I}_s\right)\mathbf{Y}_s\end{aligned}$$

donde $\mathbf{Z}_s$ es una matriz diagonal de nxn tal que $\mathbf{Z}_s\mathbf{1}_s = \mathbf{X}_s\boldsymbol{\beta}$. Brewer (2002) aproxima $\hat{V}(\hat{T}_{CC})$ con $\hat{V}(\hat{T}_{CC}) = \frac{n}{n-p}\sum_{i=1}^{n}\pi_i^{-1}\left(\pi_i^{-1} - 1\right)\left(Y_i - \hat{\hat{Y}}\right)^2$.

2. Estudio de simulación

En cada estrato, Y, se genera con un modelo de efectos aleatorios

$$Y_{ij}^h = \beta_0 + \beta_1 x_{1ij}^h + \beta_2 x_{2ij}^h + \beta_3 x_{3ij}^h + A_i^h + E_{ij}^h$$

$h = 1, ..., 10$; $i = 1, ..., N^h$; $j = 1, ..., N_i^h$ y $A_i^h \sim N(0, \sigma_A^2)$ es independiente de $E_{ij}^h \sim N(0, \sigma_E^2)$. Las $x_k^h \sim N(\mu_k^h, \boldsymbol{\sigma}_k^2)$, $k = 1, 2, 3$ con $\mu_1^h = 1000 + 100000h$, $\mu_2^h = 5000 + 100000h$, $\mu_3^h = 9000 + 100000h$ y $\boldsymbol{\sigma}_k^2 = (400^2, 1000^2, 30000^2)$ y $\boldsymbol{\beta} = (5, k_b400, k_b1000, k_b3000)^T$. Por tanto $Var_M(\mathbf{Y}) = diag(\mathbf{V_1}, \ldots, \mathbf{V}_{N^h})$ con $\mathbf{V}_i = \sigma^2\mathbf{R}_i$ y $\mathbf{R}_i$ es la matriz de correlaciones intercambiable donde $\rho = \sigma_A^2/\sigma^2 = \sigma_A^2/(\sigma_A^2 + \sigma_E^2)$ es el coeficiente de correlación intraconglomerado que cuantifica la variación dentro de los clusters. Dever y Valliant (2006) afirman que el grado de asociación lineal entre la variable de interés y las variables auxiliares puede afectar el desempeño de los ERT. Una medida de esta asociación es $R^2 = SCR/SCT$, donde $SCR = \boldsymbol{\beta}^T\mathbf{X}^T\mathbf{y} - N^{-1}(\mathbf{1}^T\mathbf{y})^2$ y $SCT = \mathbf{y}^T\mathbf{y} - N^{-1}(\mathbf{1}^T\mathbf{y})^2$.

Para obtener semejanza débil dentro de los conglomerados, $\rho = 0.33$, se toma $\sigma_A^2 = 25000$ y $\sigma_E^2 = 5000$ mientras que para la fuerte, $\rho = 0.95$, se toma $\sigma_A^2 = 25000$ y $\sigma_E^2 = 1000$. La asociación débil y fuerte , $R^2 = 0.33$ y 0.95, entre la variable respuesta y las variables auxiliares se induce con $k_b = 0.00002, 0.0001, 0.00005, 0.00001$.

Los ERT dependen de $\mathbf{V}_s$, en el trabajo se estudia el desempeño de los estimadores considerando $\mathbf{V}_s = \mathbf{I}_s$, $\mathbf{\Pi}_s$ y $\mathbf{G}_s$ donde $\mathbf{G}$ indica que se usan las ecuaciones de estimación generalizadas. $\mathbf{G}_s$ se determina a través del siguiente proceso iterativo:

1. Obtener $\mathbf{b}_0$ el estimador por mínimos cuadrados ordinarios de $\boldsymbol{\beta}$, con $\mathbf{V}_s = \mathbf{I}_s$,

2. Usar $\mathbf{b}_0$ para calcular los residuos, $e_i^h = y_i^h - (\mathbf{x}_i^h)^T\mathbf{b}_0$, $\widehat{\mathbf{R}}$ y $\widehat{\mathbf{G}}_s$ con

$$\widehat{\rho} = (n-p)^{-1}\sum_{i=1}^{n^h}\sum_{j\leq n_i^h-1} e_{i,j}e_{i,j+1} \quad \text{y} \quad \widehat{\sigma}^2 = tr\sum_{i=1}^{n^h}(n-p)^{-1}(\mathbf{y}-\mathbf{X}_s\mathbf{b}_0)(\mathbf{y}-\mathbf{X}_s\mathbf{b}_0)^T$$

3. Calcular $\mathbf{b} = (\mathbf{X}_s^T\widehat{\mathbf{R}_s}\mathbf{X}_s)^{-1}\mathbf{X}_s^T\widehat{\mathbf{R}_s}\mathbf{y}_s$,

4. Regresar a 2 e iterar el proceso hasta lograr la convergencia.

En la simulación se consideran cuatro escenarios de acuerdo a los valores de ρ y R^2: 1) $0.33 - 0.33$, 2) $0.33 - 0.95$, 3) $0.95 - 0.33$ y 4) $0.95 - 0.95$. En cada uno se toman 15000 muestras de tamaño 200 (en cada estrato se seleccionan cuatro UPM y cinco USM en las UPM seleccionadas) con ambos esquemas. Los estimadores estudiados se comparan en función del Sesgo Relativo (SR) $\widehat{E}((\widehat{T}-T))/T$, Error Cuadrático Medio Relativo (ECMR) $\widehat{E}((\widehat{T}-T)^2)/T$ y Cubrimiento (C) de los intervalos de la forma $\widehat{T} \pm 2\sqrt{\widehat{V}(\widehat{T})}$). Se usa la notación BMI, BMP y BMG para representar el estimador BM cuando $\mathbf{V}_s$ es sustituida por $\mathbf{I}_s$, $\mathbf{\Pi}_s$ y $\mathbf{G}_s$ respectivamente y los otros estimadores de manera semejante.

2.1. Resultados principales

El parámetro de interés es T y los otros son parámetros de ruido. Los totales poblacionales para los escenarios mencionados son 248244266, 1241498843, 620637533 y 124313457. Debido a que son valores diferentes para cada escenario, solo se presentan resultados relativos. Cuando hay mas asociación lineal ($R^2 = 0.95$) los estimadores tienen menor ECMR, SR y

	33	33			33	95			95	95			95	33	
Est	$\mathbf{V}_s$	E	C	Est	$\mathbf{V}_s$	E	C	Est	$\mathbf{V}_s$	E	C	Est	$\mathbf{V}_s$	E	C
BM_1	$\mathbf{I}_s$	A	0,973	BM_1	$\mathbf{I}_s$	B	0,973	AM	$\mathbf{I}_s$	A	0,950	AM	$\mathbf{I}_s$	A	0,946
BM_1	$\mathbf{I}_s$	B	0,976	BM_2	$\mathbf{I}_s$	A	0,974	AM	$\mathbf{\Pi}_s$	A	0,949	AM	$\mathbf{\Pi}_s$	A	0,943
BM_2	$\mathbf{I}_s$	A	0,977	BM_1	$\mathbf{I}_s$	B	0,976					AM	$\mathbf{I}_s$I	B	0,976
BM_2	$\mathbf{I}_s$	B	0,979	BM_2	$\mathbf{I}_s$	A	0,978					BM_1	$\mathbf{I}_s$	A	0,978

Est: Estimador, BM_1 (BM_2): BM con varianza estimada 1 (2), E: Esquema.

Tabla 1: Mejores estimadores de regresión del total

es más frecuente que sobreestimen T. En el esquema B, BM tiene ligeramente menor SR excepto con $\mathbf{V}_s = \mathbf{\Pi}_s$. Los estimadores CC y AM tienen significativamente menor SR con el esquema A. En el escenario 33-33 y 95-95, los estimadores tienen menor ECMR con el esquema B, excepto AM. Por el contrario en los escenarios 33-95 y 95-33 menor ECMR ocurre con el esquema A excepto para BM. Generalmente BM tiene menor ECMR con el esquema B mientras que AM tiene siempre menor ECMR con el esquema A.

Solo catorce de noventa y seis estimadores tuvieron cuando mucho una desviación de 0.03 del valor nominal 0.95. De estos, doce estimadores usan $\mathbf{V}_s = \mathbf{I}_s$ y dos $\mathbf{V}_s = \mathbf{\Pi}_s$. En relación al enfoque, nueve estimadores usan el BM y cinco el AM. Mientras que nueve se obtuvieron con el Esquema A y cinco con el Esquema B (Tabla 1).

3. Conclusiones

Según los resultados obtenidos, los estimadores AM, BM y CC son igualmente eficientes en términos de sesgo relativo y error cuadrático medio relativo. Pero esto no es cierto en términos de su varianza estimada y por tanto de su cubrimiento. Apenas un quince por ciento de los estimadores, donde no aparecen los CC, tienen cubrimientos cercanos al nominal. El estimador BM fue ligeramente mejor que AM tal vez por que los datos se generan de acuerdo a un modelo. Se obtienen mejores resultados si $\mathbf{V}_s = \mathbf{I}_s$ es decir bajo el supuesto incorrecto de que las observaciones son independientes y que es usado en muchos paquetes estadísticos. El uso de $\mathbf{V}_s = \mathbf{G}_s$ tampoco produce mejores resultados lo que contradice a Godínez y Méndez (2008) aunque el uso de $\mathbf{V}_s = \mathbf{\Pi}_s$ tuvo mejores resultados sin ser satisfactorios. El esquema A fue mejor que el B tal vez por que las probabilidades de inclusión de las UPM

no fueron lo suficientemente diferentes (en el caso mas extremo la máxima probabilidad de inclusion fue 0.135 que es aproximadamente tres veces la mínima, 0.047). En el esquema A, a diferencia del esquema B, no se necesita conocer los tamaños de los conglomerados condición que a veces es difícil de cumplir. En tal situación es mejor usar el el esquema A dado que mejores resultados se obtienen en términos de cubrimiento.

Referencias

Brewer, K. 2002. Combined Survey Sampling Inference, Weighing Basu's Elephants. Arnold: London.

Dever, J. A. y Valliant, R. 2006. A comparison of Model-Based and Model-Assisted Estimators under Ignorable and Non-Ignorable Nonresponse. *Proceedings of the Survey Research Methods Section. American Statistical Association.*

Godínez, J. F. y Méndez, R. I. 2008. Desempeño en muestras complejas de tres estimadores de regresión del total. *Memorias del XXII Foro Nacional de Estadística.* INEGI: Mexico.

Raj, D. 1968. Sampling Theory. McGraw-Hill: New York.

Särndal, C.E., Swensson, B. y Wretman, J. 1992. Model Assisted Survey Sampling. Springer Verlag: New York.

Valliant, R., Dorfman, A.H. y Royall, R.M. 2000. Finite Population Sampling and Inference: A Prediction Approach. John Wiley & Sons.: New York.

Memoria FONE XXIII (2009), Sección II, 107–112

Una propuesta alternativa al algoritmo EM para la estimación máximo verosímil con datos incompletos

Ernesto Menéndez Acuña[a]
Facultad de Matemáticas – Universidad Veracruzana

Ernestina Castells Gil[b]
Área Económico Administrativo – Universidad Veracruzana

1. Introducción

Un método que ha resultado de mucha utilidad ante la situación de datos incompletos es el llamado algoritmo EM, acuñado por Dempster et al. (1977). Pero aún en situaciones donde la no completitud de los datos no es evidente o natural e incluso donde la muestra no es incompleta, este algoritmo puede ser usado.

A partir de la aparición del algoritmo EM muchas han sido las aplicaciones que se han realizado, generándose diferentes extensiones o modificaciones de dicho algoritmo. Se sugiere por ejemplo, revisar los trabajos de Dempster et al.(1977); Albert y Baxter (1995); Meng and Rubin (1993); Liu and Rubin (1994); Meng y van Dyk (1995). Una de las extensiones que llama nuestra atención es aquella conocida como algoritmo Monte Carlo EM (MCEM) (Wei y Tanner, 1990). Con este algoritmo se salva la dificultad que en ocasiones suele presentarse en la ejecución del paso E en el algoritmo EM. Ejemplos de aplicación de este algoritmo pueden verse en Sinha et al. (1994) y Chan y Ledolter(1995). La propuesta de este trabajo es un algoritmo del tipo de los llamados algoritmos de simulación iterativos. En la subsección (2.1) se explica la propuesta en detalle.

[a] emenendeza@gmail.com
[b] ernestinacg@yahoo.com

2. Marco teórico

Sea $\boldsymbol{x}$ la variable aleatoria que representa la información completa y $\boldsymbol{y}$ la información de la cual se dispone, es decir, la información incompleta; en consecuencia $\boldsymbol{x} = (\boldsymbol{z}, \boldsymbol{y})$, donde $\boldsymbol{z}$ es la información faltante. Sea $\boldsymbol{\theta}$ el parámetro que se desea estimar y $\widehat{\boldsymbol{\theta}} = \boldsymbol{h}(\boldsymbol{z}, \boldsymbol{y})$ la expresión para obtener la estimación máximo verosímil de $\boldsymbol{\theta}$ con la muestra completa.

El algoritmo EM (Dempster et al., 1977) es un algoritmo iterativo que consta de dos pasos:
Paso E: Determinar en la iteración k la función $Q(\boldsymbol{\theta}; \boldsymbol{\theta}^{(k)}) = E_{\boldsymbol{\theta}^{(k)}}[logp(\boldsymbol{\theta}; \boldsymbol{z}, \boldsymbol{y})|\boldsymbol{y})]$.
Paso M: Maximizar respecto a $\boldsymbol{\theta}$ la función $Q(\boldsymbol{\theta}; \boldsymbol{\theta}^{(k)})$. Sea este valor $\boldsymbol{\theta}^{(k+1)}$. Se repite el paso E con el valor del parámetro $\boldsymbol{\theta}^{(k+1)}$ y el paso M para hallar otro $\boldsymbol{\theta}^{(k+2)}$, y así sucesivamente hasta que el algoritmo converja. El algoritmo MCEM consta también de los pasos E y M, sólo que en el paso E lo que se determina es una función que aproxima a la función $Q(\boldsymbol{\theta}; \boldsymbol{\theta}^{(k)})$. Esta función está dada por $Q^*(\boldsymbol{\theta}, \boldsymbol{\theta}^{(k)}) = \frac{1}{M} \sum_{m=1}^{M} logp(\boldsymbol{\theta}; \boldsymbol{z}^{(m)}, \boldsymbol{y}; \boldsymbol{\theta}^{(k)})$. Como se aprecia en esta última expresión, el algoritmo MCEM genera M cantidades faltantes, incorporando cada una de ellas a la función $logp(\boldsymbol{\theta}; \boldsymbol{z^{(m)}}, \boldsymbol{y}; \boldsymbol{\theta}^{(k)})$, para posteriormente maximizar el promedio de estas funciones.

2.1. Algoritmo que se propone en este trabajo

La propuesta de este trabajo también considera la generación de la cantidad faltante M veces partiendo de un valor inicial del parámetro. El promedio de estas cantidades se utiliza para completar la muestra y estimar el valor del parámetro según la expresión $\widehat{\boldsymbol{\theta}} = \boldsymbol{h}(\boldsymbol{z}, \boldsymbol{y})$. Con este valor del parámetro se repite la generación de la cantidad faltante M veces y con el promedio de ellas se completa nuevamente la muestra y se estima el valor del parámetro como se indicó anteriormente. Este proceso se repite N veces y con el promedio de estas N repeticiones se determina la estimación del parámetro. Considerar el promedio de las M cantidades que se generan como la cantidad faltante, tiene como objetivo obtener una estimación de esta cantidad con una variabilidad más pequeña. De igual manera al tomarse como estimación de $\boldsymbol{\theta}$ el promedio de las N repeticiones se pretende brindar esta estimación con una menor variabilidad.

2.2. Simulaciones

Para evaluar el comportamiento del algoritmo que se ha propuesto se efectuaron algunas simulaciones. Todas ellas fueron realizadas con el paquete Matlab (versión 6.5. Release 13). El primer ejemplo considera una distribución multinomial con cuatro celdas. Este ejemplo es debido a Rao (1973, pp. 368-369) y utilizado por Dempster et al. (1977) y por Pawitan (2001, p. 343), para mostrar el funcionamiento del algoritmo EM, obteniendo como estimación máximo verosímil del parámetro el valor 0.627. McLachlan y Krishnan (1997, p. 215) y Wei y Tanner (1990) lo utilizaon para ilustrar el funcionamiento del algoritmo MCEM.

La evaluación no se limitó a la aplicación en este único caso, sino que se consideró el mismo modelo para diferentes frecuencias, probabilidades por celdas, valores de M, repeticiones (N) y valores iniciales para comenzar el algoritmo. En todos estos casos se obtuvieron resutados satisfactorios. En este ejemplo la estimación máximo verosímil del parámetro $\boldsymbol{\theta}$ en el paso k-ésimo del algoritmo, está dada por $\widehat{\theta}^{(k)} = \frac{z_2+y_4}{z_2+y_2+y_3+y_4}$, donde z_2 es la observación faltante. El valor de esta observación faltante en las M generaciones aleatorias que considera el algoritmo se realiza a partir de la ley de ditribución condicional $\boldsymbol{z}|\boldsymbol{y};\boldsymbol{\theta} \sim binomial(y_1, \frac{\frac{1}{4}\theta}{\frac{1}{2}+\frac{1}{4}\theta})$.

En la tabla 1 se muestran solamente los resultados obtenidos en la aplicación del ejemplo considerado por McLachlan y Krishnan (1997, p. 215) y Pawitan (2001, p. 343) por considerarse que son representativos de todos los resultados obtenidos en las simulaciones realizadas. Como se observa en la tabla 1 para valores de M iguales a 50 y 100 y valores de N iguales a 100, 300, 500 y 1000 la estimación del parámetro es muy adecuada en relación con el valor obtenido por Pawitan (2001) al aplicar al mismo ejemplo el algoritmo EM.

M	N	$\widehat{\theta}$	M	N	$\widehat{\theta}$
50	100	0.6269	100	100	0.6270
50	300	0.6268	100	300	0.6269
50	500	0.6268	100	500	0.6268
50	1000	0.6267	100	1000	0.6268

Tabla 1: Estimaciones ejemplo Multinomial

El segundo ejemplo que se considera es el de una mezcla de dos distribuciones Normales con media y varianza desconocidas. A partir de la modelación de un conjunto de n ob-

servaciones $y_1, \ldots, y_n$ iid se tiene que $p_\psi(y_i) = \sum_{j=1}^{n} \pi_j \phi_j(y_i|\mu_j, \sigma_j)$, donde π_1 y π_2 son las proporciones del total de observaciones que pertenecen a las poblaciones I y II respectivamente; ϕ la función de densidad de la distribución Normal con medias μ_1 y μ_2 y desviaciones típicas σ_1 y σ_2 correspondientes a las poblaciones I y II y ψ el vector de parámetros desconocidos $(\pi_1, \mu_1, \mu_2, \sigma_1, \sigma_2)$. Si se conociese a qué población pertence cada una de las observaciones, la estimación máximo verosímil de cada uno de los parámetros es inmediata. Así, una información muestral completa estaría dada por $(x_1, \ldots, x_n) = ((y_1, z_1), \ldots, (y_n, z_n))$, donde $z_i = 1$ indica que la observación y_i pertenece a la población I y $z_i = 0$, que pertenece a la población II. De donde las estimaciones de los parámetros se obtendrían mediante, $\widehat{\boldsymbol{\pi}}_{\mathbf{1}} = \frac{1}{n}\sum_{i=1}^{n} z_i$, $\widehat{\mu}_1 = \frac{1}{n}\sum_{i=1}^{n} z_i y_i$, $\widehat{\mu}_2 = \frac{1}{n-\sum_{i=1}^{n} z_i}\sum_{i=1}^{n}(1-z_i)y_i$, $\widehat{\sigma}_1 = \{\frac{1}{n}\sum_{i=1}^{n} z_i(y_i-\mu_1)^2\}^{\frac{1}{2}}$ y $\widehat{\sigma}_2 = \{\frac{1}{n-\sum_{i=1}^{n}}\sum_{i=1}^{n}(1-z_i)(y_i-\mu_2)^2\}^{\frac{1}{2}}$. La aplicación del algoritmo que se propone en este trabajo se implementa mediante el completamiento de la muestra, al generarse los n valores faltantes que indican a qué población pertenece cada una de las observaciones. La generación aleatoria de estos valores se logra con la distribución Bernoulli con probabilidad $p = \frac{\pi_1\phi_1(y_1;\mu_1,\sigma_1)}{\sum_{j=1}^{2}\pi_j\phi_j(y_i;\mu_j;\sigma_j)}$. En la tabla 2 se muestran los resultados obtenidos al aplicarse el algoritmo con los datos del ejemplo que considera Pawitan(2001, p. 350), para ilustrar el funcionamiento del algoritmo EM al caso de una mezcla de dos distribuciones.

Estimación por	EM	MCEMA				
Par. a estimar		N=50	100	300	500	1000
π_1	0.308	0.308	0.305	0.306	0.307	0.306
μ_1	54.203	54.220	54.118	54.155	54.179	54.152
σ_1	4.952	4.955	4.894	4.912	4.927	4.907
μ_2	80.360	80.308	80.290	80.324	80.339	80.316
σ_2	7.508	7.481	7.568	7.537	7.525	7.547

Tabla 2: Estimaciones mezcla de dos distribuciones

En esta ocasión se considera M = 1 y varios valores de las repeticiones. Las estimaciones de los 5 parámetro son muy parecidas a las logradas por la aplicación del algoritmo EM. En este caso no parece razonable generar varios vectores de ceros y unos para luego tomar el promedio de los diferentes elementos del vector y conformar el vector que indica la pertenencia de

cada observación a una de las dos clases, ya que este promedio no tiene necesariamente que producir un vector con valores ceros y unos.

3. Conclusiones

El algoritmo propuesto ha demostrado su factibilidad en las situaciones consideradas, al producir estimaciones satisfactorias del parámetro o de los parámetros de interés. La aplicación del algoritmo al ejemplo de mezcla de dos distribuciones demuestra de alguna manera su factibilidad en el caso donde exista más de un parámetro como objeto de estimación. Con esta propuesta, se evita también el cálculo de las funciones $Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(k)})$ $Q^*(\boldsymbol{\theta}, \boldsymbol{\theta}^{(k)})$ y su maximización, como sucede en la aplicación de los algoritmos EM y MCEM. No obstante, a tenor de los resultados es recomendable que se siga investigando sobre el valor de M a considerar cuando la situación permita un valor mayor que 1 y sobre la cantidad de repeticiones N.

Referencias

Albert, J.R.G., and Baxter, L.A. 1995. *Applications of the EM algorithm to the analysis of life length data*. Applied Statisitics , 44(3), 323-342.

Chan, K.S., and Ledolter, J. 1995. *Monte Carlo estimation for times series models involving counts*. Journal of the American Statistical Association, 90, 242-252.

Demspter, A., Laird, N.M., and Rubin, D.B. 1977. *Maximum likelihood from incomplete data via the EM algorithm (with discussion)*. Journal of the Royal Statistical Society, 39, 1-38

McLachlan, G.J., and Krishnan, T. 1997. *The EM algorithm and extensions*. New York: Wiley.

Liu, C., and Rubin, D.B. 1994. *The ECME algorithm: a simple extension of EM and ECM with faster monotone convergence*. Biometrika, 81, 633-648.

MATLAB Version 6.5. 2002. *The MathWorks, Inc. www.mathworks.com.*

Meng, X.L, and Rubin, D.B. 1993. *Maximum likelihood estimation via the ECM algorithm: a general framework*. Biometrika, 80, 267-278.

Meng, X.L., and van Dyk, D. 1995. *The EM algorithm - An old folk song sung to a fast new tune*. Technical Report, No. 408, Dept. of Statistics. Un. of Chicago.

Pawitan, Y. 2001. *In all likelihood*. New York: Oxford University Press.

Rao, C.R. 1973. *Linear statistical inference and its applications*. 2nd. ed. New York: Wiley.

Sinha, D., Tanner, M.A.,and Hall, W.J. 1994. *Maximization of the marginal likelihood of group survival data*. Biometrika, 81, 53-60.

Wei, G.C.G., and Tanner, M.A. 1990. *A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithms*. Journal of the American Statistical Association, 85, 699-704.

Memoria FONE XXIII (2009), Sección II, 113–118

Un modelo bayesiano para regresión circular–lineal

Gabriel Nuñez-Antonio[a]
Instituto Tecnológico Autónomo de México. Universidad Autónoma Metropolitana – Iztapalapa

Eduardo Gutiérrez-Peña
IIMAS – Universidad Nacional Autónoma de México

Gabriel Escarela
Universidad Autónoma Metropolitana – Iztapalapa

1. Antecedentes

La teoría para modelos de regresión cuando la respuesta es una variable circular no se ha desarrolado completamente, a pesar de que se pueden encontrar aplicaciones en varias áreas del conocimiento, particularmente en biología, geología y meteorología. La mayoría de los modelos propuestos en la literatura para una respuesta direccional presentan problemas como falta de identificabilidad y dificultades computacionales que los hacen difíciles de analizar en la práctica. Para una revisión sobre el tema el lector se puede referir, por ejemplo, a Nuñez-Antonio *et al.* (2008), Fisher (1993) y las referencias ahí incluidas.

2. El modelo

Sean $(\boldsymbol{x}_1, \boldsymbol{u}_1), \ldots, (\boldsymbol{x}_n, \boldsymbol{u}_n)$ observaciones independientes, donde $\boldsymbol{x}_i$ es un vector de covariables y $\boldsymbol{u}_i$ es la correspondiente variable circular, con dirección media $\boldsymbol{\eta}_i$ y longitud resultante media ρ_i, la cual puede depender de $\boldsymbol{x}_i$. Para datos circulares, θ_i y ω_i son las representaciones de $\boldsymbol{u}_i = (\cos\,\theta_i, \sin\,\theta_i)^t$ y $\boldsymbol{\eta}_i = (\cos\,w_i, \text{sen}\,w_i)^t$, respectivamente. El modelo normal

[a]gabriel@itam.mx

proyectado, $PN(\cdot|\boldsymbol{\mu} = \boldsymbol{B}^t\boldsymbol{x}, \boldsymbol{I})$, considera la siguiente representación (ver, Presnell *et al.*, 1998) : $\boldsymbol{U}_i = \boldsymbol{Y}_i/R$, con $R = \|\boldsymbol{Y}_i\|$, para $i = 1, ..., n$, donde $\boldsymbol{Y}_1, ..., \boldsymbol{Y}_n$ son vectores aleatorios independientes con distribución de probabilidad normal bivariada con vector de medias $\boldsymbol{\mu}_i = \boldsymbol{B}^t\boldsymbol{x}_i$ y matriz de varianzas y covarianzas la matrix identidad $\boldsymbol{I}$. Aquí $\boldsymbol{B}_{p\times 2} = (\boldsymbol{\beta}^1, \boldsymbol{\beta}^2)$ es la matriz de coeficientes de regresión. De esta manera, los componentes de $\boldsymbol{\mu}_i$ resultan ser $\mu_i^j = \boldsymbol{x}_i^t\boldsymbol{\beta}^j$ para $j = 1, 2$. Se debe notar que, dado que $\boldsymbol{U}$ es un vector aleatorio bidimensional, también puede ser definido a través de un sólo ángulo. La especificación del modelo queda completa proponiendo una distribución inicial $f(\boldsymbol{B})$ sobre la matrix de coeficientes de regresión.

El problema central se puede establecer de la siguiente manera: dado una muestra $\{\theta_1, \ldots, \theta_n\}$ de $PN(\cdot|\boldsymbol{\mu} = \boldsymbol{B}^t\boldsymbol{x}, \boldsymbol{I})$, y un conjunto de covariables $(\boldsymbol{x}_1, ..., \boldsymbol{x}_n)$, ¿cómo realizar inferencias sobre la matriz $\boldsymbol{B}$ y, si es necesario, cómo usar el modelo para otros fines tales como predicción?

2.1. Datos faltantes

Supóngase que una porción de los datos muestrales son faltantes. En nuestro caso, sea $\boldsymbol{z} = (\boldsymbol{\theta}, \tilde{\boldsymbol{\theta}})$ un conjunto de observaciones del modelo $\{PN(z|\boldsymbol{B}, \boldsymbol{I}),\ f(\boldsymbol{B})\}$, donde $\boldsymbol{\theta}$ y $\tilde{\boldsymbol{\theta}}$ denotan datos observados y no observados, respectivamente. Si consideramos $\tilde{\boldsymbol{\theta}}$ como un parámetro adicional, entonces se puede incluir en un esquema de muestreo de Gibbs con las densidades condicionales completas $f(\boldsymbol{\beta}^j|\theta_1, \cdots, \theta_s, \tilde{\theta}_{s+1}, \cdots, \tilde{\theta}_n)$ para $j = 1, 2$, $f(\tilde{\boldsymbol{\theta}}|\boldsymbol{B}, \tilde{\boldsymbol{x}}, \boldsymbol{\theta}, \boldsymbol{r}) = f(\tilde{\boldsymbol{\theta}}|\boldsymbol{B}, \tilde{\boldsymbol{x}}, \boldsymbol{r})$ y $f(\boldsymbol{r}|\boldsymbol{z}, \boldsymbol{B}, \tilde{\boldsymbol{x}})$. De esta manera se pueden generar observaciones de la distribución final $f(\boldsymbol{B}|\boldsymbol{\theta})$.

3. Inferencias vía MCMC

La propuesta en este trabajo está basada en la introducción de variables latentes apropiadas para definir una distribución conjunta aumentada con $\boldsymbol{B}$. Esta distibución conjunta se construye de tal forma que permita simular de todas las densidades condicionales requeridas para un muestreo de Gibbs.

Si $(\Theta_1, R_1), \ldots, (\Theta_n, R_n)$ pudieran ser observadas, entonces sería posible realizar inferencias sobre la matriz $\boldsymbol{B}$ de manera relativamente simple; sin embargo, el problema es que

realmente sólo se observan las direcciones $\{\theta_1, \ldots, \theta_n\}$. La estructura del modelo sugiere tratar los $R_i = ||\boldsymbol{Y}_i||$, $i = 1, \ldots, n$, no observados como variables latentes. Así, el modelo para los datos completos resulta ser el usual modelo de regresión normal multivariado. Esta fue la aproximación seguida por Presnell *et al.* (1998), quienes se enfocaron principalmente en la estimación de máxima verosimilitud de $\boldsymbol{B}$ vía algoritmos EM.

En el modelo de regresión normal multivariado, si $\boldsymbol{D}_n = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n)$ es una muestra de $N_2(\cdot|\boldsymbol{\mu} = \boldsymbol{B}^t\boldsymbol{x}, \boldsymbol{I})$ y se considera la inicial conjugada $N_p(\boldsymbol{\beta}^j|\boldsymbol{\beta}_0^j, \boldsymbol{\Lambda}_0^j)$, entonces la distribución final de $\boldsymbol{\beta}^j$ resulta ser $f(\boldsymbol{\beta}^j|\boldsymbol{D}_n) = N_p(\cdot|\ \boldsymbol{\mu}_F^j, \boldsymbol{\Lambda}_F^j)$, para $j = 1, 2$. Si ahora se considera la variable latente R definida en $(0, \infty)$, desde $\boldsymbol{Y} \sim N_2(\cdot|\boldsymbol{\mu} = \boldsymbol{B}^t\boldsymbol{x}, \boldsymbol{I})$ se puede definir su densidad conjunta con Θ como

$$f(\theta, r|\boldsymbol{\mu} = \boldsymbol{B}^t\boldsymbol{x}) = (2\pi)^{-1} \exp\{-\tfrac{1}{2}||\boldsymbol{\mu}||^2\} \exp\{-\tfrac{1}{2}[r^2 - 2r(\boldsymbol{u}^t\boldsymbol{\mu})]\}|J|, \tag{1}$$

donde $|J| = r^2$ es el Jacobiano de la transformación $\boldsymbol{y} \to (\theta, r)$ y $\boldsymbol{u}_i = (\cos\theta_i, \sin\theta_i)^t$.

3.1. Distribuciones condicionales

La densidad condicional de $\boldsymbol{\beta}^j$ está dada por $f(\boldsymbol{\beta}^j|\theta_1, \ldots, \theta_n, \boldsymbol{r}) = N_p(\cdot|\boldsymbol{\mu}_F^j, \boldsymbol{\Lambda}_F^j)$, donde $\boldsymbol{r} = (r_1, \ldots, r_n)$ es un vector n-dimensional.

Se debe notar que las R_i $(i = 1, \ldots, n)$ son condicionalmente independientes dado las Θ_i $(i = 1, \ldots, n)$. Así, de (1) se tiene que $f(r|\theta_i, \boldsymbol{\mu}_i = \boldsymbol{B}^t\boldsymbol{x}_i) \propto r^2 \exp\{-\frac{1}{2}r^2 + b_i\, r\} 1_{(0,\infty)}(r)$, donde $b_i = \boldsymbol{u}_i^t\boldsymbol{\mu}_i$. De esta manera, se pueden generar R_i de $f(r|\theta_i, \boldsymbol{\beta}^1, \boldsymbol{\beta}^2)$ de tal forma que es factible simular un vector aleatorio $\boldsymbol{R}$ de $f(\boldsymbol{r}|\theta_1, \ldots, \theta_n, \boldsymbol{\beta}^1, \boldsymbol{\beta}^2)$. De hecho, este último paso se lleva a cabo vía un algoritmo de Metropolis-Hastings. La versión del algoritmo considerada fue la de *independencia*, es decir, se utilizó una distribución de transición $N(|\hat{\boldsymbol{\vartheta}}, \kappa\boldsymbol{V}(\hat{\boldsymbol{\vartheta}}))$, con $\kappa \geq 1.0$ un factor de sobredispersión. Aquí, $\hat{\boldsymbol{\vartheta}}$, $\boldsymbol{V}(\hat{\boldsymbol{\vartheta}})$ denotan a la media y a la varianza de la aproximación normal asintótica para la distribución final de $ln(\boldsymbol{r})$.

Finalmente, se pueden usar las condicionales completas $f(\boldsymbol{\beta}^j|\cdot)$ y $f(\boldsymbol{r}|\cdot)$ en un muestreo de Gibbs para obtener una muestra de la distribución conjunta final $f(\boldsymbol{B}, \boldsymbol{r}|\theta_1, \ldots, \theta_n)$. De lo anterior, se sigue que la muestra de $\boldsymbol{B}$ obtenida de esta manera tiene la distribución (marginal) requerida $f(\boldsymbol{B}|\theta_1, \ldots, \theta_n)$.

3.2. Datos faltantes

En el modelo para datos completos la densidad condicional completa del vector $\boldsymbol{\beta}^j$ se define por $f(\boldsymbol{\beta}^j|(r,\theta)_1,\ldots,(r,\theta)_s,(\tilde{r},\tilde{\theta})_{s+1},\ldots,(\tilde{r},\tilde{\theta})_n) = N_2(\cdot|\boldsymbol{\mu}_F^j,\boldsymbol{\Lambda}_F^j)$. En este caso, es necesario determinar las densidades condicionales completas para $\tilde{\theta}$ y todas las variables latentes $\tilde{r}_i$. Sin embargo, se puede tomar ventaja del esquema del muestreador de Gibbs para generar observaciones conjuntas de $(\tilde{r},\tilde{\theta})$. Lo anterior, reconociendo después de una transformación apropiada que la densidad condicional completa de $(\tilde{r},\tilde{\theta})$, $f((\tilde{r},\tilde{\theta})|\boldsymbol{B},\tilde{\boldsymbol{x}})$, es el modelo normal bivariado, y por lo tanto $f((\tilde{r},\tilde{\theta})|\boldsymbol{B},\tilde{\boldsymbol{x}}) = N_2(\tilde{r}\tilde{\boldsymbol{u}}|\boldsymbol{\mu} = \boldsymbol{B}^t\tilde{\boldsymbol{x}},\boldsymbol{I})$. La densidad condicional completa de las R's asociadas a los ángulos realmente observados está dada por $f(\boldsymbol{r}|\cdot)$.

4. Ejemplo

Se utilizó el modelo de regresión propuesto para analizar un conjunto de datos reales de tamaño 31 referentes a las distancias en centímetros (x) y direcciones (θ) tomadas por cierta especie de caracoles de mar, después de que ellos fueron movidos de la altura a la cual normalmente viven (ver Fisher, 1993, Apéndice B20). Se tomó la siguiente especificación propuesta por Presnell *et al.*, (1998), la cual se define como $\boldsymbol{\mu} = (\mu_1,\mu_2)^t = (1+x,1+x)^t$.

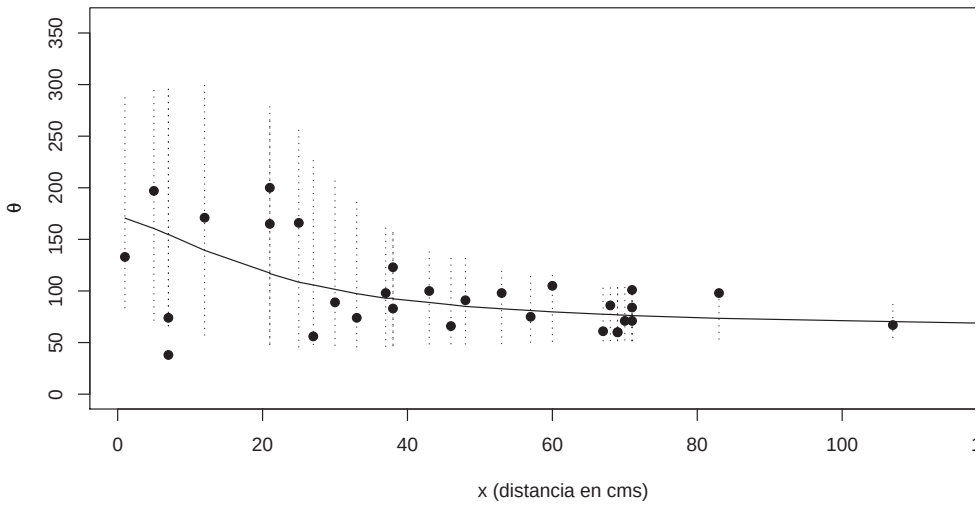


Figura 1: *Intervalos predictivos finales al 95 % para los datos de caracoles. La línea continua representa la dirección media ajustada.*

Se consideró $\boldsymbol{\beta}_0^j = \mathbf{0}$, $\boldsymbol{\Lambda}_0^j = \text{Diag}(0.0001, 0.0001)$ para $j = 1, 2$, $\boldsymbol{r} = (1, ..., 1)$ y un factor de sobredispersión de $\kappa = 1.0$. El esquema de Gibbs se implementó con una fase de calentamiento en la que se consideraron valores de R_i dados por su esperanza condicional final, $E(r_i|\boldsymbol{u}_i^t\boldsymbol{\mu}_i)$, para toda $i = 1, ..., n$. Posteriormente, para cada iteracion del Gibbs sampler, el algoritmo de Metropolis-Hastings se itero cinco veces para cada variable R_i. Los resultados de convergencia y autocorrelación para este proceso resultaron satisfactorios. Además, el desempeño del algoritmo resultó eficiente en términos del tiempo de convergencia, que para este ejemplo fue de 30 minutos en un procesador con 1GB de RAM a 2.0 Ghz.

Las distribuciones finales marginales obtenidas para cada componente de los vectores $\boldsymbol{\beta}^1$ y $\boldsymbol{\beta}^2$ concuerdan con los resultados obtenidos (específicamente, los estimadores puntuales de los coeficientes de regresión así como sus correspondientes varianzas) por los análisis previos de estos datos. En la Figura 1 se muestran los intervalos predictivos finales al 95 % para cada valor de x, así como, la dirección media ajustada con el correspondiente modelo normal proyectado.

Con el fin de ilustrar el análisis para el problema de datos faltantes supóngase que las respuestas $\theta_{13} = 197°$, $\theta_{23} = 75°$ y $\theta_1 = 67°$ con $x_{13} = 5$, $x_{23} = 57$ y $x_1 = 107$, respectivamente, son datos faltantes. Utilizando los resultados de las secciones previas se obtuvieron las distribuciones finales para θ_{13}, θ_{23} y θ_1. Estas distribuciones finales se presentan en la Figura 2. Se puede notar que, con la metodología propuesta, se obtienen inferencias apropiadas para el problema de datos faltantes.

5. Conclusiones

En este trabajo se presenta un análisis Bayesiano completo de un modelo de regresión para datos circulares basado en la distribución normal proyectada. Aunque la versión de la normal proyectada considerada no es la más general, ésta es bastante flexible y no pierde aplicabilidad práctica comparada con los modelos de regresión que suponen una distribución von Mises para la variable respuesta. Además, esta metodología se puede llevar a cabo de manera relativamente simple vía el muetreo de Gibbs. Se debe resaltar que el problema de datos faltantes se puede analizar de una manera natural y simple bajo el análisis Bayesiano propuesto. Adicionalmente, a diferencia de la mayoría de los análisis actualmente disponibles de modelos de regresión para datos circulares, la propuesta que se presenta en este trabajo

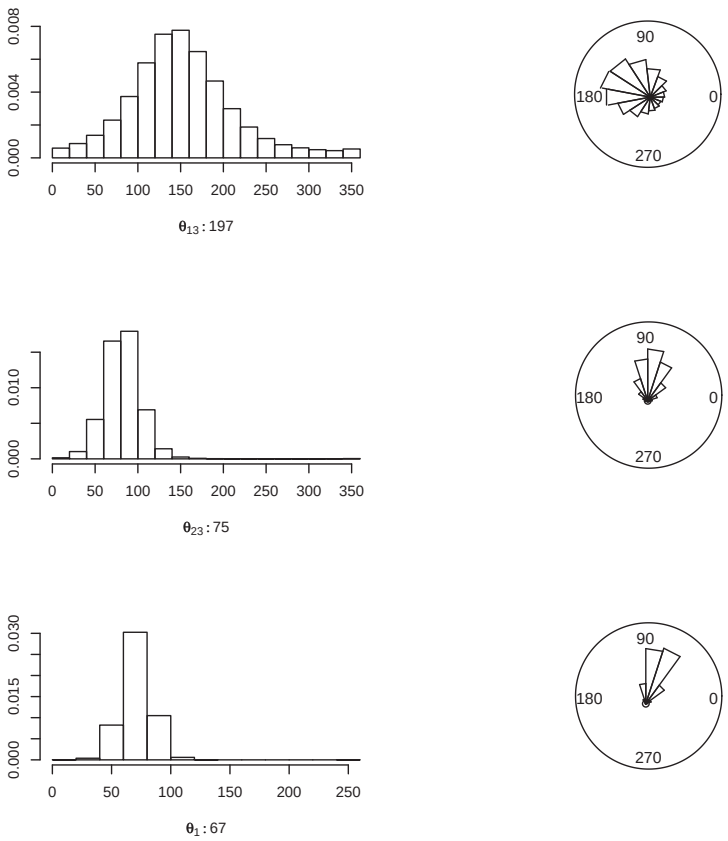


Figura 2: *Distribuciones finales de los valores faltantes θ_{13} (en $x = 5$), θ_{23} (en $x = 57$) y θ_1 (en $x = 107$), para los datos de caracoles.*

es computacionalmente fácil de implementar.

Referencias

Fisher, N.I. 1993. *Statistical Analysis of Circular Data*. Cambridge: University Press.

Nuñez-Antonio, G., Gutiérrez-Peña, E. y Escarela, G. 2008. A Bayesian regression model for circular data based on the projected normal distribution. Documento de trabajo del *Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas*, **145**.

Presnell, B., Morrison S.P. y Littell, R.C. 1998. Projected multivariate linear model for directional data. *Journal of the American Statistical Association*, **93**, 443, 1068-1077.

Memoria FONE XXIII (2009), Sección II, 119–124

El análisis multivariado aplicado al cultivo del cirián

Emilio Padrón Corral[a]
Universidad Autónoma de Coahuila

Ignacio Méndez Ramírez
IIMAS – Universidad Nacional Autónoma de México

Armando Muñoz Urbina
Universidad Autónoma Agraria Antonio Narro

1. Introducción

La estadística multivariada es útil para examinar matrices de datos sobre todo cuando se estudian gran número de variables y la inspección visual es poco práctica. En este experimento se trabajó con la especie arbórea Cirián (Crescentia alata H.B.K) la cual se utiliza como tratamiento medicinal para controlar enfermedades, es originaria de México y se cultiva en los estados de: Michoacán, Colima, Guerrero, Jalisco y Nayarit; los datos se obtuvieron de un experimento que se realizó en el área denominada, El Llano, Municipio de Coahuayana, Michoacán, México; Avila (1999), formando 30 cuadrantes en una superficie de 120 hectáreas, con 279 árboles muestreados. Las variables a medir fueron: altura del árbol , número de ramas, diámetro de ramas, cobertura, diámetro ecuatorial, diámetro polar, número de frutos, peso de frutos y rendimiento; se observa que el rendimiento de fruto por hectárea se obtuvo al multiplicar los siguientes factores: Densidad de árbol y peso promedio de frutos por árbol.

El objetivo es desarrollar un análisis de correlación canónica, a fin de determinar aquellas variables que más influyen en la producción y un análisis de coeficientes de sendero para estudiar las relaciones entre las componentes del rendimiento y características relacionadas.

[a]epadron@mate.uadec.mx

2. Metodología

Para el análisis de correlación canónica, la matriz de datos originales se particionó de acuerdo a la naturaleza de sus variables en dos submatrices, cada una correspondiente a dos grupos, esto es:

$X = [X, Y]$ tal que { $X = [x_1, x_2, ..., x_p]$ vector p-dimensional, $Y = [y_1, y_2, ..., y_q]$ vector q-dimensional} (1)

Los conjuntos de variables X y Y son descritos por dos conjuntos de k nuevas variables U y V obtenidos como combinaciones lineales de los x y y variables, respectivamente, tal que en cada par de nuevas variables U_i y V_i el coeficiente de correlación $r_i(U_iV_i)$ sea máximo bajo la restricción de que las varianzas sean iguales a uno, esto es:

$$Var(U_i) = Var(V_i) = 1 \qquad \text{para } i = 1, 2, ..., k$$
$$r_1(U_1, V_1) \geq r_2(U_2, V_2) \geq ... \geq r_k(U_k, V_k)$$

Los elementos esenciales en este procedimiento son clasificados como sigue:

$r_i(U_i, V_i)$ coeficiente de correlación canónico entre las variables U_i y V_i (1)

$U_i = \sum_j W_{1ij} X_j; V_i = \sum_l W_{2il} X_l;$ variables canónicas $i = 1, 2, ..., k;\ k \leq min(p, q)$ (2)

Los coeficientes del análisis de sendero con efectos directos e indirectos, fueron estimados de acuerdo a Wright (1934), los que posteriormente fueron descritos por Dewey y Lu (1959) y por Li (1975). Wright, ideó la manera de interpretar ecuaciones normales para resolver coeficientes de regresión estandarizados en problemas de regresión múltiple. El análisis de sendero o método de coeficientes de sendero, es una forma de análisis de regresión estructurado, varios modelos de regresión ligados, y considerando variables estandarizadas a media cero y varianza uno, en un sistema cerrado.

Es decir, los efectos directos son coeficientes de regresión estandarizados que aplicados al mejoramiento de plantas permite un avance más rápido en la selección de genotipos sobresalientes en la variable de estudio, los cuales son denominados coeficientes de sendero (b), (denotado por una línea con una flecha); cada variable predictora tiene un efecto directo y un efecto indirecto para cada una de las otras variables asociadas. El efecto indirecto es aquel que se obtiene a traves de otras variables y se estima del producto del coeficiente de

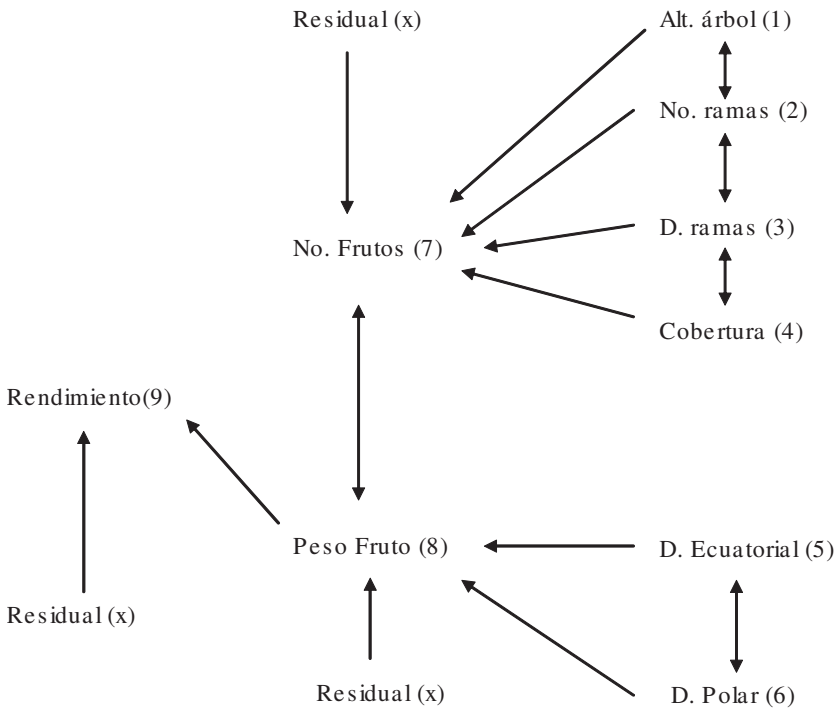


Figura 1: Diagrama o modelo gráfico de sendero mostrando las correlaciones ($\longleftrightarrow$) y los efectos directos ($\longrightarrow$) entre variables de árbol y fruto en el cultivo del Cirián.

correlación y su respectivo efecto directo y nos permite detectar su efecto correspondiente en la variable de estudio. (denotado por una línea con dos flechas). Figura 1, Los datos se analizaron con el paquete computacional MATLAB.

En el Cuadro 1. Se presenta el análisis de correlación canónica; donde se estimó una correlación altamente significativa ($R = 0.985 **$) entre el primer par de variables canónicas (U_1, V_1) por lo tanto, el primer par de variables canónicas tienen la más alta correlación posible y es por lo tanto la más importante. El segundo par de variables canónicas (U_2, V_2) presentaron una correlación significativa ($R = 0.663*$) y es la segunda en importancia. Para los restantes pares de variables canónicas (U_3, V_3) y (U_4, V_4) no se obtuvieron correlaciones estadísticas significativas.

Las correlaciones entre U_1 y sus cinco variables X (Cuadro 2) muestra que U_1 esta negativamente correlacionada con cobertura y número de frutos por árbol. En las correlaciones entre V_1 y sus cuatro variables Y, se observa que V_1 esta negativamente asociada con rendimiento de fruto; interpretando a U_1 y V_1 basados en sus correlaciones, esto suguiere que la

Raíz removida	R	R^2	χ^2cal	g.l.	Probabilidad
0	$0.985**$	0.970	109.77	20	$0.000**$
1	$0.663*$	0.439	24.82	12	$0.015*$
2	0.480	0.230	10.91	6	0.090
3	0.418	0.175	4.61	2	0.090

*, ** Significativo al 5 % y al 1 % de probabilidad, respectivamente.

Tabla 1: Análisis de correlación canónica y prueba de χ^2 con raíces sucesivas removidas.

Variable grupo X	U_1	U_2	U_3	U_4
Altura de árbol	−0.3431	−0.1433	0.5000	−0.4310
Número de ramas	−0.2854	0.3227	0.6688	0.5198
Diámetro de ramas	−0.3303	−0.1388	−0.1259	−0.9174
Cobertura	−0.7659	0.5443	0.1184	−03186
Número de frutos/árbol	−0.9998	0.0044	0.0010	0.0180
Variable grupo Y	V_1	V_2	V_3	V_4
Diámetro ecuatorial	−0.0576	0.3293	0.9335	0.1293
Diámetro polar	0.0540	0.1926	0.5375	0.8193
Peso de fruto	−01679	−0.2732	0.8215	0.1828

Tabla 2: Correlaciones entre las U_i y sus variables; y entre las V_i y sus varibales.

disminución en la cobertura y en número de frutos por árbol esta asociado con una disminución en el rendimiento de fruto.

3. Conclusiones

1. El análisis de correlación canónica mostró que los dos grupos de variables (de árbol y fruto) se encuentran relacionados. Estimándose una correlación altamente significativa entre el primer par de variables canónicas ($R = 0.985**$) y una correlación significativa entre el segundo par de variables canónicas ($R = 0.663*$).

2. De la relación entre el primer par de variables canónicas (U_1, V_1) se desprende que un

Caracter	Altura de árbol	Número de ramas	Diámetro de ramas	Cobertura	Correlación con número de frutos
a)					
Altura de árbol	**0.0875**	-0.0023	-0.0111	0.2589	0.333
Número de ramas	-0.0036	**0.0552**	0.0136	0.2327	0.298
Diámetro de ramas	0.0340	-0.0262	**-0.0282**	0.3338	0.313
Cobertura	0.0311	0.0177	-0.0131	**0.7273**	0.763**
Residual=0.6391					
$R^2 = 1 - (0.6391)^2 = 0.59$					
Caracter	Diámetro ecuatorial	Diámetro polar	Correlación con peso de fruto		
b)					
Diámetro ecuatorial	**0.4255**	0.3225	0.748**		
Diámetro polar	0.2842	**0.4828**	0.767**		
Residual=0.5581					
$R^2 = 1 - (0.5581)^2 = 0.69$					
Caracter	Número fruto	Peso de fruto	Correlación con rendimiento		
c)					
Número de frutos	**0.8938**	0.0.492	0.943**		
Peso de fruto	0.1511	**0.2909**	0.442**		
Residual=0.1689					
$R^2 = 1 - (0.5581)^2 = 0.69$					

Tabla 3: Efectos directos (negreado) e indirectos del análisis de coeficientes de sendero para: a)número de frutos; b) peso de fruto; c) rendimiento de fruto, en árbol de Cirián.

decremento en el número de frutos repercutirá en un decremento en el rendimiento de fruto. En el segundo par de variables canónicas (U_2, V_2) se observa que un incremento en la cobertura esta asociado a una disminución en el peso del fruto.

3. El análisis de coeficientes de sendero mostró que la cobertura fue un factor importante para determinar el número de frutos.

4. Los efectos directos de diámetro ecuatorial y diámetro polar sobre peso de fruto, manifestaron buena relación, y explicaron en un 69 porciento la variación en el peso de fruto.

5. El análisis de sendero para rendimiento de fruto, muestra que el incremento en el número de frutos es el factor más importante para mejorar el rendimiento de fruto por árbol. El coeficiente de determinación fue alto y muestra que el número de frutos y el peso de fruto explicaron en un 97 porciento la variación en el rendimiento de fruto.

Referencias

Avila, R.A. 1999. *Ecología y Evaluación del Fruto del Cirián (Crescentia alata H.B.K.) Como Recurso Forrajero en la Localidad el Llano, Municipio de Coahuayana, Michoacán, México.*. Universidad Autónoma Agraria Antonio Narro, Buenavista, Saltillo, Coahuila, México.

Dewey, D.R. y Lu, K.H. 1959. *A Correlation and Path Coefficient Analysis of Components of Crested Wheatgrass Seed Production.*Agronomy Journal.

Li, C.C. 1975. *Path Analysis: A Primer. Boxwood Press, Pacific Grove, C.A.*

Manly, B.F.J. 1986. *Multivariate Statistical Methods A Primer. Chapman and Hall.* Universidad Autónoma Agraria Antonio Narro, Buenavista, Saltillo, Coahuila, México.

MATLAB *The Language of Technical Computing. Version 7.0.0. 19920 (R14). Copyright (1984-2004). The MathWorks. Inc.*

Memoria FONE XXIII (2009), Sección II, 125–129

Cuantificación de variables categóricas mediante análisis multivariante no lineal

Mª Carmen Patino Alonso[a], Purificación Vicente Galindo, Elena Vicente Galindo, Purificación Galindo Villardón
Universidad de Salamanca – España

1. Introducción

En las ciencias sociales, económicas, de la salud y del comportamiento, es muy frecuente encontrarse con variables de carácter mixto, es decir, que suelen poseer diferentes escalas de medición: nominal, ordinal, de intervalo y/o de razón.

Para los casos en los que se tengan variables expresadas en distintos niveles de medida y sea necesario utilizar una técnica estadística que requiera datos cuantitativos es posible cuantificar las variables categóricas tomando como escenario el sistema Gifi.

2. Sistema Gifi de análisis multivariante no lineal

Aunque hay intentos de cuantificación desde los años 40, Guttman (1941), Torgerson (1958), Hayashi (1950) y De Leeuw (1973), es a principios de la década de los 80 cuando se hace referencia al proceso de cuantificación como escalamiento óptimo por Young en un trabajo titulado Análisis Cuantitativo de Datos Cualitativos (Young, 1981), pero es en 1981 cuando la escuela holandesa por parte del grupo de investigación en Teoría de Datos de la Facultad de Ciencias Sociales de la Universidad de Leiden, bajo el pseudónimo de Albert Gifi, publica el libro *Nonlinear Multivariate Analysis* donde explica con detalle los principios y aplicaciones del escalamiento óptimo y amplía los avances sobre métodos multivariantes no lineales, mostrándolos como componentes de un sistema. Posteriormente el texto fue reeditado en 1990 (Gifi, 1981, 1990).

[a]carpatino@usal.es

Los autores denominan **escalamiento óptimo** a cualquier técnica generadora de transformaciones que minimicen una función de pérdida.

El análisis de homogeneidad (HOMALS) es uno de los modelos básicos de la familia del Escalamiento Óptimo del sistema Gifi, el cual comprende una serie de técnicas exploratorias de análisis multivariante no lineal, extensiones del Análisis en componentes Principales y de Correlación Canónica al caso de variables nominales.

Tiene como objetivo la representación de la estructura de datos multivariantes categóricos, describiendo las relaciones entre dos o más variables nominales en un espacio de pocas dimensiones que contiene las categorías de las variables así como los individuos pertenecientes a dichas variables.

Sea X la matriz N x p, que contiene las coordenadas de los N individuos e Y la matriz M x p, que contiene las coordenadas de las M categorías. Llamaremos a X la "matriz de cuantificaciones de los objetos" e Y la "matriz de cuantificaciones de las categorías".

El objetivo es minimizar la función de pérdida simultáneamente respecto a X (matriz de las coordenadas de los objetos) y las Y (matriz de las coordenadas de las categorías de una variable) que se define como:

$$\sigma\left(X;Y\right)=\sum_{i=1}^{N}\sum_{j=1}^{M}\left[G_{ij}\sum_{k=1}^{p}\left(X_{ik}-Y_{jk}\right)^{2}\right] \tag{1}$$

donde G es la matriz indicadora (los datos se codifican mediante matrices indicadoras). Utilizamos el algoritmo del método de los Mínimos Cuadradros Alternados (ALS) (Michailidis y De Leeuw, 1998) para minimizar la función de pérdida. Se imponen restricciones sobre las cuantificaciones de las categorías y en algunos casos sobre la codificación de los datos. Las restricciones son:

$$X'X=NI_{p} \tag{2}$$

$$u'X=0 \tag{3}$$

donde u es el vector unitario. La primera restricción es para evitar soluciones triviales referente a X=0 e Y=0. La segunda restricción de normalización requiere que el gráfico de coordenadas sea centrado alrededor del origen. El algoritmo ALS itera en los siguientes pasos hasta que converge:

Primero, la función de pérdida es minimizada con respecto a Y para ajustar X. La ecuación normal es:

$$CY = G'X \tag{4}$$

donde G es la matriz transpuesta de G, C es la matriz diagonal que contiene las sumas de la columna de G. La solución de (4) es:

$$\widehat{y} = C^{-1}G'X \tag{5}$$

Segundo, la función de pérdida es minimizada con respecto a X para fijar Y. La ecuación normal es:

$$RX = GY \tag{6}$$

donde R es la matriz diagonal que contiene la suma de las filas de G. Por tanto, lo que conseguimos es:

$$\widehat{X} = R^{-1}GY \tag{7}$$

Tercero, las coordenadas están centradas y normalizadas por el procedimiento Gram-Schmidt (Golub y Van Loan, 1989).

$$X = \sqrt{N}GRAM(W) \tag{8}$$

donde

$$W = \widehat{X} - u\left(\frac{u'\widehat{X}}{N}\right) \tag{9}$$

Esta solución es denominada solución HOMALS (Homogeneity Analysis by Means of Alternating Least Squares).

2.1. Aplicación práctica

En este trabajo hemos utilizado el método HOMALS para cuantificar las categorías de las variables nominales y/o ordinales y para describir las relaciones entre esas variables en un subespacio de baja dimensión, de tal forma que la homogeneidad sea máxima. El procedimiento se aplica a la búsqueda del perfil multivariante de mujeres en situación laboral irregular en España, entendiendo por tal que no cumplen con las cotizaciones a la Seguridad Social.

El estudio se ha realizado sobre la información recogida en 3100 encuestas realizadas en 144 municipios de la provincia de Salamanca (Castilla y León, España), utilizando un muestreo multietápico en el cual la etapa final fue un muestreo en bola de nieve, proporcional. Los estratos se han elegido utilizando un HJ-BIPLOT (Galindo, 1986) sobre los datos tomados del Anuario Social de España. Los detalles sobre cómo se ha llevado a cabo el muestreo y la descripción de las variables, ya han sido publicados y pueden ser consultados en Galindo et al. (2007).

Para conocer las variables con mayor influencia en la clasificación, analizamos los índices de discriminación; es decir los ítems que más han contribuido a la formación de los clusters. El análisis de las contribuciones pone de manifiesto que las diferencias entre los clusters 1, 2 y 5 que son los que están más diferenciados por el eje I, se deben a las diferencias en las respuestas a las variables **i1** En que trabaja actualmente, con una contribución de 0,581 (son valores acotados entre cero y uno), **i4** Nivel de estudios, con una contribución de 0,354, **i8** Cotiza a la Seguridad Social, con una contribución de 0,584, **i9** Tiene contrato, con una contribución de 0,515, **i17** Ingresos percibidos por su trabajo, con una contribución de 0,483, **i20** Tipo de tarjeta sanitaria, con una contribución de 0,322, **i25.6** Trabajo en otros, con una contribución de 0,447 e **i25.7** Empleada de Hogar, con una contribución de 0,308.

Las respuestas que más han influido en la diferenciación de los clusters que son los que más se diferencian por eje 2 son las siguientes: **i3** Estado civil, con una contribución de 0,717, **i10** Edad, con una contribución de 0,599, **i11** Donde vive actualmente, con una contribución de 0,547, **i15.4** Otros familiares realizan las tareas del hogar, con una contribución de 0,451, **i16.1** Mi marido tiene trabajo remunerado, con una contribución de 0,485 e **i16.4** Mis padres tienen trabajo remunerado, con una contribución de 0,675.

Realizando un análisis de conglomerados de K-medias sobre las variables categóricas cuantificadas, ha sido posible describir 5 clusters bien diferenciados: Por un lado, tres constituidos por mujeres regularizadas: las que tienen estudios primarios o carecen de ellos; las jóvenes que deciden concluir sus estudios e incorporarse al mundo laboral; y las mujeres con un alto nivel académico bien situadas en la sociedad. Y por otro, dos bloques de mujeres irregulares: el constituido por mujeres jóvenes españolas e inmigrantes que buscan un complemento monetario para sus gastos personales, pero dependen económicamente de sus padres y viven con ellos, desempeñando normalmente su actividad en el sector de la hostelería; y el

grupo de mujeres españolas e inmigrantes dedicadas al servicio doméstico.

Referencias

De Leeuw, Jan. 1973. *Canonical analysis of categorical data*. Leiden (The Netherlands): University of Leiden.

Galindo Villardón, Mª Purificación. 1986. Una alternativa de representación simultánea: HJ-Biplot. *Questiió*, **10** (1), 13-23.

Galindo Villardón, Mª Purificación, Purificación Vicente Galindo, Mª Carmen patino Alonso y Jose Luis Vicente Villardón. 2007. Caracterización Multivariante de los perfiles de las mujeres en situación laboral irregular: el caso de Salamanca, *Pecunia*, **4** (1), 49-79.

Gifi, Albert. 1981.*Nonlinear Multivariate Analysis*. Leiden (The Netherlands): University of Leiden.

Gifi, Albert. 1990.*Nonlinear Multivariate Analysis*. Chichester (England): John Wiley & Sons.

Golub, Gene Howard y Charles Francis Van Loan. 1989.*Matrix Computations*. Baltimore: Johns Hopkins University Press.

Guttman, Louis. 1941. The quantification of a class of attributes: A theory and method of scale construction. In: P. Horst (Ed.).*The Prediction of Personal Adjustment*. New York: Social Science Research Council, 319-348.

Hayashi, Chushiro. 1950. On the quantification of qualitative data from the mathematico-statistical point of view.*Annals of the Institute of Statistical Mathematics*, **2** (1), 35-47.

Michailidis, George y Jan De Leeuw. 1998. The Gifi System of Descriptive Multivariate Analysis.*Statistical Science*, **13**, 307-336.

Torgerson, Warren. 1958. *Theory and methods of scaling*. New York: Wiley.

Young, Forrest William. 1981. Quantitative analysis of qualitative data.*Psychometrika*, **46** (4), 357-388.

Memoria FONE XXIII (2009), Sección II, 131–135

Programación cuadrática usando la técnica de ramificación y acotamiento

Blanca Rosa Pérez Salvador[a]
Universidad Autónoma Metropolitana – Iztapalapa

1. Introducción

El modelo de regresión lineal con restricciones lineales sobre los parámetros, se formula como $Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_m X_m + \varepsilon$ con $R\boldsymbol{\beta} \leq r$, donde R es una matriz de rango completo. Los estimadores de este modelo se encuentran mediante un problema de programación cuadrática. De igual manera, encontrar la respuesta óptima en la metodología de superficie de respuesta, cuando en el sistema existen restricciones lineales sobre los factores, se formula como: estimar el óptimo de $Y = X^T BX + b^T X + \beta_0 + \varepsilon$, sujetos a la restricción $RX \leq r$. También se resuelve como un problema de programación cuadrática. Por esta razón, es importante estudiar el problema de programación cuadrática para resolver algunos problemas en estadística. El problema de programación cuadrática ha sido estudiado por diferentes autores como Escobar, et al (1984) y Quintana et al (1987) quienes utilizaron el algoritmo del cono para encontrar la solución; sin embargo, estos procedimientos no llegan a la solución óptima en el 100 % de los casos. En este trabajo se explora la aplicación de la técnica de ramificación y acotamiento en la solución del problema de programación cuadrática. Esta técnica es un método éxacto de optimización el cual permite revisar sólo algunas soluciones de manera sistemática hasta llegar a la solución óptima. El trabajo consta de cinco secciones: la primera es esta introducción, en la segunda se presenta el problema de programación cuadrática, en la tercera sección se presenta la técnica de ramificación y acotamiento, en la sección cuarta se presentan los elementos para adaptar la técnica de ramificación y acotamiento al problema de programación cuadrática y en la sección 5 se presentan las conclusiones.

[a]psbr@xanum.uam.mx

2. El problema de programación cuadrática

El Problema de programación cuadrática en el caso convexo se formula así:

$$\text{Minimizar} \qquad H(\mathbf{x}) = \mathbf{x}^T B\mathbf{x} + \mathbf{b}^T\mathbf{x} + b_0, \qquad \text{sujeto a} \qquad R\mathbf{x} \leq \mathbf{r} \tag{1}$$

donde $\mathbf{x}, \mathbf{b} \in \mathbf{R}^n$; $\mathbf{r} \in \mathbf{R}^m$; $b_0 \in \mathbf{R}$, la matriz $B_{n\times n}$ es positiva definida y $R_{m\times n}$ $(m \leq n)$ es de rango completo. El mismo problema se reduce a su forma canónica mediante una transformación de translación y reescalamiento, de esta manera queda como:

$$\text{Minimizar} \qquad H(\mathbf{x}) = \mathbf{x}^T\mathbf{x}, \text{sujeto a} \qquad R\mathbf{x} \leq \mathbf{r} \tag{2}$$

donde $R_{m\times n}$ es una matriz de rango completo, $\mathbf{x} \in \mathbf{R}^n$ y $\mathbf{r} \in \mathbf{R}^m$, $(n \geq m)$. La solución analítica de este problema se encuentra en el siguiente teorema que se enuncia sin demostración.

Teorema 2.1. *La solución del problema de programación cuadrática en su forma canónica está dada por* $\mathbf{x}_{op} = R^T\boldsymbol{\omega}$ *y* $H(x_{op}) = \boldsymbol{\omega}^T RR^T\boldsymbol{\omega}$*, donde los vectores* $\boldsymbol{\omega}$ *y* $\boldsymbol{\nu}$ *satisfacen las relaciones de complementaridad lineal* $RR^T\boldsymbol{\omega} = \mathbf{r} + \boldsymbol{\nu}$*,* $\boldsymbol{\omega}$*,* $\boldsymbol{\nu} \leq 0$ *y* $\boldsymbol{\omega}^T\boldsymbol{\nu} = 0$*.*

Las condiciones $\boldsymbol{\omega}, \boldsymbol{\nu} \leq 0$ y $\boldsymbol{\omega}^T\boldsymbol{\nu} = 0$ implican que cuando una coordenada de $\boldsymbol{\omega}$ es diferente de cero, la correspondiente coordenada de $\boldsymbol{\nu}$ necesariamente es cero, y viceversa. De esta manera, si se conociera qué coordenadas de $\boldsymbol{\omega}$ son diferentes de cero, se puede encontrar la solución para $\boldsymbol{\omega}$ y $\boldsymbol{\nu}$ en la ecuación $RR^T\boldsymbol{\omega} = \mathbf{r} + \boldsymbol{\nu}$. Para ejemplificar como se encontraría la solución, se supone que las primeras k coordenadas de $\boldsymbol{\omega}$ son diferentes de cero, $(k < n)$ y el resto son iguales a cero, $(\boldsymbol{\omega}^T = (\boldsymbol{\omega}_1^T, 0))$ y que RR^T se puede escribir en forma particionada como:

$$RR^T = \left(\begin{array}{c|c} A_{11} & A_{12} \\ \hline A_{12}^T & A_{22} \end{array}\right) \begin{array}{l} k \\ m-k \end{array}$$

entonces la ecuación $RR^T\boldsymbol{\omega} = \mathbf{r} + \boldsymbol{\nu}$ es igual a

$$\left(\begin{array}{c|c} A_{11} & A_{12} \\ \hline A_{12}^T & A_{22} \end{array}\right)\left(\frac{\boldsymbol{\omega}_1}{\mathbf{0}}\right) = \left(\frac{\mathbf{r}_1}{\mathbf{r}_2}\right) + \left(\begin{array}{c|c} I & \mathbf{0} \\ \hline \mathbf{0} & I \end{array}\right)\left(\frac{\mathbf{0}}{\boldsymbol{\nu}_2}\right)$$

Esta ecuación es equivalente a

$$\left(\begin{array}{c|c} A_{11} & \mathbf{0} \\ \hline A_{12}^T & -I \end{array}\right)\left(\frac{\boldsymbol{\omega}_1}{\boldsymbol{\nu}_2}\right) = \left(\frac{\mathbf{r}_1}{\mathbf{r}_2}\right) \tag{3}$$

de donde se sigue que la solución para $\boldsymbol{\omega}_1$ y $\boldsymbol{\nu}_2$ es: $\boldsymbol{\omega}_1 = A_{11}^{-1}\mathbf{r}_1$, y $\boldsymbol{\nu}_2 = A_{12}^T\boldsymbol{\omega}_1 - \mathbf{r}_2$

Entonces, una forma de encontrar la solución del problema de programación cuadrática es probar con todas las posibles ecuaciones de la forma (3); estas ecuaciones se identifican con los elementos del conjunto $E = \{z \in \mathbf{R}^m \mid z_i = 0 \text{ ó } z_i = 1\}$, esto es, para $z \in E$ se define la matriz cuadrada $A_z = (a_1|a_2|\ldots|a_n)$ tal que a_i es igual a la i-esima columna de la matriz RR^T cuando $z_i = 1$ y $c_i = -e_i$ cuando $z_i = 0$ (e_i es el i-esimo vector de la base canónica de $\mathbf{R}^m$). La solución del problema de programación cuadrática corresponde al vector $z \in E$ tal que $A_z^{-1}r < 0$.

Entonces, la solución del problema de programación cuadrática se obtiene calculando el vector $A_z^{-1}r$ para todos los elementos $z \in E$ hasta encontrar el vector $z_0 \in E$ tal que $A_{z_0}^{-1}r < 0$. Y como la cardinalidad de E es 2^m se tiene un problema de optimización combinatoria.

3. Técnica de ramificación y acotamiento

La técnica de ramificación y acotamiento es un algoritmo para encontrar el mínimo de una función discreta $\mathcal{H}: E \to \mathbf{R}$, con E conjunto finito, cuyos pasos son

1. Se propone una cota superior M para la solución mínimo factible de $\mathcal{H}(E)$. Si no se tiene información que le ayude, se puede comenzar con $M = \infty$.

2. Se establece una partición de E, (la ramificación) E_1, E_2, E_3, ..., E_N, del conjunto de posibles soluciones.

3. Se encuentra para cada subconjunto de la partición el valor mínimo de la función objetivo aunque no corresponda a una solución factible: $min(E_i) = \text{mín}_{e \in E_i}\, \mathcal{H}(e)$.

4. Se Realiza un sondeo por todos los elementos de la partición y se elimina de futura consideraciones a aquellos elementos que estén dentro de alguno de los supuestos siguientes:

   a) Sí $\text{mín}(E_i) \geq M$.

   b) Sí en el subconjunto no hay soluciones factibles.

   c) Sí $\text{mín}(E_i)$ corresponde a una solución factible, en cuyo caso se actualiza la cota superior $M = \text{mín}(E_i)$.

5. Se Escoje de entre los elementos de la partición que no fueron eliminados, (los elementos restantes) al que presente el mín(E_i) de menor valor. El elemento así elegido, se particiona en subconjuntos más pequeños y de nuevo se realiza un sondeo con los elementos de la partición restantes y con los de la nueva partición, de acuerdo al criterio del punto anterior.

6. Se Detiene cuando ya no hay más subconjuntos restantes para ramificar. La solución óptima corresponde al último valor asignado a M.

4. Aplicación de la técnica de ramificación y acotamiento para resolver el problema de programación cuadrática

Para resolver el problema de programación cuadrática usando esta técnica se debe primero definir una función objetivo adecuada. Segundo, se debe calcular el valor M. Tercero, se debe establecer la forma de construir la partición en cada paso y por último, se debe formular una regla para establecer el orden.

La función objetivo es $\mathcal{H} : \ E \rightarrow \ \mathbf{R}$ donde $E = \{z \in \mathbf{R}^n \mid z_i = 0 \ \ ó \ \ z_i = 1\}$ $\mathcal{H}(z) = r_z^T C_z^{-1} r_z$, con C_z definida igual que antes y $r_{zi} = r_i \times z_i$, sujeto a $C_z r \leq 0$.

Para calcular el valor inicial de M se observa que $x_u = R^T(RR^T)^{-1}u$ es una solución factible siempre que $u \leq r$ ya que $Rx_u \leq r$; entonces en particular $x_r = R^T(RR^T)^{-1}r$ es una solución factible cuyo valor de la función objetivo es $x_r^T x_r = r^T(RR^T)^{-1}r$. Ahora, se busca otras soluciones factibles con el valor de la función objetivo menor. Si el vector a es mayor que 0, entonces $r - a \leq r$ y por lo tanto $R^T(RR^T)(r-a)$ es otra solución factible y ahora se busca un vector a tal que $(r-a)^T(RR^T)^{-1}(r-a) \leq r^T RR^T)r$ este vector debe satisfacer la desigualdad $-2a^T(RR^T)^{-1}r + a^T(RR^T)^{-1}a < 0$. Es fácil determinar soluciones factibles con menor valor en la función objetivo si el vector a tiene una única coordenada diferente de cero, $a^T = (0, 0, \ldots, 0, a_i, 0, \ldots, 0)$ en este caso se tiene que

$$-2a^T(RR^T)^{-1}r + a^T(RR^T)^{-1}a = -2a_i b_i + a_i^2 c_{ii} = c_{ii}(a_i - b_i/c_{ii})^2 - b_i^2/c_{ii} < 0$$

donde b_i es la i-sima coordenada del vector $(RR^T)^{-1}r$ y c_{ii} es el i-esimo valor en la diagonal de $(RR^T)^{-1}$. El mínimo valor de $c_{ii}(a_i - b_i/c_{ii})^2 - b_i^2/c_{ii}$ es cuando $a_i = b_i/c_{ii}$ y para tener

una solución factible se requiere que $b_i > 0$.

Entonces, se escoge como M el valor $M = min_{b_i>0}\{r^T(RR^T)^{-1}r - b_i/c_{ii}\}$.

Finalmente, si A_1 y A_2 son dos matrices positivas definidas tales que $A_2 = \begin{pmatrix} A_1 & a \\ a^T & a_2 \end{pmatrix}$ y x_1 y x_2 son tales que $x_2^T = (x_1^T, x)$ entonces $x_1 A_1^{-1} x_1 \leq x_2^T A_2^{-1} x_2$; esto se sigue de que

$$\begin{pmatrix} A_1 & a \\ a^T & a_2 \end{pmatrix}^{-1} = \begin{pmatrix} A_1^{-1} + 1/\alpha A_1^{-1} a a^T A_1^{-1} & -1/\alpha A_1^{-1} a \\ -1/\alpha a^T A_1^{-1} & 1/\alpha \end{pmatrix} \quad \text{con} \quad \alpha = a_2 - a^T A_1^{-1} a.$$

De este resultado se sigue que si z_1 y $z_2 \in E$ son tales que $z_1 \leq z_2$ entonces $\mathcal{H}(z_1) \leq \mathcal{H}(z_2)$, y por lo tanto, si $z_0 \in E$ se tiene que $\mathcal{H}(z_0) > M$. Entonces, se deben eliminar de futuras consideraciones todos los elementos del subconjunto $\{z \in E \mid z > z_0\}$.

La partición de E en la etapa i esta dada por los elementos de $A_i = \{z \in E \mid \sum_{j=1}^{m} z_j = i\}$; esto significa que las ecuaciones que se deben probar son las correspondientes a $z \in E$ tales que tienen una única coordenada igual a cero. Las ecuaciones que se deben probar en la segunda etapa son los correspondientes a $z \in E$ con únicamente dos coordenadas diferentes o cero, etc.

5. Conclusiones

La técnica de ramificación y acotamiento adaptada al problema de programación cuadrática se probó con 50 ejemplos en los que la matriz R de 10×8 y el vector r de dimensión 8 fueron generados al azar, en todos los casos se llegó a la solución en no más de 100 exploraciones. Se considera que con una programación eficiente del algoritmo, puede mejorar la eficiencia del método.

Referencias

Escobar, L. A. And Skarpness, B. 1984. "A closed form solution for the least squares regression problem with linear inequality constraints"; Communications in Statistics, Theory and Method, 13 (9), pp 1127-1134.

Quintana, J., OŔeilly, F. and Gómez, S. 1987. "Least Squares with Inequality Restrictins: A Symmetric Positive-Definite Linear Complementary Problem Algorithm"; J. Statistics Computation and Simulation; Vol. 28, pp 128/143.

Memoria FONE XXIII (2009), Sección II, 137–142

Pruebas de bondad de ajuste para la distribución normal asimétrica

Paulino Pérez Rodríguez[a], José A. Villaseñor Alva
Colegio de Postgraduados

1. Introducción

Las distribuciones normales asimétricas constituyen una familia de distribuciones de tres parámetros: localización, escala y forma, la cual contiene a la familia normal cuando el parámetro de forma es 0 y a la distribución media-normal cuando dicho parámetro tiende a infinito. Esta familia de distribuciones tiene algunas de las propiedades de la familia normal, lo que la hace atractiva desde el punto de vista de aplicaciones. Esta familia apareció de forma independiente varias veces en la literatura estadística [Roberts(1966), O'Hagan y Leonard (1976)]; sin embargo, Azzalini (1985) estudió sus principales propiedades, propuso algunas generalizaciones y le dio el nombre con el cual se le conoce actualmente. Una revisión completa sobre esta distribución se encuentra en Azzalini (2005).

En este trabajo se proponen dos pruebas de bondad de ajuste y se comparan en términos de su potencia con otras pruebas existentes en la literatura utilizando simulación.

2. La distribución normal asimétrica

Definición 2.1. *Una v.a. Z tiene distribución normal asimétrica con parámetro de forma γ si su función de densidad es:*

$$f_Z(z;\gamma) = 2\phi(z)\,\Phi(\gamma z) I_{(-\infty,\infty)}(z) \tag{1}$$

donde $\phi(\cdot)$ y $\Phi(\cdot)$ denotan la función de densidad y de distribución normal estándar, $\gamma \in \mathbb{R}$.

[a]perpdgo@colpos.mx

Si Z tiene la función de densidad (1) entonces usualmente se escribe $Z \sim SN(\gamma)$. Si $Y = \xi + \omega Z$ con $\xi \in \mathbb{R}$ y $\omega \in \mathbb{R}^+$, entonces $Y \sim SN(\xi, \omega, \gamma)$ y su función de densidad es:

$$f_Y(y; \xi, \omega, \gamma) = 2\frac{1}{\omega}\phi\left(\frac{y-\xi}{\omega}\right)\Phi\left[\gamma\left(\frac{y-\xi}{\omega}\right)\right] I_{(-\infty,\infty)}(y).$$

Una propiedad importante de la distribución normal asimétrica es la siguiente: Si $Z \sim SN(\gamma)$ entonces $Z^2 \sim \chi^2_1$ para cualquier valor del parámetro γ (Azzalini, 2005).

3. Pruebas de bondad de ajuste

3.1. Planteamiento del problema

Sea $Y_1, ..., Y_n$ una muestra de una distribución F con densidad $f(y)$, con soporte en $\mathbb{R}$ y media finita. Supóngase que se desea probar el siguiente juego de hipótesis:

$$H_0 : f(y) = f_Y(y; \xi, \omega, \gamma), \text{ con } \xi \in \mathbb{R}, \omega \in \mathbb{R}^+, \gamma \in \mathbb{R} \quad \text{vs} \quad H_1 : f(y) \neq f_Y(y; \xi, \omega, \gamma) \quad (2)$$

Para probar (2) se proponen dos pruebas basadas en el coeficiente de correlación muestral.

3.2. Procedimiento de prueba 1

De las propiedades mencionadas de la distribución normal asimétrica, $X := (Y - \xi)^2 = \omega^2 Z^2 \sim \Gamma(1/2, 2\omega^2)$. Entonces para ξ fijo, digamos $\xi = \xi_0$, X tiene distribución con un parámetro de escala, i.e., $P(X \leq x) = F_1\left(\frac{x}{2\omega^2}\right)$, donde F_1 es la función de distribución de una variable aleatoria $\Gamma(1/2, 1)$. Entonces dada la muestra y ξ_0, se calculan las $x_i's$, con las que se estima F_1 utilizando la función de distribución empírica, es decir, $F_1\left(\frac{x}{2\omega^2}\right) \approx F_n(x)$ y por lo tanto $u := F_1^{-1}\left(F_n(x)\right) \approx \frac{x}{2\omega^2}$.

Entonces bajo la hipótesis nula se espera tener una relación lineal entre las $x_i's$ y las $u_i's$, y también se espera que esta relación se mantenga aún cuando se estime ξ por un estimador consistente $\hat{\xi}$. Para probar la existencia de esta relación lineal se plantea utilizar el estimador de momentos del coeficiente de correlación:

$$r_n = Corr(X, U) = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(U_i - \bar{U})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2 \sum_{i=1}^{n}(U_i - \bar{U})^2}} \quad (3)$$

La prueba r_n rechaza la hipótesis nula al nivel de significancia α si $r_n \leq C_n(\alpha)$, donde $C_n(\alpha)$ es tal que:

$$\alpha = \max_{\gamma} P(r_n \leq C_n(\alpha)|H_0).$$

Nótese que la distribución de r_n bajo H_0 no depende de ω ya que r_n es invariante bajo cambios de escala. La distribución de la estadística de prueba se puede obtener por simulación, usando el procedimiento siguiente:

1. Fijar $n, \xi = 0, \omega = 1$ y γ es un valor arbitrario.

2. Simular una muestra de tamaño n de $SN(\xi, \omega, \gamma)$

3. Obtener el estimador de máxima verosimilitud del parámetro ξ

4. Calcular x_i, $i = 1, ..., n$

5. Ordenar las $x_i's$ en forma ascendente

6. Calcular $u_i = F_1^{-1}\left(F_n(x_i)\right)$, $i = 1, n...$, donde F_1^{-1} es la función de cuantiles de la distribución $\Gamma(1/2, 1)$

7. Calcular r_n utilizando la ecuación (3), y los datos x_i, u_i generados en los pasos 5 y 6

8. Repetir los pasos 2 a 7 B veces

La distribución de la estadística de prueba depende fuertemente del valor desconocido del parámetro γ, como se observa en la Figura 1 (izquierda), lo cual sugiere que las constantes críticas deben ser tales que:

$$\alpha = \max_{\gamma} P(r_n \leq C_n(\alpha)|H_0) = \max_{\gamma \geq 0} P(r_n \leq C_n(\alpha)|H_0)$$

y se obtienen con los cuantiles 100α de la distribución empírica de r_n con $\gamma \to \infty$.

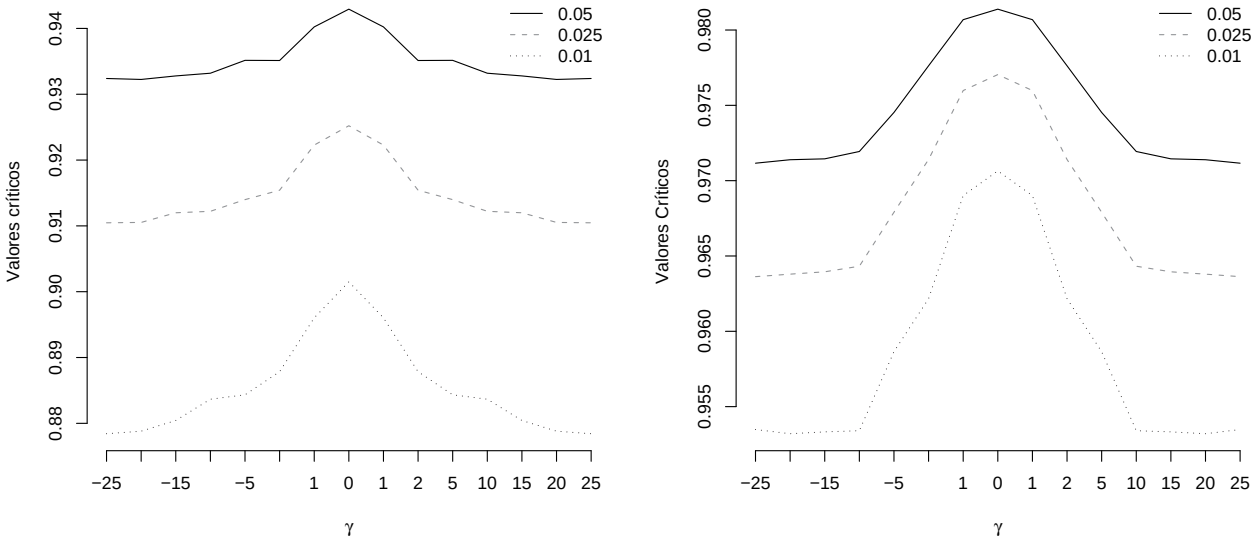


Figura 1: Valores críticos como función de γ para n=50, $B = 5,000$ para las estadísticas r_n (izquierda) y r_n^* (derecha).

3.3. Procedimiento de prueba 2

Para γ fijo, digamos $\gamma = \gamma_0$, Y tiene una distribución de localización y escala, es decir, $P(Y \leq y) = F_{\gamma_0}\left(\frac{y-\xi}{\omega}\right)$, donde F_{γ_0} es la función de distribución de $Z \sim SN(\gamma_0)$. Dadas las observaciones, la distribución empírica es un estimador consistente de $P(Y \leq y)$, entonces, $F_{\gamma_0}\left(\frac{y-\xi}{\omega}\right) \approx F_n(y)$, por lo tanto, $v := F_{\gamma_0}^{-1}\left(F_n(y)\right) \approx \frac{y-\xi}{\omega}$.

Entonces bajo la hipótesis nula se espera tener una relación lineal entre las $y_i's$ y las $v_i's$, y también se espera que esta relación se mantenga aún cuando se estime γ_0 por un estimador consistente $\hat{\gamma}$. Para probar la existencia de esta relación lineal se usa el coeficiente de correlación muestral r_n^*. Para obtener la estadística de prueba r_n^* se emplea el estimador de momentos de γ. El procedimiento para obtener las constantes críticas es similar al caso de r_n con $\gamma \to \infty$. La distribución de r_n^* no depende de ξ ni de ω ya que r_n^* es invariante bajo cambios de localización y escala.

4. Potencia de las pruebas

Las potencias de las pruebas propuestas se compararon con las estudiadas por Mateu *et al.* (2007) y $\tilde{T}_{n,a}$ de Meintanis (2007), la cual utiliza bootstrap paramétrico y está basada en

una estadística tipo Kolmogorov. Los resultados se muestran en la Tablas 1 y 2.

		Alternativa	A^2	W^2	U^2	D	V	D_n	r_n^*	r_n
Simétricas	Cola ligera	Logística estándar	0.1058	0.0849	0.1108	0.0672	0.1062	0.0284	0.1212	**0.1607**
		t(12)	0.0564	0.0464	0.0665	0.0419	0.0664	0.0256	0.0709	**0.1156**
	Cola pesada	t(4)	0.3293	0.2862	0.3228	0.2286	0.2933	0.0874	**0.3963**	0.3896
		Cauchy estándar	0.9946	0.9942	**0.9950**	0.9904	0.9936	0.9606	0.9942	0.9710
Asimétricas	Cola ligera	Exp. estándar	**0.7412**	0.7184	0.5231	0.6667	0.5156	0.7292	0.5148	0.3310
		Ji Cuadrada(4)	0.0999	0.1064	0.0828	0.1065	0.0744	0.1356	**0.1893**	0.1704
	Cola pesada	Weibull(0.75,1)	**0.9967**	0.9920	0.9666	0.9885	0.9656	0.9740	0.8856	0.5430
		Gumbel estándar	0.0565	0.0514	0.0661	0.0543	0.0637	0.0394	0.1397	**0.1600**
		Log-Normal(0,0.5)	0.1326	0.1387	0.1280	0.1335	0.1145	0.1480	**0.3064**	0.2672
		Sta(1.6,0.25,1,0;0)	0.5895	0.5435	0.5729	0.4859	0.5389	0.3342	**0.6613**	0.6606
		Sta(1.8,0.25,1,0;0)	0.2773	0.2398	0.2648	0.2147	0.2511	0.1296	0.3603	**0.4036**
		Sta(1.8,0.50,1,0;0)	0.2704	0.2344	0.2577	0.2092	0.2492	0.1340	0.3588	**0.4070**

Tabla 1: Comparación de potencias de las pruebas estudiadas por Mateu *et al.* y las propuestas para $n = 50$.

$\gamma = 0.25$		$\gamma = 0.50$		$\gamma = 0.75$		$\gamma = 1.0$		$\gamma = 2.0$	
6	4	6	3	5	3	4	5	4	5
$\nu = 2$		$\nu = 3$		$\nu = 5$		$\nu = 10$		$\nu = 15$	
48	**87**	33	**65**	11	**30**	7	**15**	6	**11**
$g = 0.25$		$g = 0.50$		$g = 1.0$		$g = 1.5$		$g = 2.0$	
3	**11**	25	**31**	85	**91**	98	**99**	**100**	**100**
$\phi = 1.0$		$\phi = 1.25$		$\phi = 1.55$		$\phi = 1.75$		$\phi = 2.0$	
9	**56**	20	**45**	27	**46**	34	**43**	**42**	**42**

Tabla 2: Comparación de potencias de las estadísticas $\tilde{T}_{n,a}$ (izquierda), r_n^* (derecha) para las distribuciones normal asimétrica $SN(\gamma)$, t asimétrica $ST(\gamma, \nu)$, g de Tukey $TU(g)$ y Laplace asimétrica $AL(\phi)$ con $n = 50$, $\alpha = 0.05$, $B = 100$ muestras bootstrap.

Mateu *et al.* (2007) mencionan que si $n <= 50$ sus pruebas resultan conservadoras. Por otro lado, para las pruebas estudiadas por Mateu cuando $\gamma < 10$ y $n < 50$, $\hat{\alpha}$ es prácticamente igual a la mitad del α nominal. Esto sugiere que si las pruebas de Mateu son calibradas para sostener el tamaño especificado entonces sus potencias pueden ser mayores que las reportadas.

5. Conclusiones

- Las pruebas de bondad de ajuste propuestas en general tienen buena potencia y respetan los tamaños especificados.
- El procedimiento de prueba 2 resultó ser más potente que el procedimiento propuesto por Meintanis(2007) y requiere mucho menos tiempo de cómputo para su cálculo.

Referencias

Azzalini, A. 1985. A class of distributions which includes the normal ones. *Scandinavian Journal of Statistics*, **12**, 171-178.

Azzalini, A. 2005. The skew normal distribution and related multivariate families. *Scandinavian Journal of Statistics*, **32**, 159-188.

O'Hagan, A. y Leonard, T. 1976. Bayes Estimation Subject to Uncertainty About Parameters Constraints. *Biometrika*, **63**, 201-203.

Mateu, G., P. Puig y Pewsey, A. 2007. Goodness-of-fit tests for the skew-normal distribution when the parameters are estimated from data. *Communications in Statistics-Theory and Methods*, **36**, 1735-1755.

Meintanis, S. G. 2007. A Kolmogorov-Smirnov Type Test for Skew Normal Distributions Based on the Empirical Moment Generating Function. *Journal of Statistical Planning and Inference*, **137**, 2681–2688.

Roberts, C. 1966. A correlation model useful in the study of twins. *Journal of the American Statistical Association*, **61**, 1184-1190.

Memoria FONE XXIII (2009), Sección II, 143–148

Pruebas exactas de no-inferioridad para probabilidades binomiales

Cecilia Ramírez Figueroa[a], David Sotres Ramos
Colegio de Postgraduados

Félix Almendra Arao
UPIITA – Instituto Politécnico Nacional

1. Introducción

La comparación de los efectos de dos tratamientos es un tópico que a menudo se utiliza en muchos campos de aplicación de la estadística. Es muy común comparar la superioridad de eficacia del tratamiento nuevo con respecto a un control; pero si los dos tratamientos tienen perfiles idénticos de eficacia, no es posible demostrar equivalencia exacta, por lo tanto se diseña un ensayo para mostrar que el tratamiento en prueba (con menores efectos secundarios o menor costo) es "no-inferior" al tratamiento control por una cantidad aceptable.

La metodología estadística para este problema es probar las hipótesis que estén asociadas con lo que se quiere demostrar, que en este caso es una diferencia de proporciones, $H_0 : p_1 - p_2 \geq d_0$; donde d_0 es una cantidad positiva previamente especificada. El objetivo de este trabajo es comparar los niveles de significancia de las pruebas exactas que se usan comúnmente para no-inferioridad: Blackwelder (1982); Farrington-Manning (1990) y de Razón de Verosimilitudes. Para realizar los cálculos se verificó la condición de convexidad de Barnard y de simetría en la misma cola.

2. Marco teórico

Sean X_1 y X_2 dos variables aleatorias independientes con distribución binomial con parámetros (n_1,p_1) para la primera y (n_2, p_2) para la segunda. Con p_1 se representa la probabilidad

[a]ceciliarf@colpos.mx

de éxito del tratamiento estándar (control) y con p_2 la del tratamiento nuevo. Se considera el problema de probar la hipótesis nula:

$$H_0 : p_1 - p_2 \geq d_0 \quad \text{vs} \quad H_a : p_1 - p_2 < d_0 \tag{1}$$

donde d_0 es el margen de no-inferioridad; el cual es una constante positiva y conocida. En el contexto de ensayos clínicos los valores usuales para d_0 son 0.10, 0.15 y 0.20, FDA (1992).

2.1. Prueba de Blackwelder (1982)

La estadística de prueba se define como:

$$T_B(x_1, x_2) = \frac{\widehat{p}_1 - \widehat{p}_2 - d_0}{\widehat{\sigma}_B} \tag{2}$$

donde $\hat{p}_i = \frac{X_i}{n_i}$ es el estimador de máxima verosimilitud de p_i para $i = 1, 2$, y $\hat{\sigma}$ es el estimador de la desviación estándar de $\hat{d} = \hat{p}_1 - \hat{p}_2$, dado por

$$\hat{\sigma}_B = \left(\frac{\hat{p}_1(1 - \hat{p}_1)}{n_1} + \frac{\hat{p}_2(1 - \hat{p}_2)}{n_2} \right)^{1/2}$$

2.2. Prueba de Farrington y Manning (1990)

La estadística de prueba tiene la misma forma, y varía en el estimador de la desviación estándar de $\hat{d} = \hat{p}_1 - \hat{p}_2$ y en este caso tal estimador está dado por

$$\hat{\sigma}_{FM} = \left(\frac{\check{p}_1(1 - \check{p}_1)}{n_1} + \frac{\check{p}_2(1 - \check{p}_2)}{n_2} \right)^{1/2}$$

donde $\check{p}_i$ es el estimador de máxima verosimilitud restringida bajo la hipótesis nula.

2.3. Prueba de razón de verosimilitudes

La estadística está dada por

$$T_{LR}(X_1, X_2) = \frac{\sup_{H_0} L(p_1, p_2;\ X_1, X_2)}{\sup_{H_0 \cup H_a} L(p_1, p_2;\ X_1, X_2)} \tag{3}$$

2.4. Nivel de significancia real

El espacio muestral está compuesto por $\chi = \{(x_1, x_2) \in \{0, ..., n_1\} \times \{0, ..., n_2\}\}$ y el espacio de parámetros es $\Theta = \left\{(p_1, p_2) \in [0, 1]^2\right\}$.

La región de rechazo (R_T) para un nivel nominal α para la prueba exacta determinada por la estadística de prueba T está dada por

$$R_T(\alpha) = \{(x_1, x_2) \in \{0, ..., n_1\} \times \{0, ..., n_2\} : T(x_1, x_2) \leq k\} \tag{4}$$

La distribución de las v.a.'s independientes es binomial, por lo tanto su distribución conjunta es:

$$L\left(p_1, p_2; x_1, x_2\right) = \binom{n_1}{x_1} p_1^{x_1} \left(1 - p_1\right)^{n_1 - x_1} \binom{n_2}{x_2} p_2^{x_2} \left(1 - p_2\right)^{n_2 - x_2}$$

y la función de potencia es $\beta_T(p_1, p_2) = \sum_{(x_1, x_2) \in R_T(\alpha)} L(p_1, p_2; x_1, x_2)$. En consecuencia el nivel de significancia es $\sup_{(p_1, p_2) \in \Theta_0} \beta_T(p_1, p_2)$ donde $\Theta_0 = \{(p_1, p_2) \in \Theta : p_1 - p_2 \geq d_0\}$ es el espacio nulo. $k = T(x_1^*, x_2^*)$ es la constante que define la región crítica y se obtiene como

$$T(x_1^*, x_2^*) = \text{máx} \left\{ T(a, b) : \sup_{(p_1, p_2) \in \Theta_0} \left(\sum_{T(x_1, x_2) \leq T(a, b)} L\left(p_1, p_2; x_1, x_2\right) \right) \leq \alpha \right\} \tag{5}$$

Chan (1998) investigó el comportamiento del nivel de significancia para la prueba asintótica de Farrington-Manning calculando el máximo únicamente en $\Theta_0^* = \{(p_1, p_2) \in \Theta : p_1 - p_2 = d_0\}$, el cual es solamente una parte de la frontera del espacio nulo. Röhmel (2005) presenta una prueba formal del procedimiento utilizado por Chan (1998) de lo que hay que resaltar que si la región de rechazo cumple con la Condición de Convexidad de Barnard la forma de cálculo está justificada.

Definición 2.1. *Se dice que una prueba estadística, para el problema en (1), con región de rechazo R_T cumple la* ***condición de convexidad de Barnard*** *(C) si satisface las dos propiedades siguientes:*

1. $(x, y) \in R_T \Longrightarrow (x - 1, y) \in R_T \quad \forall \quad 1 \leq x \leq n_1,\ 0 \leq y \leq n_2$

2. $(x, y) \in R_T \Longrightarrow (x, y + 1) \in R_T \quad \forall \quad 0 \leq x \leq n_1,\ 0 \leq y \leq n_2 - 1$

Definición 2.2. *Si $n_1 = n_2 = n$, se dice que una región de rechazo R cumple* ***la condición de simetría en la misma cola*** *si $(x, y) \in R \Longrightarrow (n - y,\, n - x) \in R$.*

Almendra-Arao (2008) probó el siguiente resultado.

Proposición 2.1. *Sean $n_1 = n_2 = n$ y $R(\alpha)$ una región crítica para el problema de prueba de hipótesis en (1), si $R(\alpha)$ cumple la condición de convexidad de Barnard y la condición de simetría en la misma cola, entonces el nivel de significancia exacto de la prueba $R(\alpha)$ está dado por*

$$\alpha^* = \max_{\substack{p_2=p_1-d_0 \\ p_1\in\left[d_0,\frac{1+d_0}{2}\right]}} \sum_{x_1=0}^{n_1}\sum_{x_2=0}^{n_2} \binom{n_1}{x_1} p_1^{x_1}\left(1-p_1\right)^{n_1-x_1} \binom{n_2}{x_2} p_2^{x_2}\left(1-p_2\right)^{n_2-x_2} I_{[(x_1,x_2)\in R(\alpha)]} \qquad (6)$$

En la expresión en (6) el intervalo de $p_1 \in [d_0, (1 + d_0)/2]$ se aproxima mediante el conjunto de puntos $\{d_0 + .001k : k = 0, 1, 2, \ldots, 500(1 - d_0)\}$. Cabe señalar que ésta es la única aproximación usada en la investigación.

La prueba de Farrington y Manning (1990) satisface las condiciones de la proposición 2.1 debido a que Röhmel (2005) lo probó analíticamente para cualesquiera n_1 y n_2 la condición (C). Para las pruebas de Blackwelder (1982) y de razón de verosimilitudes se probo numéricamente la condición (C) para n_1 y n_2 iguales. En cuanto a la condición de simetría en la misma cola de la proposición 2.1 se verificó numéricamente para las tres pruebas exactas de no inferioridad . La verificación numérica de las condiciones de la proposición 2.1 se realizó usando programas de cómputo escritos en S-PLUS®.

Para realizar los cálculos del nivel de significancia de las pruebas estudiadas se comienza por calcular los valores de la estadística de prueba correspondiente, se ordenan en forma ascendente los valores de la estadística de prueba calculada para todos los puntos (x_1,x_2) del espacio muestral que cumplan con ($\hat{p}_1$-$\hat{p}_2$)¡d_0. Se continua con formar la región critica con el par (x_1,x_2) que tenga el valor más pequeño de la estadística de prueba seleccionada y que cumpla la condiciones de la proposición 2.1. Se calcula (6) con este par (x_1,x_2) y sucesivamente se añadirán pares en el orden definido por la estadística de prueba seleccionada hasta que se obtenga una región crítica R_T cuyo tamaño α^*, se acerque lo más posible al nivel de significancia nominal sin excederlo α, es decir $\alpha^* \leq \alpha$. Este proceso fue implementado en S-PLUS®, los programas están disponibles a petición.

2.5. Resultados

Para evaluar el comportamiento de las pruebas en la Tabla 1 se presentan los porcentajes de niveles de significancia que se encuentran en el intervalo [0.8α,α].

alfa(α)	Intervalo	d_0	Porcentaje de N.S. que pertenecen al Intervalo		
	$[0.8\alpha, \alpha]$		Blackwelder	Farrington-Manning	Razón de Vero.
0.01	[0.008,0.01]	0.10	15.63	96.88	76.84
		0.15	25.00	97.92	83.16
		0.20	29.17	95.83	86.32
0.05	[0.04,0.05]	0.10	28.13	98.96	81.05
		0.15	39.58	95.83	88.42
		0.20	46.88	93.75	92.63
0.10	[0.08,0.10]	0.10	41.67	90.63	86.32
		0.15	48.96	96.88	90.53
		0.20	61.46	94.79	92.63

Tabla 1: Porcentaje de tamaños de muestra entre 5 y 100 donde los niveles de significancia pertenecen al intervalo [0.80α, α]

En la Figura 1 se muestra el nivel de significancia de las pruebas para el nivel nominal $\alpha = 0.05$ según el tamaño de muestra $5 \leq n \leq 100$.

3. Conclusiones

No obstante que la prueba Blackwelder es más fácil de aplicar que la prueba de Farrington y Manning, motivo por el cuál es muy utilizada, sus niveles de significancia están muy alejados del nivel nominal para un gran porcentaje de los tamaños de muestra y mientras menor sea el nivel de significancia, más difícil será rechazar la hipótesis nula y declarar al tratamiento en prueba como no inferior.

Todas las pruebas investigadas mantienen el nivel nominal. Pero los niveles de la prueba de Farrington y Manning y de la de razón de verosimilitudes siempre están por encima de la prueba de Blackwelder.

Se recomienda la prueba de Farrington y Manning debido a que los valores del nivel de

significancia están muy cercanos al nivel nominal lo que indica que no es una prueba muy conservadora.

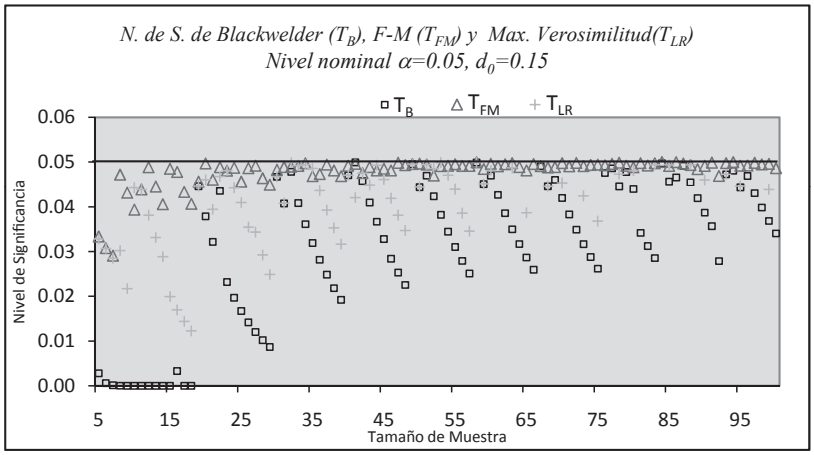


Figura 1: Niveles de Significancia de las pruebas Blackwelder (T_B), Farrington-Manning (T_{FM}) y Máxima Verosimilitud (T_{LR}) para el nivel nominal $\alpha = 0.05$ y $d_0 = 0.15$

Referencias

Almendra-Arao, F. 2008. *A Study on the Classical Asymptotic Non-inferiority Test for Two Binomial Proportions.* Drug Information Journal. En prensa.

Blackwelder, W. C. 1982. *Proving the null hypothesis in clinical trials.* Controlled Clinical Trials. 3, 345-353.

Chan, I. 1998. *Exact tests of equivalence and efficacy with a non zero lower bound for comparative studies.* Statistics in Medicine. **17**, 1403-1413.

Farrington, C. and G. Manning. 1990. *Tests statistics and sample size formulae for comparative binomial trials with null hypothesis of non-zero risk difference or non-unity relative risk.* Statistics in Medicine, **9**, 1447-1454.

FDA. 1992. *Points to consider. Clinical development and labeling of anti-infective drug products.* U.S. Dept. of Health and Human Services. FDA, CDER.

Röhmel, J. 2005. *Problems with existing procedures to calculate exact unconditional p-values for noninferiority and confidence intervals for two binomials and how to resolve them.* Biometrical Journal, **47**, 37-47.

Memoria FONE XXIII (2009), Sección II, 149–154

Modelos lineales generalizados con restricciones lineales de desigualdad en los parámetros

Silvia Ruiz Velasco Acosta[a], Federico O'Reilly Togno
IIMAS – Universidad Nacional Autónoma de México

1. Introducción

1.1. Modelos lineales generalizados

Los modelos lineales generalizados fueron propuestos por Nelder y Weederburn (1970), estos modelos están especificados por tres componentes:

- El *componente aleatorio* que consta de observaciones independientes de $\mathbf{Y} = (Y_1, \ldots, Y_n)^t$ donde Y_i es una variable aleatoria con distribución perteneciente a la familia exponencial que incluye un parámetro de escala ϕ y se expresa de la siguiente manera:

$$f(x_i, \theta_i, \phi) = \exp\Big[\frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y_i, \phi)\Big]. \tag{1}$$

- El *componente sistemático* $\eta(\cdot)$ al que se le conoce como predictor lineal, es decir, que

$$\eta(\cdot) = \mathbf{X}\beta \tag{2}$$

- La *función liga*, $g(\cdot)$ una función monótona y diferenciable, que describe la relación entre la media de la i-ésima observación y su predictor lineal

$$\eta_i = g(\mu_i) = \sum_{j=1}^{p} \beta_j x_{ij} \quad i = 1, \ldots, n \tag{3}$$

[a] silvia@sigma.iimas.unam.mx

A $g(\cdot)$ se le conoce con el nombre de liga canónica cuando se cumple que $\theta = \eta$, siendo θ el parámetro canónico. En este caso, los parámetros desconocidos de la estructura lineal, el vector β, tienen como estadísticas suficientes $X^T Y$.

Algunos ejemplos en la literatura de situaciones en las que es deseable imponer restricciones de desiguldad en los parámetros son:

- Un modelo logístico en donde la variable respuesta es "creditabilidad", que toma el valor de cero cuando un individuo es considera para otorgarle un crédito y de uno cuando se considerada que no debe entregársele un crédito. La variable explicativa que denota la cantidad de crédito a otorgarse se categorizó y lo que se quieres es imponer una restriccion de orden en los parámetros correspondientes a las diferentes categorías. (Fahrmeir y Klinger, (1994))

- Un modelo Poisson para explicar el número de incidentes en un barco, entre las variables explicativas se encuentra el año de construcción categorizado en períodos de cinco años y la restricción que se impone es que el año de construcción influye en la seguridad del barco, en particular que barcos más viejos son más seguros. (Kredler, (1993))

- Muertes por cáncer en trabajadores expuestos a arsénico. En este caso el modelo propuesto utiliza una reparametrización de los niveles y el tiempo de exposición, de tal forma que cada nivel se compara con el previo. La restricción es que todos los parámetros deben ser mayores ó iguales a cero, es decir a mayor exposición tanto en tiempo como cantidad, hay mayor riesgo. (McDonald y Diamond, (1990))

1.2. Estimación

La estimación usual en el caso de Modelos Lineales Generalizados es por máxima verosimilitud. La función log-verosimilitud para la familia exponencial está dada por

$$\ell = \sum_{i=1}^{n} \frac{y_i \theta_i - b(\theta_i)}{a(\phi)} - c(y_i, \phi). \tag{4}$$

La primera y segunda derivada de ℓ con respecto a β, estan dados por:

$$\frac{\partial \ell}{\partial \beta_j} = \frac{1}{a(\phi)} \sum_{i=1}^{n} \left[y_i \frac{\partial \theta_i}{\partial \beta_j} - \frac{\partial b(\theta_i)}{\partial \theta_i} \frac{\partial \theta_i}{\partial \beta_j} \right] \sum_{i=1}^{n} \frac{(y_i - \mu_i)}{a(\phi) v(\mu_i)} \frac{\partial \mu_i}{\partial \eta_i} x_{ij} \tag{5}$$

y

$$\frac{\partial^2 l}{\partial\beta_j\partial\beta_k} = -\sum_{i=1}^{n}\frac{1}{a(\phi)}\Big[\frac{1}{v(\mu_i)}(\frac{\partial\mu_i}{\partial\eta_i})^2\Big]$$

$$-(\mu_i - y_i)\Big[\frac{1}{v(\mu_i)^2}(\frac{\partial\mu_i}{\partial\eta_i})^2\frac{\partial v(\mu_i)}{\partial\mu_i} - \frac{1}{v(\mu_i)}\frac{\partial^2\mu_i}{\partial\eta_i^2}\Big]x_{ij}x_{ik}. \tag{6}$$

Las ecuaciones máximo verosímiles para β generalmente se resuelven por un método iterativo de mínimos cuadrados ponderado.

$$\boldsymbol{\beta}^{(r)} = (\boldsymbol{X}^T\boldsymbol{W}^{(r-1)}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{W}^{(r-1)}\boldsymbol{z}^{(r-1)} \tag{7}$$

con

$$\boldsymbol{W} = diag(\frac{1}{v(\mu)a(\phi)}(\frac{\partial\mu}{\partial\eta})^2) \tag{8}$$

y

$$\boldsymbol{z} = (y-\mu)(\frac{\partial\eta}{\partial\mu} + \eta), \tag{9}$$

con $\boldsymbol{W}$ y $\boldsymbol{z}$ son evaluadas en $\boldsymbol{\beta}^{(r)}$. Si la liga es canónica, esto corresponde a la solución utilizando Newton-Raphson y si no lo es, corresponde a remplazar la matriz de segundas derivadas por su valor esperado.

1.3. Algoritmo del cono

El algoritmo del cono propuesto por Quintana et. al.(1987) y revisado en O'Reilly et.al. (2008), tiene aplicación en modelos de regresión con restricciones en los parámetros. Es decir si consideramos el problema

$$min||Y - X\beta||^2 \tag{10}$$

sujeto a $A\beta > b$, con Y n-dimensional X de nxp y A de mxp. El problema no restringido de β es encontrar u y v de dimensión m tal que

$$Mu + q = v \tag{11}$$

con u y $v > 0$ y $u'v = 0$, $M = A(X'X)^{-1}A'$ y $q = A\beta_u - a$ donde β_u es el estimador no restringido.

2. Estimación con restricciones

Algunas de las sugerencias para resolver el problema planteado:

- McDonald y Diamond (1990) sugieren para el caso de la restricción de no negatividad $A\beta >= 0$ realizar una búsqueda sobre todas las posibles soluciones. Esto es, si la solución máximo verosímil no satisface las restricciones, al menos una de las β 's debe ser cero. La idea es ajustar sub modelos con menos variables y utilizar las condiciones de Kuhn Tucker para determinar si un submodelo que cumple las restricciones es el óptimo.

- Fahrmeir y Klinger (1994) proponen un algoritmo general aplicable a problemas de estimación de máxima verosimilitud con restricciones llamado SQP (sequence of quadratic problems), que incorporan la restricción al proceso iterativo de estimación, la estimación de modelos lineales generalizados con restricciones es un caso particular.

Desde el punto de vista Bayesiano, ver por ejemplo Gelfand et.al, (1992); Chen et.al (2000). Existen otros métodos que resuelven el caso particular de restricciones de igualdad, por ejemplo Nyquist, (1991).

Nuestra propuesta es obtener el estimador de mínimos cuadrados iterativos ponderados del modelo lineal generalizado sin restricciones y utilizar el algoritmo del cono a partir de las últimas actualizaciones de la matriz de pesos y de la variable de trabajo, es decir, resolvemos (7) y si la convergencia se da en la iteración s, obtenemos $W^{(s)}$ y $z^{(s)}$ dados por (8) y (9) y evaluadas en $\boldsymbol{\beta}^{(s)}$.

Aplicamos el algoritmo del cono al modelo de regresión ponderada de la última iteración. Para comprobar si el valor obtenido para β es óptimo podemos utilizar las condiciones de Kuhn Tucker, las cuales son condiciones suficientes para un máximo local si la verosimilitud es estrictamente cóncava. En este caso para toda j se cumple una de las dos siguientes condiciones $\beta_j^* > 0$ y $\frac{\partial \ell}{\partial \beta_j} = 0$ o $\beta_j^* = 0$ y $\frac{\partial \ell}{\partial \beta_j} <= 0$. Las parciales con respecto a β en este caso están dadas por el producto ponderado de las variables explicativas y los residuales.

Utilizamos nuestra propuesta con varios ejemplos en literatura y los reproduce los resultados publicados. Para probar como funciona en un modelo de mayor dimensión generamos 50 observaciones de un modelo Poisson con media $exp(1 - .47x_1 + .6x_2 + .7x_3 - .8x_4 - .4x_5 +$

$-.7x_6 + .3x_7 - .6x_8 - .9x_9 - x_{10} + 1.2x_{11} + 1.4x_{12} - 1.6x_{13} - 1.8x_{14})$, en donde las x's son variables generadas como normales.

Los estimadores del modelo sin restricciones son:

$$\begin{array}{lllll} \beta_0 = 0.967, & \beta_1 = -0.456, & \beta_2 = 0.572, & \beta_3 = 0.730, & \beta_4 = -0.786, \\ \beta_5 = 0.344, & \beta_6 = -0.689, & \beta_7 = 0.299, & \beta_8 = -0.574, & \beta_9 = -0.886, \\ \beta_{10} = -0.989, & \beta_{11} = 1.216, & \beta_{12} = 1.415, & \beta_{13} = -1.628, & \beta_{14} = -1.834. \end{array}$$

Utilizando nuestra propuesta obtenemos en el modelo con restricciones los siguientes estimdadores puntuales:

$$\begin{array}{lllll} \beta_0 = 7.376, & \beta_1 = 0, & \beta_2 = 0.7232, & \beta_3 = 0.4384, & \beta_4 = 0.389, \\ \beta_5 = 0, & \beta_6 = 1.452, & \beta_7 = 1.372, & \beta_8 = 0, & \beta_9 = 0, \\ \beta_{10} = 0, & \beta_{11} = 0, & \beta_{12} = 0, & \beta_{13} = 0, & \beta_{14} = 0. \end{array}$$

que claramente no corresponde al ajuste del modelo restringido quitando los parámetros menores a cero. En este caso para analizar todas los posibles submodelos (más de 16000) es necesario tener un algoritmo de búsqueda muy eficiente.

En el caso de que la liga es canónica tanto en nuestra propuesta como el método SQP los estimadores convergen. Aquí falta comparar la eficiencia computacional de los dos métodos. En el caso en que la liga no es canónica el SQP reporta "buenos resultados númericos", aunque la convergencia no esta probada. Nuestra propuesta puede ser modificada para incluir en la matriz de pesos el segundo término de la matriz de segundas derivadas de β.

Referencias

Chen, M.H., Shao, Q.M. and Ibrahim, J.G. 2000. Monte Carlo methods in bayesian computation. Springer-Verlag, New York, UAS. 191-208.

Fahrmeir, L. Klinger, J. 1994. Estimating and testing generalized linear models under inequality restriccions. *Statistical Papers* 35, 211-229

Gelfand A.E., Smith A.F.M., Lee, T.M. 1992. Bayesian Analysis of constrained parameter and truncated data problems using Gibbs sampling. *Journal of the American Statistical Association*, **87**, 523-532.

Kredler, C. 1993. The SQP-method for linearly constrained maximum likelihhod problems. *Technical Report Nr.IAMSI1994.5TUM, Technical University Munich*

McDonald, J.M. Diamond I. 1990. On the fiting of generalized linear models with non-negative parmeter constrains. *Biometrics*, **46**, 201-206.

Nelder, J.A. and Wedderburn, R.W.M. 1972. Generalized linear models. *J. R. Statist. Soc. A*, **135**, 370-384.

Nyquist, H. 1991. Restricted estimation of Generalized linear models. *Appl. Statist.*, **40**, 133-141.

O'Reilly F., Quintana J. 2008. Non-parametric regression the cone algorithm. In *Preimpreso IIMAS* **149**.

Quintana J., O'Reilly F., Gomez S. 1987. Least Squares with inequality restriccions a symmetric positive definite linear complementary algorithm . *J. Statist. Comput. Simul.*, **28**, 127-143

Memoria FONE XXIII (2009), Sección II, 155–163

Estudio de algunas distribuciones multivariadas mediante el apoyo de un programa de computo matemático

Ismael Sosa Galindo, Agustín Jaime García Banda
Facultad de Ciencias Administrativas y Sociales – UV

Luis Cruz-Kuri
Instituto de Ciencias Básicas – UV

1. Introducción

En los cursos básicos de estadística multivariada, se presentan una variedad de modelos probabilistas, tanto de tipo absolutamente continuo como de tipo puramente discreto. De los modelos absolutamente continuos, el más utilizado, corresponde a la familia de la distribución normal multivariada con vector de medias de dimensión p y matriz de varianzas y covarianzas de dimensiones $p \times p$. De manera natural, también se presentan otros modelos de tipo continuo, tales como el de la T-cuadrada de Hotelling. Para los modelos de tipo discreto aparecen las generalizaciones de los modelos univariados, tales como los de la familia binomial o hipergeométrica. Ya sean distribuciones teóricas o empíricas, es de importancia el poder estudiarlas en sus detalles probabilistas y estadísticos. Programas de cómputo estadístico, tales como *SPSS*, *STATISTICA* o *BMDP*, tienen poderosas funciones numéricas de procesamientos multivariados y de presentaciones gráficas en dos y tres dimensiones. Por otra parte, para hacer un estudio con mayor flexibilidad, hemos encontrado que un programa de computo matemático, tal como *Mathematica*, facilita notablemente los procesamientos tanto de tipo numérico como gráfico, y permite, dadas las fórmulas funcionales de los modelos multivariados bajo consideración, su visualización al menos para dos y tres dimensiones. Lo mismo se puede hacer para el procesamiento de datos empíricos. Así, por ejemplo, para la distribución normal de tres dimensiones, se presentan, entre otros objetos geométricos, la

colección de los elipsoides de concentración. De manera análoga, estudiamos algunas distribuciones discretas, tales como las familias de la distribución binomial y de la distribución hipergeométrica (con dos o más variables).

2. Utilización del programa *Mathematica*

La función de densidad definida en (1) corresponde a la Distribución Normal Bivariada con vector de medias (μ_1, μ_2) y matriz de variancias y covarianzas {{$\sigma_{1,1}$,$\sigma_{1,2}$},{$\sigma_{2,1}$,$\sigma_{2,2}$}}

$$f := (x, y) \to \frac{1}{2} \frac{e^{\left(-\frac{\frac{(x-\mu_1)^2}{\sigma_{1,1}} + \frac{(y-\mu_2)^2}{\sigma_{2,2}} - 2\frac{\rho(x-\mu_1)(y-\mu_2)}{\sqrt{\sigma_{1,1}\sigma_{2,2}}}}{2-2\rho^2} \right)}}{\pi\sqrt{\sigma_{1,1}\sigma_{2,2}(1-\rho^2)}} \tag{1}$$

Con las instrucciones en *Mathematica* que siguen se genera la superficie correspondiente a la gráfica de la función definida en (1) para valores selectos de los parámetros (Figura 1). Con las instrucciones que siguen se generan las curvas de nivel de dicha superficie (Figura 2).

Plot3D [Exp [$-x^2 - y^2$],{x, -2, 2},{y, -2, 2}, Background → RGBColor[0, .3, 0],
PlotPoints→ 40, Axes → Automatic, Boxed → False, LightSources→ {{{1, 2, 1.5},
RGBColor[1,0,0]}, {{-3, -2, 1}, RGBColor[0, 0, 1]}, {{20, 0, 9}, RGBColor[1, 0, 1]}}]
ContourPlot[Exp [$-x^2 - y^2$], {x, -2, 2}, {y, -2, 2}, Background →
RGBColor[1, 1, 1], PlotPoints→100]

De manera semejante, con las instrucciones *Mathematica* que siguen se generan la superficie y las curvas de nivel para una distribución normal bivariada con ρ=-0.8 (ver figuras 3 y 4).

Plot3D [Exp [$-0.1x^2 - y^2 - 0.5x * y$], {x,-5,5},{y,-3,3}, Background →
RGBColor[0,0.3,0], PlotPoints → 60, LightSources→ {{{1.3,0.4,2},
RGBColor[1,0,0]}, {{2.7,0,2}, RGBColor[-2.3,-1.4,2]}, {{0,0,1}}]

Nota. Se tomaron las desviaciones estándar σ_1 y σ_2 de tal forma que $\sigma_1\sigma_2$=40/9. Hay muchas opciones para σ_1 y σ_2; por ejemplo, si se escoge $\sigma_{2,2}$=100/72, se obtiene $\sigma_{1,1}$=1152/81. Con estos valores, la forma cuadrática (incluido el factor $0.5/(1-\rho^2)$) corresponde a $0.10x^2 + y^2 + 0.50xy$ y el coeficiente de correlación es $\rho = -0.8$.

Plot3D [Exp [$-0.1x^2 - y^2 - 0.5x * y$], {x,-5,5},{y,-3,3}, Background →
RGBColor[0,0.3,0], PlotPoints → 60, LightSources→ {{{1.3,0.4,2},

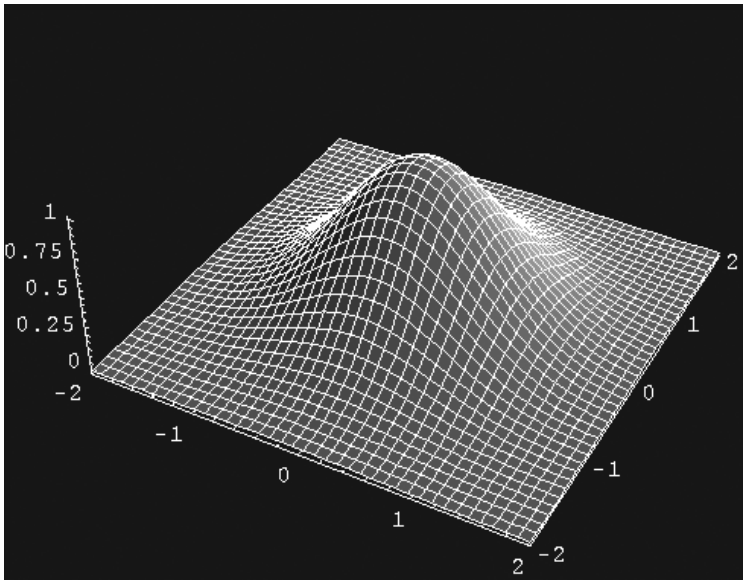


Figura 1: Superficie de la distribución normal bivariada con $\rho = 0$.

RGBColor[1,0,0]}, {{2.7,0,2}, RGBColor[-2.3,-1.4,2]}, {{0,0,1}}]

Aunque para el artículo presente las gráficas aparecen en tonos de gris, se dejan las instrucciones para controlar colores para los lectores interesados en usar esta opción.

ContourPlot[Exp $[-0.1x^2 - y^2 - 0.5x * y]$, {x,-5,5}, {y,-3,3},Background →
RGBColor[1,1,1], PlotPoints→100]

Instrucciones de *Mathematica* para generar distribución hipergeométrica multivariada con parámetros N, n, m_1, m_2, donde N=tamaño de la población, n=tamaño de la muestra, m_1=número de objetos en la categoría 1, m_2=número de objetos en la categoría 2.

combi[n_,r_]:=Factorial[n]/(Factorial[r]*Factorial[n-r]) (2)

hipergeo[x_,y_,npob_,m1_,m2_,n_]:=combi[m1,x]*combi[m2,y]
*combi[npob-m1-m2,n-x-y] /combi[npob,n] (3)

Instrucciones de *Mathematica* para generar distribución polinomial multivariada, también conocida como distribución multinomial, con parámetros n, p_1, p_2, donde n = número de pruebas de Bernoulli.

polin[x_,y_,n_,p1_,p2_]:=Factorial[n]*p1^x*p2^y*
(1-p1-p2)^(n-x-y)/(Factorial[x]* Factorial[y]* Factorial[n-x-y]) (4)

Ejemplo adicional. Dos catadores profesionales clasifican muestras de vino, dando cada

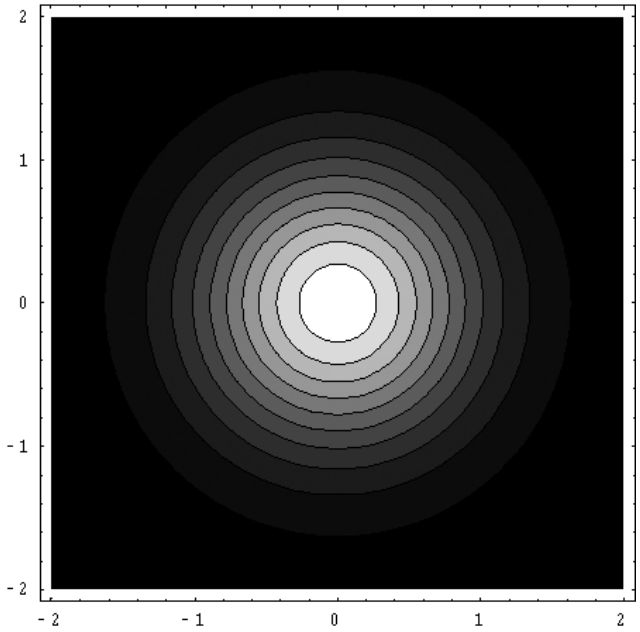


Figura 2: Curvas de nivel de la distribución normal bivariada con $\rho = 0$.

uno de ellos una calificación de 1 a 5. Sea X = valor asignado por el primer catador a la muestra de vino y sea Y = valor asignado por el segundo catador a la misma muestra de vino. Sigue la distribución de probabilidad conjunta de X & Y (Tabla 3).

BarChart3D[{{0.03,0.02,0.01,0.00,0.00}, {0.02,0.08,0.05,0.02,0.01},
{0.01,0.05,0.25,0.05,0.01}, {0.00,0.02,0.05,0.20,0.02}, {0.00,0.01,0.01,0.02,0.06} }]

3. Discusión

Aparte de la utilización de los programas de cómputo estadístico especializados, tales como *SPSS* o *STATISTICA*. ¿Qué más se puede hacer para estudiar algunos modelos probabilistas multivariados, al menos para dos y tres dimensiones? Hemos encontrado bastante útil un programa de cómputo matemático, tal como *Mathematica*, que permita procesamientos a niveles numérico, algebraico y gráfico. ¿Qué se puede hacer entonces en forma más específica? Nosotros hemos utilizado el paquete *Mathematica* para presentar, dada la función de densidad de probabilidad de un modelo multivariado absolutamente continuo, su gráfica correspondiente, al menos para el caso bivariado. Así mismo, en una misma gráfica, hemos

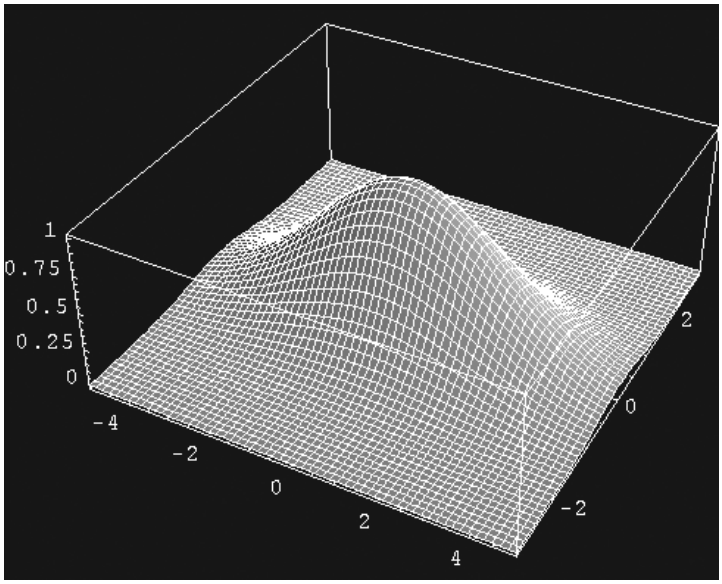


Figura 3: Superficie de la distribución normal bivariada con ρ=-0.8

presentado dos modelos bivariados para propósitos de comparación. Por otra parte, también para modelos discretos multivariados, hemos podido elaborar las gráficas correspondientes (tipo Histograma). Para la visualización de gráficas que requieren un espacio tridimensional, el programa *Mathematica* permite la construcción de curvas de nivel, por ejemplo.

Referencias

Blachman, N. 1992. *Mathematica: A Practical Approach.*. Prentice Hall, Inc.

Derman, C., Gleser, L. y Olkin I. 1973. *A Guide to Probability Theory and Application*. Holt, Rinehart and Winston, Inc.

Morrison, D.F. 1990. *Multivariate Statistical Methods*. Third Edition.McGraw-Hill.

Shawn, W. y Tigg, J. 1994. *Applied Mathematica - Getting Started, Getting It Done*. Addison Wesley. New York, USA.

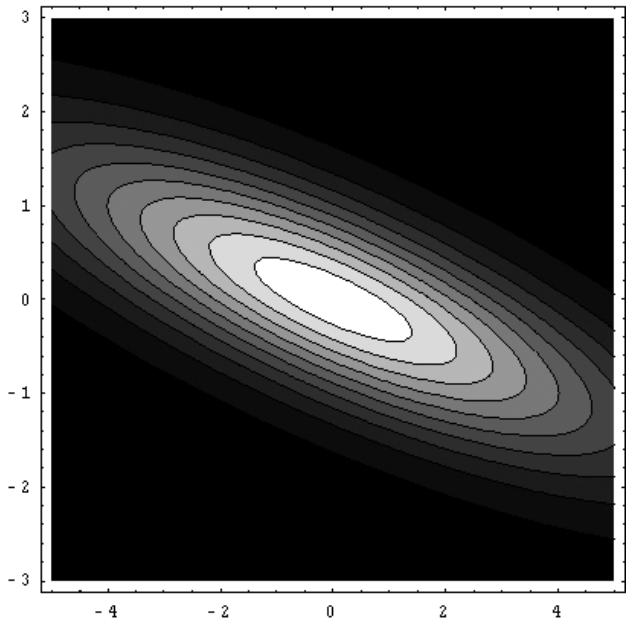


Figura 4: Curvas de nivel de la distribución normal bivariada con ρ=-0.8.

$$\begin{pmatrix} 0 & 0 & 2 & 2 \\ 0 & 8 & 24 & 8 \\ 4 & 36 & 36 & 4 \\ 8 & 24 & 8 & 0 \\ 2 & 2 & 0 & 0 \end{pmatrix}$$

Tabla 1: Matriz: Table [168*hipergeo [x,y,10,4,3,5],x,0,4,y,0,3]//MatrixForm

$$\begin{pmatrix} .00243 & .0162 & .0432 & .0576 & .0384 & .01024 \\ .01215 & .0648 & .1296 & .1152 & .0384 & 0 \\ .0243 & .0972 & .1296 & .0576 & 0 & 0 \\ .0243 & .0648 & .0432 & 0 & 0 & 0 \\ .01215 & .0162 & 0 & 0 & 0 & 0 \\ .00243 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

Tabla 2: Matriz generada: Table[polin[x,y,5,.3,.4], {x,0,5}, {y,0,5}] //MatrixForm. Ver (4).

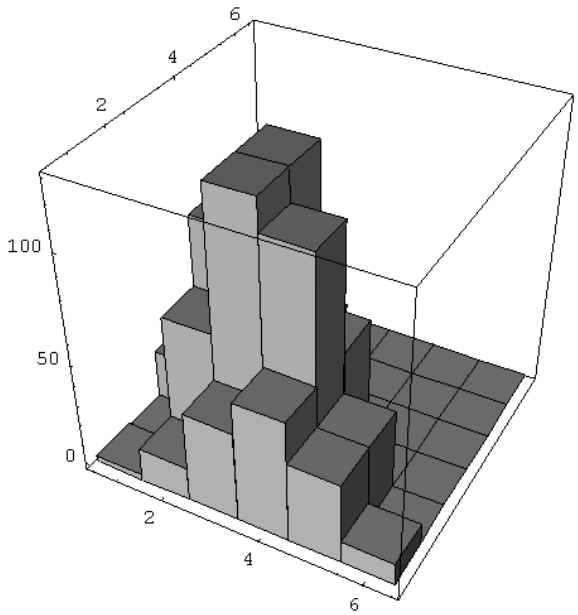


Figura 5: Distribución discreta hipergeométrica bivariada. Ver Tabla 1.

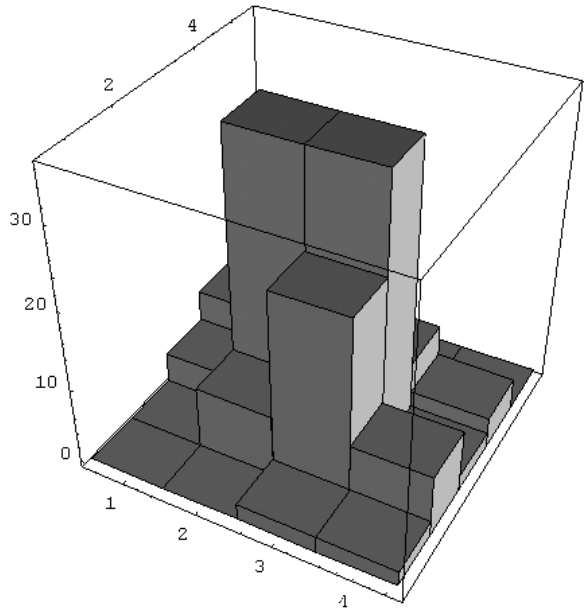


Figura 6: Distribución de probabilidades polinomiales bivariadas. Ver Tabla 2.

	1	2	3	4	5
1	0.03	0.02	0.01	0.00	0.00
2	0.02	0.08	0.05	0.02	0.01
3	0.01	0.05	0.25	0.05	0.01
4	0.00	0.02	0.05	0.20	0.02
5	0.00	0.01	0.01	0.02	0.06

Tabla 3: Distribución conjunta para calificaciones de dos catadores profesionales

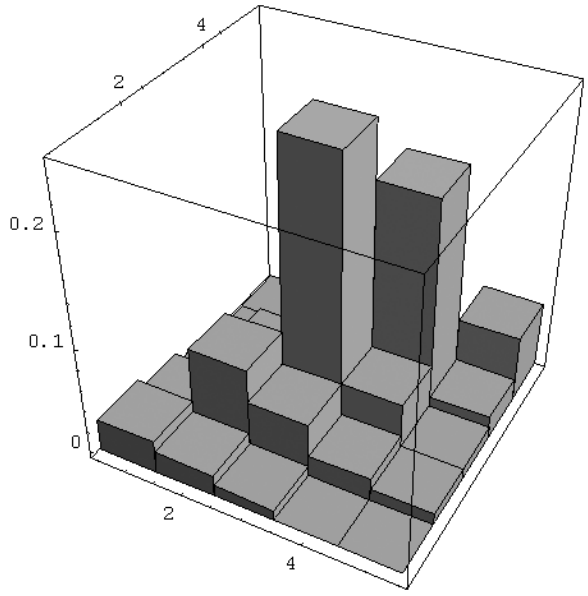


Figura 7: Probabilidades empíricas bivariadas, experimento de dos catadores de vino. Ver Tabla 3

Memoria FONE XXIII (2009), Sección II, 165–171

Tablas para la prueba exacta de no inferioridad de Farrington-Manning

David Sotres Ramos[a], Cecilia Ramírez Figueroa
Colegio de Postgraduados

Félix Almendra Arao
UPIITA – Instituto Politécnico Nacional

1. Introducción

Las pruebas estadísticas de no-inferioridad se utilizan muy frecuentemente en ensayos clínicos. Estas pruebas sirven para ver si existe evidencia muestral de que una terapia nueva (con mínimos efectos secundarios o bajo costo) no es sustancialmente inferior en eficacia a la terapia estándar, ver Chen et al. (2000). La prueba exacta de no inferioridad de Farrington-Manning (FM) para la comparación de dos proporciones independientes ha sido estudiada y recomendada por varios autores e.g. Farrington y Manning (1990), Martin y Herranz (2004).

Sin embargo, la prueba de FM es numéricamente difícil de aplicar debido a que la constante crítica de la prueba es complicada de calcular. El principal objetivo de este trabajo es elaborar tablas para la constante crítica y el tamaño de la prueba exacta de FM. Estas tablas simplifican considerablemente el uso de esta prueba. Además por cálculo directo se ha verificado que el nivel de significancia de esta prueba está muy cercano al nivel nominal de la prueba, para los niveles nominales $\alpha = 0.01$ y $\alpha = 0.05$.

2. Marco teórico

Sean X_1 y X_2 dos variables aleatorias independientes con distribución binomial y con parámetros (n_1,p_1) y (n_2,p_2) respectivamente, donde p_1 y p_2 representan las probabilidades de

[a]sotres.davida@kendle.com

respuesta de los tratamientos estándar y nuevo. La hipótesis de interés (hipótesis de no-inferioridad) a ser probada es la alternativa (H_a) en el siguiente juego de hipótesis:

$$H_0 : p_1 - p_2 \geq d_0 \; vs \; H_a : p_1 - p_2 < d_0 \tag{1}$$

donde d_0 es el margen de no-inferioridad el cual es una constante positiva y conocida. En el contexto de ensayos clínicos los valores usuales para d_0 son 0.10, 0.15 y 0.20.

La estadística de prueba de FM se define como:

$$T(X_1, X_2) = \frac{\widehat{p}_1 - \widehat{p}_2 - d_0}{\widehat{\sigma}} \tag{2}$$

donde $\hat{p}_i = \frac{X_i}{n_i}$ es el estimador de máxima verosimilititud de p_i para $i = 1, 2$ y $\hat{\sigma}$ es el estimador de la desviación estándar de $\hat{d} = \hat{p}_1 - \hat{p}_2$; dado por:

$$\hat{\sigma} = \left(\frac{\check{p}_1(1 - \check{p}_1)}{n_1} + \frac{\check{p}_2(1 - \check{p}_2)}{n_2} \right)^{1/2}$$

donde $\check{p}_i$ es el estimador de máxima verosimilitud restringida bajo la hipótesis nula de p_i, ver Farrington y Manning(1990).

Para un nivel nominal de significancia igual a α, la región crítica para la prueba exacta de FM tiene la forma:

$$R_T(\alpha) = \{(x_1, x_2) \in \{0, ..., n_1\} \times \{0, ..., n_2\} : T(x_1, x_2) < T(x_1^*, x_2^*)\} \tag{3}$$

donde (x_1^*, x_2^*) se define en la siguiente sección.

Las tablas se calcularon para diseños balanceados, es decir, para $n_1 = n_2 = n$. Los niveles de significancia nominales empleados en este trabajo fueron $\alpha = 0.01$ y $\alpha = 0.05$. Las pruebas estadísticas serán simbolizadas igual que sus correspondientes estadísticas de prueba.

2.1. Estrategia para el cálculo del nivel de significancia

De acuerdo al modelo Binomial empleado en este trabajo tenemos que el espacio muestral es $\chi = \{(x_1, x_2) \in \{0, ..., n_1\} \times \{0, ..., n_2\}\}$, el espacio de parámetros es $\Theta = \left\{(p_1, p_2) \in [0, 1]^2\right\}$ y en virtud de que X_i tiene distribución Binomial con parámetros (n_i, p_i) para $i = 1, 2$, se tiene que la función de verosimilitud conjunta es:

$$L\left(p_1, p_2; x_1, x_2\right) = \binom{n_1}{x_1} p_1^{x_1} \left(1 - p_1\right)^{n_1 - x_1} \binom{n_2}{x_2} p_2^{x_2} \left(1 - p_2\right)^{n_2 - x_2}$$

y la función de potencia es $\beta_T(p_1, p_2) = \sum_{(x_1,x_2)\in R_T(\alpha)} L(p_1, p_2; x_1, x_2)$, además, el espacio nulo es $\Theta_0 = \{(p_1, p_2) \in \Theta : p_1 - p_2 \geq d_0\}$ y el nivel de significancia queda dado por $\sup_{(p_1,p_2)\in\Theta_0} \beta_T(p_1, p_2)$. El punto (x_1^*, x_2^*) de la ecuación (3) queda definido con la siguiente expresión

$$T(x_1^*, x_2^*) = \text{máx}\left\{T(a, b) : \sup_{(p_1,p_2)\in\Theta_0}\left(\sum_{T(x_1,x_2)\leq T(a,b)} L\left(p_1, p_2; x_1, x_2\right)\right) \leq \alpha\right\} \qquad (4)$$

Chan (1998) calculó el nivel de significancia para la prueba asintótica de Farrington-Manning tomando el supremo no en todo el espacio nulo (Θ_0) sino calculando el máximo únicamente en $\Theta_0^* = \{(p_1, p_2) \in \Theta : p_1 - p_2 = d_0\}$, el cual es solamente una parte de la frontera del espacio nulo. Computacionalmente ésto representa una inmensa ventaja, pues el tiempo de cómputo se reduce aproximadamente al 0.22 % del tiempo original. Sin embargo, el autor mencionado no justificó formalmente la validez de este argumento. Fue hasta 2005 cuando Röhmel (2005) presenta una prueba formal que justifica el procedimiento utilizado por Chan (1998). En este trabajo se siguió la misma estrategia de Chan(1998), para lo cual en lo que resta de esta sección se verifica la validez de la llamada condición de convexidad de Barnard y de la condición de simetría en la misma cola (ver definiciones abajo) para la prueba exacta de FM.

Definición 2.1. *Se dice que una prueba estadística, para el problema en (1), con región de rechazo R_T cumple la* ***condición de convexidad de Barnard*** *(C) si satisface las dos propiedades siguientes:*

1. $(x, y) \in R_T \Longrightarrow (x - 1, y) \in R_T \quad \forall \quad 1 \leq x \leq n_1,\ 0 \leq y \leq n_2$

2. $(x, y) \in R_T \Longrightarrow (x, y + 1) \in R_T \quad \forall \quad 0 \leq x \leq n_1,\ 0 \leq y \leq n_2 - 1$

Definición 2.2. *Si $n_1 = n_2 = n$, se dice que una región de rechazo R cumple* ***la condición de simetría en la misma cola*** *si $(x, y) \in R \Longrightarrow (n - y, n - x) \in R$.*

Almendra-Arao (2008) probó el siguiente resultado.

Proposición 2.1. *Sean $n_1 = n_2 = n$ y $R(\alpha)$ una región crítica para el problema de prueba de hipótesis en (1), si $R(\alpha)$ cumple la condición de convexidad de Barnard y la condición*

de simetría en la misma cola, entonces el nivel de significancia exacto de la prueba $R(\alpha)$ está dado por

$$\alpha^* = \max_{\substack{p_2=p_1-d_0 \\ p_1\in\left[d_0,\frac{1+d_0}{2}\right]}} \sum_{x_1=0}^{n_1}\sum_{x_2=0}^{n_2} \binom{n_1}{x_1} p_1^{x_1}\left(1-p_1\right)^{n_1-x_1} \binom{n_2}{x_2} p_2^{x_2}\left(1-p_2\right)^{n_2-x_2} I_{[(x_1,x_2)\in R(\alpha)]} \qquad (5)$$

La prueba exacta de FM satisface las condiciones de la proposición 2.1 debido a que:

1. Rohmel (2005) probó que la prueba exacta de FM satisface la condición (C) para cualesquiera n_1 y n_2.

2. En este trabajo se verificó numéricamente que la prueba exacta de FM satisface la condición de simetría en la misma cola usando programas de cómputo escritos en S-PLUS®. Estos programas pueden ser obtenidos por solicitud al tercer autor.

3. Conclusiones

Con base en lo comentado en A y B, el cálculo del nivel de significancia de la prueba exacta de FM se hizo aplicando la fórmula (5) y particionando el intervalo $[d_0, (1+d_0)/2]$ en subintervalos de longitud 0.001. Esto quiere decir que aproximamos el nivel de significancia exacto α^* reemplazando en la fórmula (5) al intervalo $[d_0, (1+d_0)/2]$ por el conjunto finito de puntos (p_1, p_2) tales que $p_1 = \{d_0 + (0.001)\, i : i = 0, 1, 2, ..., 500(1-d_0)\}$ y $p_2 = p_1 - d_0$; a esta aproximación la hemos llamado el nivel de significancia real y la denotamos por α_R. Los valores calculados de α_R son los que se reportan en las tablas. Las tablas se calcularon considerando las siguientes configuraciones $\alpha = 0.01, 0.05$, con $n_1 = n_2 = \{5, 6, 7, \ldots, 200\}$, y $d_0 \in \{0.05, 0.10, 0.15, 0.20, 0.25\}$.

En este artículo solo se presenta la tabla correspondiente al nivel nominal $\alpha = 0.05$, y $n_1 = n_2 = \{5, 6, 7, \ldots, 50\}$, el resto de las tablas pueden obtenerse solicitándolas al tercer autor. El valor crítico que aparece en las tablas corresponde al valor observado de $T(x_1^*, x_2^*)$ ver ecuación (4), es decir $C_\alpha = T(x_1^*, x_2^*)$. Para ilustrar el uso de las tablas, considere el siguiente ejemplo: sea $\alpha = 0.05$, $n_1 = n_2 = n = 31$ y $d_0 = 0.10$, entonces de la Tabla 2 se obtiene que la prueba exacta de FM rechaza H_0 en favor de la alternativa si $T(x_1^*, x_2^*) < -1.70224$ y el nivel de significancia real es igual a 0.04875. Como las pruebas exactas siempre mantienen el nivel nominal, pero pueden ser demasiado conservadoras, se calculó el porcentaje de niveles

de significancia que pertenecen al intervalo $[0.80\alpha, \alpha]$; es decir, el porcentaje para el cual el nivel de significancia está muy cercano al nivel de significancia nominal elegido para la prueba. Con base en los resultados de la Tabla 1 podemos concluir que la prueba exacta de FM no es demasiado conservadora. Además por cálculo directo se verificó que el nivel de significancia de esta prueba está muy cercano al nivel nominal de la prueba. Así que para el nivel nominal $\alpha = 0.05$ se verificó que $|0.05 - \text{nivel de significancia}| \leq 0.005$, para toda $31 \leq n \leq 200$. Un resultado similar se obtuvo para el nivel nominal $\alpha = 0.01$.

α / d_0	**n**	**0.05**	**0.10**	**0.15**	**0.20**	**0.25**
	$5 \leq n \leq 30$	80.77 %	88.46 %	92.31 %	84.62 %	76.92 %
0.01	$31 \leq n \leq 100$	98.59 %	98.59 %	98.59 %	98.59 %	98.59 %
	$101 \leq n \leq 200$	100.00 %	100.00 %	100.00 %	100.00 %	100.00 %
	$5 \leq n \leq 30$	88.46 %	96.15 %	84.62 %	76.92 %	84.62 %
0.05	$31 \leq n \leq 100$	98.59 %	98.59 %	98.59 %	98.59 %	98.59 %
	$101 \leq n \leq 200$	100.00 %	100.00 %	100.00 %	100.00 %	100.00 %

Tabla 1: Porcentaje de niveles de significancia reales que caen dentro del intervalo $[0.80\alpha, \alpha]$, con base en el número total de tamaños de muestra en los rangos $5 \leq n \leq 30$, $31 \leq n \leq 100$, $101 \leq n \leq 200$.

Referencias

Almendra-Arao, F. 2008. *A Study on the Classical Asymptotic Non-inferiority Test for Two Binomial Proportions*. Drug Information Journal. En prensa.

Chan,I. 1998. *Exact tests of equivalence and efficacy with a non zero lower bound for comparative studies.* Statistics in Medicine, **17**,1403-1413.

Chen,J.,Tsong,Y.,y Kang,S. 2000. *Tests for equivalence or noninferiority between two proportions.* Drug Information Journal, **34**, 569-578.

Farrington,C. y Manning,G. 1990. *Test statistics and sample size formulae for comparative binomial trials with null hypothesis of non-zero risk difference or non-unity relative risk.* Statistics in Medicine,**9**,1447-1454.

Martin A., A., y I. Herranz T. 2004. *Exact unconditional non-classics tests on the difference of two proportions.* Computational Statistics & Data Analysis, **45**, 373-388.

Röhmel,J. 2005.*Problems with existing procedures to calculate exact unconditional p-values for noninferiority and confidence intervals for two binomials and how to resolve them.* Biometrical Journal, **47**, 37-47.

n	d_0=0.05		d_0=0.10		d_0=0.15		d_0=0.20		d_0=0.25	
	c_α	Tamaño	c_α	Tamaño	c_α	Tamaño	c_α	Tamaño	c_α	Tamaño
5	-2.05805	0.03919	-1.87238	0.04807	-2.02575	0.03334	-2.18654	0.02256	-2.14800	0.03840
6	-1.93189	0.03971	-1.86960	0.04646	-2.03617	0.03076	-1.96082	0.03922	-1.91796	0.04019
7	-1.86313	0.04434	-1.87757	0.04475	-2.05731	0.02906	-1.98421	0.03334	-1.99705	0.03886
8	-1.82293	0.04445	-1.89162	0.04915	-1.80112	0.04716	-1.86827	0.04528	-1.79902	0.04827
9	-1.79899	0.04550	-1.88769	0.04263	-1.83067	0.04312	-1.90769	0.03943	-1.88163	0.03909
10	-1.78516	0.04621	-1.87122	0.04121	-1.86444	0.03935	-1.85409	0.04511	-1.96251	0.03458
11	-1.77802	0.04659	-1.86612	0.03993	-1.81054	0.04385	-1.88198	0.03939	-1.85772	0.04224
12	-1.77553	0.04699	-1.77239	0.04792	-1.75779	0.04880	-1.84033	0.04454	-1.80587	0.04462
13	-1.87319	0.03765	-1.77023	0.04661	-1.79237	0.04449	-1.88108	0.03944	-1.78679	0.04744
14	-1.77970	0.04769	-1.77391	0.04525	-1.82849	0.04055	-1.85164	0.04995	-1.78848	0.04779
15	-1.77021	0.04801	-1.74720	0.04933	-1.73553	0.04846	-1.83392	0.04814	-1.80385	0.04854
16	-1.79154	0.04482	-1.77469	0.04656	-1.75993	0.04776	-1.74607	0.04856	-1.85122	0.04164
17	-1.73801	0.04997	-1.75946	0.04764	-1.80342	0.04331	-1.80689	0.04451	-1.85979	0.04804
18	-1.72855	0.04898	-1.72937	0.04856	-1.83863	0.04064	-1.75260	0.04613	-1.77521	0.04447
19	-1.72225	0.04895	-1.74207	0.04670	-1.76928	0.04560	-1.74800	0.04570	-1.75472	0.04654
20	-1.71848	0.04879	-1.75644	0.04485	-1.72251	0.04966	-1.76777	0.04482	-1.72516	0.04837
21	-1.71676	0.04853	-1.71288	0.04965	-1.75219	0.04601	-1.74702	0.04647	-1.75427	0.04699
22	-1.71671	0.04819	-1.73980	0.04675	-1.71831	0.04882	-1.73750	0.04809	-1.75393	0.04661
23	-1.71807	0.04779	-1.72535	0.04803	-1.72621	0.04804	-1.71493	0.04927	-1.76365	0.04920
24	-1.72059	0.04969	-1.74103	0.04603	-1.72120	0.04870	-1.74141	0.04850	-1.79357	0.04721
25	-1.77541	0.04671	-1.71931	0.04649	-1.74733	0.04560	-1.75138	0.04794	-1.70371	0.04983
26	-1.76072	0.04630	-1.70660	0.04846	-1.71191	0.04862	-1.76517	0.04749	-1.70767	0.04982
27	-1.74886	0.04649	-1.78094	0.04530	-1.71751	0.04913	-1.77985	0.04730	-1.71766	0.04943
28	-1.73929	0.04724	-1.81068	0.04565	-1.74025	0.04634	-1.70951	0.04922	-1.77422	0.04347
29	-1.72557	0.04828	-1.69629	0.04938	-1.74225	0.04492	-1.76634	0.04679	-1.72692	0.04955
30	-1.72601	0.04745	-1.71153	0.04755	-1.72873	0.04824	-1.70356	0.04975	-1.74807	0.04739
31	-1.72165	0.04744	-1.70224	0.04875	-1.72874	0.04885	-1.70339	0.04943	-1.77389	0.04809
32	-1.71854	0.04735	-1.69199	0.04986	-1.73167	0.04894	-1.70757	0.04950	-1.69675	0.04859
33	-1.71653	0.04721	-1.70694	0.04802	-1.73168	0.04910	-1.69934	0.04935	-1.74947	0.04473
34	-1.70582	0.04849	-1.69461	0.04734	-1.74444	0.04970	-1.72585	0.04814	-1.72300	0.04785
35	-1.71528	0.04681	-1.69723	0.04847	-1.75876	0.04680	-1.71810	0.04959	-1.71641	0.04921
36	-1.70604	0.04951	-1.70675	0.04819	-1.76663	0.04728	-1.75372	0.04605	-1.72504	0.04942
37	-1.71386	0.04750	-1.71181	0.04730	-1.68920	0.04938	-1.75728	0.04694	-1.74589	0.04890
38	-1.71885	0.04700	-1.69551	0.04852	-1.71704	0.04809	-1.69442	0.04925	-1.69372	0.04998
39	-1.72119	0.04649	-1.70147	0.04825	-1.74495	0.04683	-1.67628	0.04934	-1.70854	0.04984
40	-1.68549	0.04935	-1.68043	0.04950	-1.69505	0.04915	-1.70765	0.04828	-1.69709	0.04992
41	-1.68772	0.04893	-1.69084	0.04930	-1.68946	0.04957	-1.69991	0.04814	-1.71974	0.04798
42	-1.69041	0.04849	-1.69140	0.04908	-1.70535	0.04749	-1.70122	0.04947	-1.72564	0.04813
43	-1.68338	0.04989	-1.69586	0.04867	-1.68672	0.04884	-1.69963	0.04905	-1.73278	0.04917
44	-1.75182	0.04730	-1.70938	0.04709	-1.70047	0.04809	-1.71693	0.04909	-1.76518	0.04654
45	-1.74385	0.04682	-1.75261	0.04538	-1.69505	0.04846	-1.72749	0.04913	-1.68870	0.04991
46	-1.73442	0.04895	-1.69516	0.04839	-1.71165	0.04809	-1.74843	0.04761	-1.69416	0.04809
47	-1.72804	0.04844	-1.69106	0.04882	-1.70246	0.04980	-1.68201	0.04998	-1.71043	0.04851
48	-1.72259	0.04834	-1.68756	0.04850	-1.70716	0.04927	-1.69423	0.04883	-1.70793	0.05000
49	-1.71211	0.04997	-1.67050	0.04896	-1.71818	0.04954	-1.72939	0.04518	-1.73302	0.04735
50	-1.70743	0.04986	-1.68092	0.04994	-1.72232	0.04963	-1.69433	0.04956	-1.75701	0.04796

Tabla 2. Constantes críticas (C_α) y tamaños de la prueba (Tamaño) para la prueba exacta de FM, para α=0.05

Memoria FONE XXIII (2009), Sección II, 173–179

Análisis de confiabilidad para la predicción de vida útil de alimentos

Fidel Ulín-Montejo[a], Rosa Ma. Salinas-Hernández
Ciencias Básicas–Ciencias Agropecuarias – Universidad Juárez Autónoma de Tabasco

1. Introducción

La evaluación sensorial es un factor clave para determinar la vida de anaquel de muchos alimentos. Alimentos estables tendrán una vida de anaquel definida por los cambios en sus propiedades sensoriales. Muchos alimentos, como los frutos frescos, después de almacenamiento relativamente prolongado, pueden ser seguros para comerlos, pero pueden ser rechazados debido a cambios sensoriales (Hough, *et al.*, 2003; Salinas-Hernández, *et al.*, 2007). Por otro lado, el análisis de confiabilidad es una rama de la estadística usada extensamente en el análisis de tiempos de vida en estudios clínicos y de tiempos a la falla en productos y sistemas (Kalbfleish y Prentice, 1980; Lawless, 1982; Klein y Moeschberger, 1997; Meeker y Escobar, 1998; Ulín-Montejo, 2007). Estos métodos empiezan a ser utilizados en estudios de calidad sensorial y nutritiva de los alimentos (Hough, *et al*, 2003; Salinas-Hernández, *et al*, 2007). Por otro lado, (Cardelli y Labuza, 2001; Gacula y Singh, 1984; Salinas-Hernández, *et al.*, 2007) presentan el modelo Weibull, obtenido del análisis de confiabilidad, en estudios de vida de anaquel de alimentos. Aquí se define el fenómeno de censura, un concepto clave en confiabilidad y supervivencia escasamente abordado en estudios de vida útil de alimentos.

2. Censura

En estudios de vida de anaquel de alimentos, se define una variable aleatoria T como el tiempo al cual el consumidor rechaza la muestra, la función de confiabilidad $S(t)$ se define como la probabilidad de que un consumidor acepte un producto mas allá del tiempo t,

[a]fidel.ulin@basicas.ujat.mx

$S(t) = P(T > t)$. Para ilustrar los datos censurados, suponga que los tiempos de almacenamiento son 0, 9, 18,. . . , 63, 72 y 81 h. Debido a que los tiempos son discretos, T nunca será observada exactamente, en vez de ello solo se observará que $T \leq 9, 63 < T \leq 72, T \geq 81$. Ésto es considerado como información censurada en análisis de confiabilidad (Hough, *et al*, 2003): **Censura-por-la-Izquierda:** Si un consumidor rechaza la muestra con 9 h de almacenamiento, el tiempo de rechazo sera a las 9 h. Este consumidor es sensible a cambios ocurridos durante el almacenamiento y rechaza el producto en algún momento entre 0 y 9 h. **Censura por Intervalo:** Si un consumidor acepta muestras almacenadas a un tiempo de 0 a 9 h, pero rechaza la muestra almacenada por 18 h, el tiempo de rechazo estara en $9 < T \leq 18$ h. Las limitaciones de recursos, o consideraciones experimentales, impiden ofrecer muestras a 10, 11, 12, . . . , 17 y 18 h. **Censura por la Derecha:** Si un consumidor acepta todas las muestras, se diría que el tiempo de rechazo es ≥ 81 h. Esto es, si una muestra es almacenada hasta un tiempo suficientemente largo, el consumidor finalmente la rechazará.

3. Metodología

3.1. Función de Verosimilitud

El principal objetivo de la inferencia por máxima verosimilitud es ajustar modelos por medio de las combinaciones modelos-parámetros, para los cuales la probabilidad de los datos sea alta. Los métodos de máxima verosimilitud pueden aplicarse a una amplia variedad de modelos con datos censurados y co-variables explicatorias (Meeker y Escobar, 1998). La función de verosimilitud es utilizada para estimar la función de supervivencia (Lawless, 1982), la cual se define como la probabilidad conjunta de las datos obtenidos de los n consumidores:

$$L(\theta) = \prod_{i \in R} [S(r_i; \theta)] \prod_{i \in L} [1 - S(l_i; \theta)] \prod_{i \in I} [S(l_i; \theta) - S(r_i; \theta)] \tag{1}$$

Donde R es el conjunto de observaciones censuradas por la derecha r_i, L es el conjunto de observaciones censuradas por la izquierda l_i, I es el conjunto de los las observaciones censuradas por intervalo y θ el vector de parámetros (μ, σ). En (1) se muestra como cada uno de los tipos de censura contribuye de manera diferente a la función de verosimilitud.

3.2. Modelos y estimación por Máxima Verosimilitud

Con base en estudios anteriores y la información en los datos, puede suponerse una distribución adecuada; los modelos paramétricos proporcionan estimaciones más precisas para la función de confiabilidad y otras cantidades de interés.

Usualmente los tiempos a la falla no siguen una distribución normal, sino distribuciones sesgadas a la derecha. Frecuentemente se elige una distribución de log-localización-escala para T, tal que su función de distribución $F_T(t) = \Phi[(log(t) - \mu)/\sigma]$; donde Φ no depende de los parametros, μ es el parámetro de localización y σ de escala, y con funcion log-cuantil

$$y_p = log(t_p) = \mu + \Phi^{-1}(p)\sigma. \tag{2}$$

Donde *log* denotara el *logaritmo natural*. En Meeker y Escobar (1998) se muestran distribuciones para Φ; para T lognormal (LN), se tiene la distribución normal estándar Φ; si T es Weibull (W), la distribución es la de valores mínimos extremos $\Phi_{sev}(w) = exp[exp(w)]$. Si se eligen los modelos lognormal y Weibull, sus funciones de confiabilidad estan dadas por:

$$S_{LN}(t) = 1 - \Phi\left[\frac{log(t) - \mu}{\sigma}\right]; \qquad S_W(t) = \Phi_{sev}\left[\frac{log(t) - \mu}{\sigma}\right]. \tag{3}$$

Los parámetros de los modelo de log-localización-escala son obtenidos maximizando la función de verosimilitud (1). Una vez construida la verosimilitud para un modelo dado, softwares estadísticos como R y $Splus$ pueden usarse para estimar μ y σ que maximizan (1); esto se realiza numéricamente resolviendo simultáneamente el siguiente sistema de ecuaciones:

$$\partial\left[logL(\mu,\sigma)\right]/\partial\mu = 0; \qquad \partial\left[logL(\mu,\sigma)\right]/\partial\sigma = 0. \tag{4}$$

En Análisis de Confiabilidad, el tiempo medio a la falla se define como: $E(T) = \int_0^\infty S(t)dt$. En este estudio sensorial de vida de anaquel, $M(T)$ y $E(T)$ representan la *mediana* y el *tiempo medio* de almacenamiento, respectivamente, a los cuales se rechazaría el producto. Para los modelos lognormal y Weibull, para el que $\eta = exp(\mu)$ y $\beta = 1/\sigma$, estan dadas por:

$$M_{LN}(T) = \exp(\mu), \qquad E_{LN}(T) = \exp(\mu + \sigma^2/2). \tag{5}$$

$$M_W(T) = \exp(\mu)[log(2)]^\sigma, \qquad E_W(T) = \exp(\mu)\Gamma(1+\sigma). \tag{6}$$

Tiempo de almacenamiento en horas												
Consumidor	Frecuencia	0	9	18	27	36	45	54	63	72	81	Censura
1	21	S	S	S	S	S	S	S	S	N	N	Intervalo: 63 - 72
2	14	S	S	N	N	S	S	S	N	N	N	Intervalo: 9 - 63
3	7	S	S	S	S	S	S	S	S	S	S	Derecha: > 81
4	5	S	N	N	N	N	N	N	N	N	N	Izquierda: < 9
5	6	N	N	S	S	S	S	S	N	N	N	No se considera

Tabla 1: Datos de aceptación/rechazo (S/N) para los consumidores de mango 'Ataulfo'.

4. Predicción de vida de anaquel

4.1. Estudio de consumidores

Se utilizó un fruto tropical fresco cortado, mango 'Ataulfo'. Cajas de mangos fueron compradas a un distribuidor local. Los mangos fueron pelados y cortados en cubos de 8 cm^3 para ser almacenados en recipientes de plástico transparente de 150 grs, a una temperatura de 10 ^oC, en un numero considerable para cubrir 10 tiempos de almacenamiento de 9 h cada uno; esto es, 0, 9, 18, 27, 36, 45, 54, 63, 72 y 81 h. Previo análisis microbiológico se determino que las muestras fueran inocuas para su consumo.
Se consideraron 53 consumidores, proporcionandoles muestras de mango almacenados a distintos tiempos (0, 9, ..., 72 y 81 h.) una a la vez y en orden aleatorio. Los sujetos probaron cada muestra y respondieron la pregunta: ¿Usted consumiría este producto? *Si* o *No*.

4.2. Resultados

La Tabla 1 muestra los datos para 5 consumidores típicos, de los 53 considerados en total, ilustrandose la interpretación dada a la información de cada uno de los consumidores.

El consumidor 1 aceptó las muestras hasta las 72 h, teniendose entonces una observación censurada por intervalo entre 63 y 72 h; de igual modo, para el consumidor 2, quien rechazó a las 18 h, aceptó a 36 y rechazó otra vez a las 63 h. El consumidor 3 aceptó todas las muestras, considerándose información censurada por la derecha al tiempo de 81 h. Para el consumidor 4, que rechazó la muestra desde la primera prueba, su información fue censurada por la

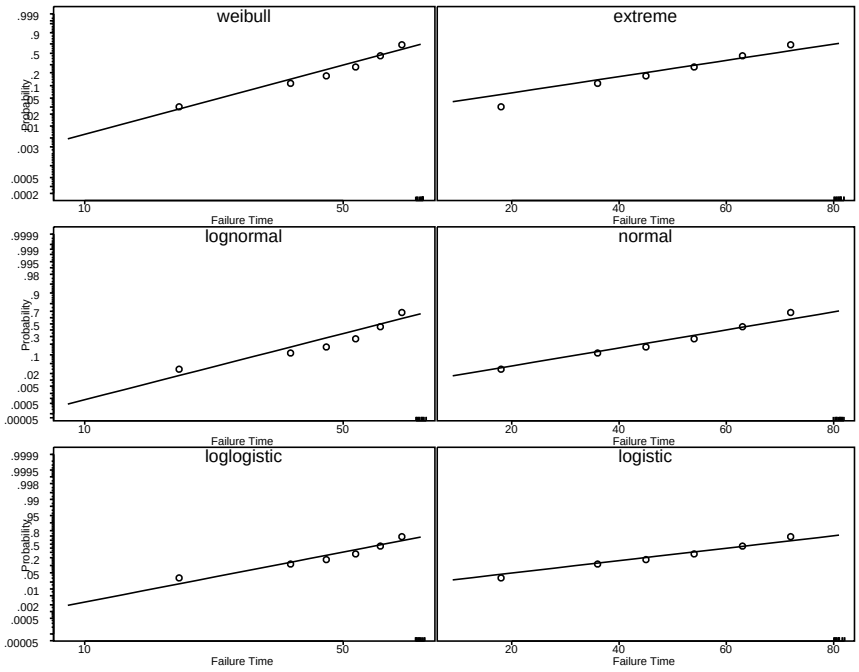


Figura 1: Probabilidad del rechazo comparado con el tiempo de almacenamiento.

izquierda a las 9 h. El consumidor 5 rechazo la muestra fresca, quizá por error o por que no gusta del mango fresco, para este estudio es razonable no considerar esta información.

Ahora, con los gráficos de probabilidad para las funciones log-cuantil (2) de seis distribuciones de log-localización-escala típicas (Figura 1) y su log-verosimilitud $[logL(\mu, \sigma)]$, se comparan los ajustes de los modelos (Meeker y Escobar, 1998). Dos de éstas fueron seleccionadas, lognormal y Weibull; por buen ajuste y porque han sido estudiadas en modelación de vida de anaquel de alimentos (Gacula y Singh, 1984). Las estimaciones, a través de máxima verosimilitud, fueron:

$$\text{lognormal: } \mu = 4.150,\ es_{\mu} = 0.096;\ \sigma = 0.587,\ es_{log(\sigma)} = 0.160;\ logL = -68.5. \quad (7)$$

$$\text{Weibull: } \mu = 4.325,\ es_{\mu} = 0.072;\ \sigma = 0.399,\ es_{log(\sigma)} = 0.188;\ logL = -67.5. \quad (8)$$

Con éstas estimaciones pueden construirse los gráficos para las funciones de confiabilidad con ambos modelos, lognormal y Weibull. Si embargo, aquí lo mas relevante es estimar algunas medidas importantes como la mediana $M(T)$, probabilidad de 0.50 del rechazo del consumidor, adoptada por algunos autores (Cardelli y Labuza, 2001) como determinante para

definir la vida de anaquel. Junto a su intervalo del 95 porciento de confianza, se obtuvieron:

$$M_{LN}(T) = 63.27\ (52.48, 76.29); \qquad M_W(T) = 65.27\ (56.13, 76.84). \tag{9}$$

Otra cantidad de interes es el tiempo de vida media de anaquel $E(T)$, que puede ser obtenida de los parámetros estimados del modelo, como en (4) y (5).

$$E_{LN}(T) = 75.36, \qquad E_W(T) = 67.05. \tag{10}$$

Como ambos modelos son sesgados, el tiempo medio de vida de anaquel es mayor que la mediana; esto es, $E(T) > M(T)$. Ambos modelos ofrecen resultados de inferencia similares.

5. Conclusiones

Para determinar la vida de anaquel sensorial de alimentos, el enfoque ha sido establecido sobre la probabilidad de rechazo del producto después de cierto tiempo de almacenamiento. Los conceptos de análisis de confiabilidad han sido presentados, mostrándose los diferentes tipos de datos censurados que deben ser considerados en un estudio de vida de anaquel. Un aspecto interesante de la metodología presentada es que la experimentación es relativamente sencilla, ya que los consumidores solo prueban muestras de mango con diferentes tiempos de almacenamiento, respondiendo únicamente *"si"* o *"no"* a consumir cada muestra. Esta información es suficiente para modelar la probabilidad de rechazo a distintos tiempos de almacenamiento y temperaturas; estimándose así la vida de anaquel.

Referencias

Cardelli, C. and Labuza, T.P. 2001. Application of Weibull hazard analysis to the determination of the shelf life of roasted coffe. *Lebensm Wiss u Technol* **34**(5): 273-278.

Gacula, M.C. and Singh, J. 1984. *Statistical methods in food and consumer research*. New York: Academic Press. 505 p.

Hough, G., Langohr, K., Gomez, G., and Curia, A. 2003. Survival analysis applied to sensory shelf life of foods. *J. of Food Sci.: Sensory and Nutritive Qualities of Food*, **68**(1):359-362.

Kalbfleisch, J.D. and Prentice, R.L. 1980. *The Statistical Analysis of Failure Time Data*, New York: Wiley. 321 p.

Klein, J.P. and Moeschberger, M.L. 1997. *Survival analysis techniques for censored and truncated data*, New York: Springer. 502 p.

Lawless, J.F. 1982. *Statistical models and methods for lifetime data*. New York: Wiley. 592 p.

Meeker, W.Q. and Escobar, L.A. 1998. *Statistical Methods for Reliability Data*. New York: Wiley. 680p.

Salinas-Hernández, R.M., González-Aguilar, G.A., Pirovani, M.E. and Ulín-Montejo, F. 2007. Modelling of the deterioration of fresh vegetables. *Universidad y Ciencia*, **23**(2): 183-196.

Ulín-Montejo, F. 2007. Análisis de Datos Censurados para Ingeniería y Ciencias Biológicas *Revista de Matemática: Teoría y Aplicaciones*, **14**(2): 239-250.

Memoria FONE XXIII (2009), Sección II, 181–185

Análisis psicométrico del funcionamiento diferencial del cuestionario QUALEFFO basado en la TRI y en regresión logística

Purificación Vicente Galindo[a], Mercedes Sánchez Barba, Purificación Galindo Villardón, Jose Luís Vicente Villardón
Universidad de Salamanca – España

1. Introducción

Es muy difícil dar una definición de *Calidad de Vida*, porque significa diferentes cosas para diferentes personas, para las mismas personas en diferentes momentos e incluso, tiene diferentes significados dependiendo del área de aplicación. Una de las áreas en las que éste término ha adquirido una mayor importancia es en el ámbito de la salud.

No hay, por el momento, una definición universalmente aceptada. En la mayoría de los trabajos se identifica el concepto con las dimensiones que evalúa el cuestionario que se utiliza; este trabajo se centra en el estudio del QUALEFFO (Questionnaire of the European Foundation for Osteoporosis, Lips *et al*, 1996), un cuestionario específico para evaluar Calidad de Vida en pacientes osteoporóticos, que fue diseñado para su uso en mujeres hospitalizadas que habían sufrido fractura de cadera. Hoy se sabe que los hombres están cada vez más afectados hasta el punto de que en el *American Journal of Medicine* (1993; pp: 646-650) se afirma que el riesgo de padecer fractura de cadera, a lo largo de la vida, en hombres, es mayor que el de padecer cáncer de próstata. Por esta razón la Organización Mundial de la Salud ha recomendado la detección de la Osteoporosis en el ámbito de Atención Primaria. Sánchez (2008), probó que el QUALEFFO sigue siendo válido al ser aplicado en *screening* en pacientes ambulatorios y demostró también que puede ser simplificado sin perder sus propiedades psicométricas.

[a]primer_purivg@usal.es

En términos de Calidad de Vida un ítem presenta Funcionamiento Diferencial (*DIF*) cuando sujetos con un mismo nivel de Calidad de Vida no eligen la misma categoría de respuesta del ítem que evalúa ese nivel. Por ejemplo hombres y mujeres para un mismo nivel de dolor pueden percibirlo unos como moderado y los otros como fuerte. Si esto ocurre para todos los niveles de Calidad de Vida, se dice que el *DIF* es *uniforme* . Si sólo ocurre para algunos niveles de Calidad de Vida, hablaremos de un *DIF no Uniforme*. Normalmente, el grupo objeto de análisis se denomina *grupo focal* y el grupo que sirve como criterio de comparación se conoce como *grupo de referencia*.

2. Método

2.1. Participantes

La muestra está formada por 741 pacientes de varios Centros de Atención Primaria de la Comunidad de Castilla y León, a los que se les realizó una medición ultrasónica en el calcáneo, obteniendo el BUA. Este último valor permitió clasificar a los pacientes estudiados en sujetos con algún grado de Osteoporosis y Normales. Así, la muestra quedó dividida de la siguiente manera: 516 pacientes fueron considerados Osteoporóticos y 192 como Normales. De los pacientes Osteoporóticos casi un 26 % son hombres y un 74 % mujeres.

2.2. Instrumento

La versión adaptada al español (Badia y Hermand, 1999) del cuestionario QUALEFFO incluye 35 ítems que consideran 7 dimensiones: 5 ítems sobre dolor, 3 ítems sobre actividades cotidianas, 5 ítems sobre tareas domésticas, 6 ítems sobre movilidad, 4 ítems sobre actividades sociales, 2 ítems sobre la calidad de vida relacionada con la salud global, y 10 ítems sobre el funcionamiento mental. La respuesta a cada ítem se presenta en escala tipo Lickert, desde 1 (ningún problema) a 5 (muchos problemas). La puntuación de cada dimensión del cuestionario se obtiene sumando los valores de respuesta de cada ítem y dividiéndolo por el total de ítems respondidos de la dimensión. La puntuación total se obtiene sumando la puntuación en cada dimensión y dividiéndola entre el número de dimensiones del cuestionario.

La media de tiempo de administración, para el QUALEFFO fue de 20 minutos, que para una consulta de Atención Primaria es mucho tiempo, debido a que el número de pacientes

que acuden diariamente a la consulta es muy elevado.

2.3. Procedimiento

La evaluación del *DIF* del cuestionario QUALEFFO se realizó utilizando dos procedimientos diferentes: el modelo *DFIT* que utiliza el nivel de Calidad de Vida como variable de equiparación, estimada bajo un Modelo de Respuesta al Ítem (TRI) y el análisis mediante regresión logística ordinal o modelo logit con respuesta ordinal basado en las puntuaciones observadas en el test.

El modelo DFIT (Differential Functioning of Items and Tests) fue propuesto por Raju, van der Linden y Fleer (1995) en el marco de la TRI que permite estudiar el funcionamiento diferencial de los ítems con el estadístico *Noncompensatory Differential Item Functioning* (*NCDIF*) . Este estadístico refleja las diferencias de la puntación verdadera para dos grupos de individuos (hombres y mujeres). Denotando por d_{ij} a la diferencia de las puntuaciones verdaderas para el grupo focal y para el grupo de referencia, $(ES_{ijF} - ES_{ijR})$ entonces:

$$NCDIF_i = \sigma^2_{d_i} + \mu^2_{d_i}$$

donde $\sigma^2_{d_i}$ y $\mu^2_{d_i}$ son las varianzas y medias de d_{ij} respectivamente.

El estadístico *NCDIF* se considera significativo en el contexto de la significación del estadístico chi-cuadrado y en asociación del índice que excede a priori en un valor crítico específico que según Raju en una comunicación en 1999 recomienda que para el valor *NCDIF* para un ítem con 5 categorías (como los ítems del cuestionario QUALEFFO) sea 0.096.

Regresión logística ordinal o modelo logit con respuesta ordinal. La ecuación general de un modelo de regresión logística para el análisis del *DIF* para ítems politómicos ordinales es la siguiente:

$$logit\left[P(Y \geq k)\right] = \log\left[\frac{P(Y \geq k)}{1 - P(Y \geq k)}\right] = \log\left[\frac{\pi_k + ... + \pi_{m-1}}{\pi_0 + ... + \pi_{k-1}}\right], k = 1, 2, ...m$$

donde Y es la variable respuesta, m es el número de categorías de respuesta del ítem considerado, $\pi_0, \pi_1, \ldots, \pi_{m-1}$, las probabilidades de respuesta categórica y P es la probabilidad de responder a la categoría k o superior del ítem. En nuestro estudio $m = 5$.

El análisis de regresión logística es uno de los métodos de comprobada eficacia para su uso en la detección tanto del *DIF uniforme* como *DIF no-uniforme* (Clauser y Mazor, 1998).

3. Resultados

Se ha evaluado el *DIF* del QUALEFFO con el modelo *DFIT*, para identificar si los hombres y las mujeres osteoporóticos (516 pacientes), con un mismo nivel de nivel de Calidad de Vida, tienen distinta probabilidad de elegir una determinada respuesta.

El ítem **C13**, de la dimensión *Tareas Domésticas* y los ítems **G27**, **G29**, **G30**, **G33**, **G34** y **G35** de la dimensión *Estado de Ánimo* presentan *DIF*, es decir, la probabilidad de responder a una categoría k o superior de estos ítems es distinta en el grupo de los hombres que en el grupo de las mujeres. Eto es debido a que para cada uno de estos ítems, el estadístico χ^2 es significativo ($p < 0.01$) y el valor del índice *NCDIF* es superior al valor crítico 0.096 recomendado por Raju (1999). Para obtenener el índice *NCDIF* se han calculado los parámetros de los ítems para el grupo de las mujeres y el grupo de los hombres por seprado con el modelo de Respuesta Graduada (Samejima, 1969) y se han puesto en la misma métrica por el método de Stocking-Lord.

El modelo de Regresión logística nos ha permitido identificar cinco ítems con *DIF uniforme* y cinco ítems con *DIF no uniforme*. El ítem **B8** de la dimensión *Actividades Cotidianas*, los ítems **C10** y **C13** de la dimensión *Tareas Domésticas*, el ítem **D19** de la dimensión *Movilidad* y el ítem **G35** de la dimensión *Estado de Ánimo* presentan *DIF uniforme*, es decir, al mismo nivel de Calidad de Vida, los hombres y las mujeres no eligen la misma respuesta categórica de estos ítems. Presentan *DIF no uniforme*, es decir, la respuesta al ítem **D15**, de la dimensión *Movilidad* y a los ítems **G27**, **G28**, **G29** y **G30** de la dimensión *Estado de Ánimo* para el grupo de los hombres y de las mujeres puede o no ser la misma, para distintos niveles de Calidad de Vida.

4. Conclusiones

Los resultados obtenidos con la regresión logística son totalmente coincidentes con los del modelo *DFIT* en las escalas: *Dolor*, *Actividades Sociales y de Tiempo Libre* y *Percepción de la Salud General*. Con la regresión logística se detecta cinco ítems con *DIF uniforme* y cinco ítems con *DIF no uniforme*, mientras con el modelo *DFIT* sólo se corrobora en tres de los ítems anteriores.

Los ítems que presentan DIF sería mejor no incluirlos en el cuestionario, ya que son percibidos de diferente forma por el grupo de mujeres y por el grupo de los hombres.

Referencias

Badía, Xavier y Michael Herdman. 1999. Adaptación Transcultural al Español de los Cuestionarios OQLQ y QUALEFFO para la Evaluación de la Calidad de Vida Relacionada con la Salud de Mujeres con Fractura Vertebral Osteoporótica, *Rev Esp Enf Met Oseas*, **8**, 135-140.

Clauser, Brian E. y Kathleen M. Mazor. 1998. Using Statistical Procedures to Identify Differentially Functioning Test Items, *Educational Measurement: Issues and Practice*, **17**, 31-44.

Lips, Paul *et al.* 1996. The Development of a European Questionnaire for Quality of Life in Patients with Vertebral Osteoporosis, *Scand J Rheumatol (Suppl)*, **103**, 84-85.

Raju, Nambury S., Wim J. Van Der Linden y Paul F. Fleer. 1995. IRT-based Internal Measures of Differential Functioning of Items and Tests, *Applied Psychological Measurement*, **19**, 353-368.

Raju, Nambury S. 1999. *DFIT5P: A Fortran Program for Calculating DIF/DTF [Computer software]*. Illinois Institute of Technology: Chicago.

Samejima, Fumiko. 1969. Calibration of Latent Ability Using a Response Pattern of Graded Scores. *Psychometrika Monograph Supplement*,**17**.

Sánchez, Mercedes. 2008. *Aportaciones al Análisis de Datos de Calidad de Vida Relacionada con la Salud, desde una Perspectiva Multivariante*. Tesis Doctoral: Universidad de Salamanca.

Memoria FONE XXIII (2009), Sección II, 187–192

Una prueba de bondad de ajuste para la distribución pareto generalizada

José A. Villaseñor Alva[a], Elizabeth González Estrada
Colegio de Postgraduados

1. Introducción

La distribución Pareto Generalizada con parámetros σ y γ (PG(σ, γ)) tiene función de distribución

$$F(x; \sigma, \gamma) = 1 - \left(1 + \frac{\gamma}{\sigma} x\right)^{-1/\gamma} \tag{1}$$

con $\sigma > 0$ y $\gamma \in R$ tales que $x > 0$ para $\gamma \geq 0$ y $0 < x < -\sigma/\gamma$ cuando $\gamma < 0$.

Cuando $\gamma \to 0$, $F(x; \sigma, \gamma) \to 1 - \exp(-x/\sigma)$, la cual es la distribución Exponencial(σ). También note que cuando $\gamma = -1$, $F(x; \sigma, \gamma) = x/\sigma$, la cual es la distribución Uniforme$(0, \sigma)$.

Esta familia de distribuciones contiene distribuciones de cola pesada, la familia de distribuciones exponencial, así como una subclase de distribuciones Beta y otras de soporte acotado. Debido a su riqueza, la familia de distribuciones PG ha sido usada para modelar probabilidades en diferentes campos como Finanzas, Ecología e Hidrología entre otras (Ver el libro de Reiss y Thomas, 2001).

Es posible argumentar la factibilidad de la distribución PG para ajustar una muestra aleatoria mediante argumentos basados en la teoría asintótica de excedencias sobre un umbral, así como mediante el uso de métodos exploratorios de datos como el uso de la función de vida media residual (Reiss y Thomas, 2001).

Sin embargo, estas técnicas no proporcionan una medida de error para aceptar la distribución PG cuando de hecho no es el modelo correcto.

Por lo tanto, se requiere contar con una prueba de bondad de ajuste eficiente para H_0 : *F es una distribución PG,* con base en una muestra aleatoria de F.

[a] jvillasr@colpos.mx

2. Estimación del parámetro γ

2.1. Estimación de Hill: caso $\gamma \geq 0$

La distribución Pareto(γ) con parámetro de forma γ se define como $F(x;\gamma) = 1 - x^{-1/\gamma}$, $x > 1$. Entonces $\lim_{x\to\infty} \dfrac{\overline{F}(x;\gamma)}{\overline{F}(x;\sigma,\gamma)} = \lim_{x\to\infty} \dfrac{x^{-1/\gamma}}{\left(1+\frac{\gamma}{\sigma}x\right)^{-1/\gamma}} = \left(\dfrac{\sigma}{\gamma}\right)^{1/\gamma}$ donde $\overline{F}(x) = 1 - F(x)$. Es decir, la distribución PG(σ,γ) es equivalente en la cola a la distribución Pareto(γ).

Por lo tanto, el estimador de Hill (1975) para γ basado en las k estadísticas extremas superiores (Embrechts et al., 1997) es dado por

$$\widehat{\gamma}_k = -\left(W_{n-k+1} - \frac{1}{k}\sum_{j=1}^{k} W_{n-j+1}\right), \tag{2}$$

donde $W_j = \log X_{(j)}, j = n-k+1, n-k+2, ..., n$ y $X_{(1)} < X_{(2)} < ... < X_{(n)}$ son las estadísticas de orden correspondientes a una muestra aleatoria $X_1, X_2, ..., X_n$ de la distribución PG(σ,γ) .

2.2. Método combinado: caso $\gamma < 0$

Sea $U = \left(\bar{F}(X)\right)^{-\gamma}$, esto es, $U = 1 + \dfrac{\gamma}{\sigma}X$. Note que U tiene distribución $Beta(-1/\gamma, 1)$.

Proponemos el siguiente procedimiento para estimar el parámetro γ.

2.2.1. Etapa 1: Método de Momentos

Sean $X_1, X_2, ..., X_n$ una muestra aleatoria de tamaño n de la distribución PG(σ,γ).

El momento muestral de primer orden de U es $m = \dfrac{1}{n}\sum_{i=1}^{n}\left(1 + \dfrac{\gamma}{\sigma}X_i\right) = 1 + \dfrac{\gamma}{\sigma}\bar{X}$ donde $\bar{X} = \dfrac{1}{n}\sum_{i=1}^{n} X_i$.

Por otro lado, el valor esperado de U es $E\{U\} = 1/(1-\gamma)$. Entonces, por el método de momentos, un estimador $\tilde{\gamma}$ debe satisfacer la ecuación: $\dfrac{1}{1-\gamma} = 1 + \frac{\gamma}{\sigma}\bar{X}$.

Resolviendo para γ se obtiene:

$$\gamma = 1 - \frac{\sigma}{\bar{X}}. \tag{3}$$

2.2.2. Máxima Verosimilitud

De la definición de la distribución PG(σ, γ), se tiene que $0 < x < \dfrac{\sigma}{-\gamma}$, cuando $\gamma < 0$. Entonces, el EMV de $\dfrac{\sigma}{-\gamma}$ es $X_{(n)} = \text{máx}\{X_1, X_2, ..., X_n\}$.

Un estimador $\hat{\sigma}$ de σ es dado por:

$$\hat{\sigma} = -\gamma X_{(n)}. \tag{4}$$

Por lo tanto, sustituyendo $\hat{\sigma}$ de (4) por σ en (3) se tiene:

$$\tilde{\gamma} = \frac{\bar{X}}{\bar{X} - X_{(n)}}. \tag{5}$$

3. Prueba de bondad de ajuste

Para una muestra aleatoria $X_1, X_2, ..., X_n$ de una distribución F con soporte en los reales positivos se desea una prueba de bondad de ajuste para $H_0 : F$ *es una distribución* $PG(\sigma, \gamma)$, con σ y γ desconocidos.

Con base en el parámetro de forma γ, se definen dos subclases de distribuciones PG:

$$A^+ = \{\text{todas las distribuciones PG con parámetro de forma } \gamma \geq 0\}$$

y

$$A^- = \{\text{todas las distribuciones PG con parámetro de forma } \gamma < 0\}.$$

Por lo tanto, la hipótesis H_0 es equivalente a $H_0 : F \in A^+ \cup A^-$.

Entonces bajo estas condiciones se propone una prueba de intersección-unión (Casella y Berger, 1990 p. 357) mediante dos pruebas simultáneas para probar $H_0^+ : F \in A^+$ y $H_0^- : F \in A^-$.

3.1. Prueba para H_0^+ $(\gamma \geq 0)$

Note que de (1)

$$\left\{\overline{F}(x; \sigma, \gamma)\right\}^{-\gamma} = 1 + \frac{\gamma}{\sigma}x, \ \sigma > 0, \ \gamma \in R. \tag{6}$$

Entonces, bajo H_0, se tiene una relación lineal entre $\left(\overline{F}(X)\right)^{-\gamma}$ y X.

Una prueba para H_0^+ puede ser basada en el coeficiente de correlación muestral de X_i y $Y_i = \left(\overline{F}_n(X_i)\right)^{-\hat{\gamma}}$, $i = 1, 2, ..., n$, donde F_n es la función de distribución empírica basada en la muestra aleatoria y $\hat{\gamma}$ es el estimador de Hill dado en (2) con $k = n/5$.

El coeficiente de correlación muestral de $X_1, ..., X_n$ y $Y_1, ..., Y_n$ es $R^+ = \frac{\sum_{j=1}^{n}\left(X_j - \bar{X}\right)\left(Y_j - \bar{Y}\right)}{n\sqrt{S_X^2 S_Y^2}}$, donde $\bar{X}$, S_X^2 y $\bar{Y}$, S_Y^2 son la media y la varianza muestral de $X_1, ..., X_n$ y $Y_1, ..., Y_n$.

Entonces la prueba rechaza H_0^+ si $R^+ < c_\alpha^+$ donde c_α^+ es el cuantil del $100\alpha\,\%$ de la distribución de R^+ bajo H_0^+.

Como la distibución nula de R^+ depende de γ, una prueba de bootstrap paramétrico puede ser usada para obtener el valor crítico c_α^+ como sigue.

1. Calcular $\hat{\gamma}$ en (2) con base en la muestra aleatoria y generar J muestras bootstrap de la distribución PG(σ, γ) con $\sigma = 1$ y $\gamma = \hat{\gamma}$.

2. Calcular el valor de R^+ para cada muestra bootstrap.

3. Sean $R_{(j)}^+$ los valores ordenados R_j^+, $j = 1, ..., J$.

4. Tomar $c_\alpha^+ = R_{(\alpha J)}^+$.

Nótese que se usa $\sigma = 1$ ya que R^+ es una estadística escala-invariante.

3.2. Prueba para H_0^- $(\gamma < 0)$

Con base en la relación (6), una estadística de prueba para H_0^- es el coeficiente de correlación muestral de X_i y $Z_i = \left(\overline{F}_n(X_i)\right)^{-\tilde{\gamma}}$, $i = 1, 2, ..., n$, donde $\tilde{\gamma}$ es el estimador combinado dado en (5).

Sea $|R^-|$ el valor absoluto del coeficiente de correlación muestral de X_i y Z_i, $i = 1, ..., n$.

Por lo tanto, se rechaza H_0^- si $|R^-| < c_\alpha^-$ donde c_α^- es el cuantil del $100\alpha\,\%$ de la distribución de $|R^-|$ bajo H_0^-.

Para obtener c_α^- usamos bootstrap paramétrico, similarmente al caso de c_α^+.

3.3. Prueba de Intersección-Unión

Esta prueba para H_0 : F *es una distribución* $PG(\sigma, \gamma)$ *con* $\sigma > 0$ y $\gamma \in R$, consiste en rechazar si ambas pruebas R^+ y $|R^-|$ rechazan. Para que la prueba sea de nivel α se requiere

que cada una de las pruebas R^+ y $|R^-|$ sean de tamaño α, ya que la región de rechazo de la prueba de intersección-unión consiste en la intersección de las regiones de rechazo de las pruebas R^+ y $|R^-|$ (ver Casella y Berger, 1990, p. 378).

Los Cuadros 1 y 2 presentan los resultados de un estudio de simulación para determinar el tamaño y la potencia de la prueba propuesta.

	γ												
n	-10	-5	-4	-3	-2	-1	0	1	2	3	4	5	10
50	0.06	0.08	0.07	0.09	0.09	0.07	0.03	0.07	0.11	0.12	0.13	0.10	0.09
100	0.07	0.09	0.09	0.10	0.10	0.07	0.04	0.08	0.10	0.10	0.10	0.10	0.08

Tabla 1: Tamaño de la prueba intersección-unión, $\alpha = 0.1$

Alternativa	$n = 50$	$n = 100$	**Alternativa**	$n = 50$	$n = 100$
Beta(1,2)	6	10	Gen-Gama(1,1/2)	36	56
Beta(2,1)	53	77	Abs(norm(2,2))	10	16
Beta(5,5)	94	100	Abs(norm(2,1))	64	92
Weibull(2,1)	35	56	Abs(norm(3,1))	93	100
Weibull(3,1)	80	98	Chisq(6)	19	23
Gama(5,1)	45	53	Chisq(15)	68	72
Gama(8,1)	68	73	Abs(Gumbel(3,2))	39	72
Gen-Gama(2,1/3)	97	100	Abs(Gumbel(5,2))	79	98
Gen-Gama(2,1/2)	79	93	Abs(Gumbel(5,5))	99	100

Tabla 2: Potencia de la prueba intersección-unión, $\alpha = 0.1$

4. Aplicación

Osterman (1993) (Reiss y Thomas, 2001) estudió un conjunto de datos que contiene 135 registros en horas por semana de televidentes. El Cuadro 3 presenta los registros que exceden las 20 horas.

Al aplicar la prueba propuesta considerando un nivel de significancia del 10 % se obtiene: $R^+ = 0.9628$, $c^+_{.1} = 0.9277$, $|R^-| = 0.9043$ y $c^-_{.1} = 0.9845$. Por lo tanto, la prueba propuesta

20.00	20.00	20.00	20.50	20.50	22.00	22.00	22.00	23.00	23.00	23.00	23.90
24.00	24.75	25.00	25.00	26.00	26.00	27.00	27.00	27.50	27.50	28.00	28.00
28.50	29.00	29.50	30.00	31.50	33.00	37.00	40.00	45.00	49.00	63.00	

Tabla 3: Horas de TV

no rechaza la hipótesis nula, ya que los datos no presentan evidencia en contra de H_0^+. Debido a que para estos datos H_0^+ no se rechaza, usando (2) una estimación para γ es $\hat{\gamma}_7 = 0.5839$.

Referencias

Casella, G. y Berger, J. 1990. *Statistical Inference*. Brooks/Cole, USA.

Embrechts, P. et al. 1997. *Modelling extremal events*. Springer.

Hill, B.M. 1975. A simple general approach to inference about the tail of a distribution. *Ann. Statist.* **3**, 1163-1174.

Reiss, R.D. y Thomas, M. 2001. *Statistical Analysis of Extreme Values with Applications to Insurance, Finance, Hydrology and Other Fields*. 2ª Ed. Birkhäuser.

Memoria FONE XXIII (2009), Part II, 193–201

Fractional factorial designs: categorical variable applications

Alexander von Eye[a]
Michigan State University

Patrick Mair
WU Wirtschaftsuniversität Wien

1. Introduction

In virtually all experimental as well as non-experimental studies in the social and behavioral sciences, the design factors are completely crossed. This tradition has two important implications. First, the number of factors that can be studied simultaneously is limited, for financial and sample size reasons. Second, the interpretation of higher order interactions is almost impossible. Therefore, effort is wasted in the collection of information that is not used or cannot be used. To give an example (von Eye 2008), consider the cross-classification of six dichotomous factors. The analysis of this design comes with 1 *df* for the intercept, 6 *df* for the main effects, 15 *df* for the two-way interactions, 20 *df* for the three-way interactions, 15 *df* for the four-way interactions, 6 *df* for the five-way interactions, and 1 *df* for the six-way interaction. Now, suppose that only the intercept, the main effects, and the first order interactions are of importance. In this case, two thirds of the degrees of freedom in this design are used to estimate parameters that are not of interest and will not be interpreted.

Based on the Sparsity of Effects Principle (Hamada and Wu 1992, Kutner et al. 2004, Wu and Hamada 2000), it can be predicted that most systems are driven largely by a limited number of main effects and lower-order interactions. Higher-order interactions are, therefore, usually relatively unimportant.

To accommodate the need for larger numbers of factors, and considering the sparsity of effects principle, optimal designs have been discussed (Pukelsheim 2006, Ledolter and

[a]voneye@msu.edu

Swersey 2007). Optimal designs are more parsimonious than fully crossed designs, yet they allow one to estimate the parameters of interest optimally and without confound. Fractional factorial designs are special cases of optimal designs. In this contribution, we discuss fractional factorial designs for studies that collect and analyze categorical dependent variables. We focus on Box-Hunter designs and dichotomous variables.

2. Box-Hunter designs

A subtype of fractional factorial designs is known as Box-Hunter designs (Box, Hunter, and Hunter 2005). These designs use only a fraction of the completely crossed design, for example, 1/2, 1/4, or an even smaller portion of the total number of possible cells (also called runs). The number of factor levels in Box-Hunter designs is 2, for each factor, and the number of runs is a power of 2. In the present context, the dependent variable has also 2 levels.

The design matrix for fractional factorial designs can be created using computer programs. Therefore, these designs are also called computer-aided designs. For the following discussion, consider p variables (consisting of, e.g., $p-1$ factors and 1 dependent variable) and $2p-k$ runs, that is a design that is not fully crossed. In this design, $p-k$ is the number of variables whose main effects can be coded as usual, in a completely crossed design. The main effects of the remaining k variables and their interactions have to be coded differently, because, in fractional factorial designs, the number of rows in the design matrix, that is, the number of runs, is reduced by at least 50% when compared to a completely crossed design. An algorithm for the creation of Box-Hunter designs involves the following steps (von Eye 2008):

1. For the first $p-k$ variables, create a design matrix in which the main effects are coded as in a completely crossed design with $2p-k$ cells (= rows in the design matrix; runs).

2. For Variable $p-k+i$, create the coding vector for the main effect as if it were the interaction among the variables in the first of the $\binom{p-k}{p-k-1} = \binom{p-k}{1}$ combinations of the first $p-k$ variables. In other words, the remaining k main effects are expressed in terms of the $(p-k)$-way interactions of those variables that can be coded as in $(p-k)$-factorial design. Thus, confounds will exist at least at the level of the $(p-k)$-way

interactions.

3. Repeat Step 2 a total of k times, until main effects are created for all p variables.

4. Generate two-way interactions as in a standard ANOVA design, that is, by element-wise multiplication of vector elements from two different variables.

5. Generate three-way interactions also as in a standard ANOVA design, that is, by element-wise multiplication of vector elements from three different variables.

6. Repeat generating interactions until either the design is saturated or all non-confounded and important interactions are included in the design matrix.

The number of designs that can be created this way is $p!/(p-k)!$. This number results from selecting different variables that are coded as in a completely crossed design with $p-k$ cells, and changing their order. This process is also called *randomizing the runs*. If runs can be randomized for fractional designs, alternative designs exist that allow one to estimate the same interactions.

3. The resolution of fractional designs

The resolution of fractional designs is defined as the degree to which main effects and interactions can be independently estimated and interpreted. In different words, the resolution of a design indicates the order of effects that can be estimated such that they are not confounded with each other.

There is a fundamental difference between the resolution in the context of the General Linear Model (GLM) and the resolution in the context of the General Log-Linear Model (GLLM). In the GLM, a main effect relates one independent and on dependent variable to each other. That is, a main effect already involves two variables. Accordingly, a GLM first-order interaction involves three variables, and so forth. In contrast, a main effect in the GLLM involves just one variable. The other "variable" is given by the frequencies in a cross-classification. A GLLM first-order interaction, therefore, involves two variables, and so forth.

From this difference, it follows that, in the GLLM, effects of one level higher than in the GLM are needed to explain main effects and interactions, and the resolution is shifted one

Resolution	**GLM**	**GLLM**
I	No effect can be independently estimated	-
II	Main effects confounded with each other	No effect can be independently estimated
III	Main effects confounded with 2-way interactions	Main effects confounded with each other
IV	Main effects not confounded; two-way effects confounded with each other	Main effects confounded with 2-way interactions
V	Main effects and 2-way effects not confounded; 3-way effects confounded	Main effects not confounded; two-way effects confounded with each other
VI	Up to 3-way effects not confounded	Main effects and 2-way effects not confounded; 3-way effects confounded

Table 1: Design resolution in GLM and GLLM contexts

step up also. Table 1 displays resolutions up to Level VI, and their interpretation in GLM and GLLM contexts.

In the study of categorical variables, main effects are rarely the only effects of interest. Researchers typically focus on interactions of various orders. In the context of GLLM applications, such studies require at least Resolution Level VI. Interestingly, in a large portion of applications, focusing on main effects and first order interactions does not constitute a major restriction. Just consider such methods as factor analysis, correspondence analysis, cluster analysis, or multidimensional. Each of these methods starts from a variance-covariance matrix. That is, each of these methods focuses exclusively on bivariate relationships. Similarly, most applications of structural equations modeling focus on multiple bivariate relationships.

In GLLM applications, consider the example of logistic regression (von Eye and Bogat 2005). Standard logistic regression models make no assumptions about predictor interac-

tions. Therefore, these models are typically saturated in the predictors, and the standard design is completely crossed (if all predictors are categorical). The effects of interest are typically bivariate predictor - criterion relationships. To examine these relationships, two - way interactions are estimated. Higher order interactions are often considered unimportant. In these cases, a fractional factorial design with Resolution Level VI will do the job, at a savings of 50% of the cells and the corresponding savings in sample size, effort, and financial costs.

4. Data example

In a study on the effects of social welfare on mental health in battered women, von Eye and Bogat (2006) asked whether, longitudinally, depression is linked to social welfare as measured by food stamp and Medicaid reception. Data from 6 observation points are available. In the following analyses, we focus on the data from the third and the following three observation points. For the illustration of the application of fractional factorial designs in the analysis of categorical outcome variables, we use the measures:

- Medicaid received at observation points 3, 4, 5, and 6 (M3, M4, M5, and M6; all scored as 1 = did not receive and 2 = did receive)

- Depression at observation point 6: D6 scored as 1 = below the cutoff for clinical-level depression and 2 = above cutoff; depression was measured using the BDI (Beck, Ward, and Mendelson 1961).

The data were collected in one-year intervals. Crossed, these 5 variables span a contingency table with $2^5 = 32$ cells. For the following analyses, we hypothesize that (1) Medicaid reception predicts itself over time; and (2) at Time 6, Medicaid reception predicts depression.

This model is depicted in Figure 1. It shows that only two-way interactions are needed to test the hypothesized relationships. The corresponding hierarchical log-linear model is

$$\ln m = \lambda + \lambda^{M3,M4} + \lambda^{M4,M5} + \lambda^{M5,M6} + \lambda^{M6,D6}$$

with design matrix

1	$M3$	$M4$	$M5$	$M6$	$M3, M4$	$M4, M5$	$M5, M6$	$M6, D6$
1	1	1	1	1	1	1	1	−1
1	1	1	1	−1	1	1	−1	−1
1	1	1	−1	1	1	−1	−1	1
1	1	1	−1	−1	1	−1	1	1
1	1	−1	1	1	−1	−1	1	1
1	1	−1	1	−1	−1	−1	−1	1
1	1	−1	−1	1	−1	1	−1	−1
1	1	−1	−1	−1	−1	1	1	−1
1	1	1	1	1	−1	1	1	1
1	−1	1	1	−1	−1	1	−1	1
1	−1	1	−1	1	−1	−1	−1	−1
1	−1	1	−1	−1	−1	−1	1	−1
1	−1	−1	1	1	1	−1	1	−1
1	−1	−1	1	−1	1	−1	−1	−1
1	−1	−1	−1	1	1	1	−1	1
1	−1	−1	−1	−1	1	1	1	1

This model can be enriched by also testing whether Medicaid reception at observation points 3, 4, and 5 are predictive of depression at Time 6. The enriched model is

$$\ln m = \lambda + \lambda^{M3,M4} + \lambda^{M4,M5} + \lambda^{M5,M6} + \lambda^{M6,D6} + \lambda^{M3,D6} + \lambda^{M4,D6} + \lambda^{M5,D6}.$$

The interactions tested in the enriched model will also be first order.

The fully crossed table of these six variables allows for the study of interactions of up to fifth order. As the three-way, four-way, and five-way interactions are not part of the model,

Figure 1: Medicaid and depression

Model	$LR-X^2; df; p$	$\Delta_1 LR-X^2; df; p$	$\Delta_2 LR-X^2; df; p$
1	218.16; 10; < 0.01		
2	2.37; 6; 0.88	215.79; 4; < 0.01	
3	0.08; 3; 0.99	218.08; 7; < 0.01	2.29; 3; 0.51

Table 2: Estimation and Comparison of Three Models of the Effects of Medicaid Reception on Depression

a fractional, Box-Hunter design with Resolution VI will be sufficient. This design requires only 16 of the 32 cells of the completely crossed table.

We estimate three models. The first is the main effect model, Model 1. It is used as a reference. The second model, Model 2, is the one depicted in Figure 1, and the third, Model 3, is the enriched model. Table 2 shows the overall goodness-of-fit $LR-X^2$ values for these models, and the results of the model comparisons.

The results in Table 2 suggest that the main effect model (Model 1) does a poor job describing the data. In contrast, the model depicted in Figure 1 (Model 2) is not only significantly better than Model 1, it also describes the data very well. In fact, its $LR-X^2$ is so small that it is impossible to improve this model significantly (see the comparison with Model 3). We thus retain Model 2. In Model 2, all interaction parameters are significant. In Model 3, none of the additional parameters is significant. We conclude that depression at Time 6 cannot be predicted from Medicaid reception at the earlier years.

This conclusion is based on a fractional factorial design. In the present example, we are able to compare the results from the fractional design with those from the fully crossed design because we possess the complete data matrix (von Eye 2008). We perform the same analyses as in Table 3, but based on the complete cross-classification. Table 3 shows the overall goodness-of-fit $LR-X^2$ and the results of the model comparisons.

The results in Table 3 are similar to the ones in Table 3. The main effect model does not describe the data well. The model depicted in Figure 1 is a significant improvement, and Model 3 does not improve Model 2 significantly. There is one notable difference between the solutions. Model 2 is not as close to the data when the complete cross-classification is used as when the fractional design is used. However, as before, all interaction parameters in

Model	$LR - X^2; df; p$	$\Delta_1 LR - X^2; df; p$	$\Delta_2 LR - X^2; df; p$
1	292.16; 26; < 0.01		
2	34.81; 22; 0.04	257.35; 4; < 0.01	
3	29.96; 19; 0.05	262.20; 7; < 0.01	4.85; 3; 0.43

Table 3: Estimation and Comparison of Three Models of the Effects of Medicaid Reception on Depression; Results Based on Completely Crossed Factors

Model 2 are significant. In addition, the parameters for the associations between Medicaid Reception during the earlier years and depression at Time 6 are, again, not significant in Model 3. We thus conclude that the fractional design reflects the data structure well. There are no significant interactions higher than first order.

5. Discussion

It was the goal of this contribution to show that fractional factorial designs can be applied without significant loss of information when outcome variables are categorical. It was shown, that the application of fractional factorial designs in the analysis of categorical outcome variables mirrors the application in the analysis of metric outcome variables only in part. Specifically, when categorical outcome variables are analyzed, resolution has to be selected one level higher than for metric outcome variables. The reason for this difference is that main effects in the context of the GLM already describe the relationships between one independent and one outcome variable. In contrast, two-way interactions are needed to do the same when the outcome variable is categorical. Therefore, fractional designs can be more parsimonious when metric outcome variables than when categorical outcome variables are analyzed.

Three implications of this methodology stand out. First, designs and data collection become far more parsimonious than when routinely completely crossed designs are used. Second, data analysis and interpretation of results are simplified because only those interactions are discussed that are of interest based on theory and prior results. Third, analyses can be conducted within the context of methods of analysis from the Generalized Linear Model. Thus, methods of ANOVA, the GLLM, and CFA can be used without significant adaptation. The popular general purpose statistical software packages, for example, R, SAS,

SYSTAT, or Minitab can be used for data analysis. In addition, some of these packages, for example, SYSTAT, Statistica or Minitab offer modules that allow the researcher to specify the design.

References

Beck, A. T., C. Ward, and M. Mendelson. 1961. "Beck Depression Inventory (BDI)". *Archives of General Psychiatry* 4:561–571.

Box, G. E. P., J. S. Hunter, and W. G. Hunter. 2005. *Statistics for experimenters: Design innovation, and discovery*. 2nd. Hoboken, NJ: Wiley.

Hamada, M., and C. F. J. Wu. 1992. "Analysis of designed experiments with complex aliasing". *Journal of Quality Technology* 24:130–137.

Kutner, M. H., C. J. Nachtsheim, J. Neter, and W. Li. 2004. *Applied linear statistical models*. 4th. Boston, MA: McGraw-Hill.

Ledolter, J., and A. J. Swersey. 2007. *Testing 1-2-3: Experimental design with applications in marketing and service operations*. Standford, CA: Stanford University Press.

Pukelsheim, F. 2006. *Optimal design for experiments*. New York: Wiley.

von Eye, A. 2008. "Fractional factorial designs in the analysis of categorical data". *InterStat* 3:1–43.

von Eye, A., and G. A. Bogat. 2006. "Mental health in women experiencing intimate partner violence as the efficiency goal of social welfare functions". *International Journal of Social Welfare* 15:31–40.

von Eye, A., P. Mair, and G. A. Bogat. 2005. "Prediction models for configural frequency analysis". *Psychology Science* 47:342–355.

Wu, C. F. J., and M. Hamada. 2000. *Experiments: Planning, analysis and parameter design optimization*. New York: Wiley.

Lista de autores

Lista de árbitros

El Comité Editorial de la Memoria del XXIII Foro Nacional de Estadística agradece la valiosa colaboración de los siguientes árbitros:

1. Aguirre Torres, Victor M. A. *ITAM*
2. Barrios Zamudio, Ernesto J. *ITAM*
3. Christen Gracia, J. Andrés . *CIMAT*
4. Contreras Cristán, Alberto. *IIMAS – UNAM*
5. Cuevas Covarrubias, Carlos *Universidad Anáhuac Norte*
6. Díaz Ávalos, Carlos. *IIMAS – UNAM*
7. Fuentes García, Ruth S. *Facultad de Ciencias – UNAM*
8. González Barrios, José María. *IIMAS – UNAM*
9. González Farias, Graciela *CIMAT*
10. González Pérez, Luis F. *ITAM*
11. Gracia-Medrano Valdelamar, Leticia. *IIMAS – UNAM*
12. Gutiérrez Peña, Eduardo. *IIMAS – UNAM*
13. Hernández Cid, Rubén. *ITAM*
14. Mena Chávez, Ramses H. *IIMAS – UNAM*
15. Méndez Méndez, Miguel A. *Universidad Anáhuac Norte*
16. Méndez Ramírez, Ignacio. *IIMAS – UNAM*

17. Mendoza Ramírez, Manuel. *ITAM*
18. Nieto Barajas, Luis E. *ITAM*
19. Núñez Antonio, Gabriel. *ITAM*
20. O'Reilly Togno, Federico. *IIMAS – UNAM*
21. Rojas Nandayapa, Leonardo. *ITAM*
22. Rueda Díaz del Campo, Raúl. *IIMAS – UNAM*
23. Ruiz-Velasco Acosta, Silvia. *IIMAS – UNAM*
24. Sánchez García, Oliva, *Universidad Anáhuac Norte*
25. Trejo Valdivia, Belem. *INSP*

Esta publicación consta de 769 ejemplares y se terminó
de imprimir en septiembre de 2009 en los talleres gráficos del
Instituto Nacional de Estadística y Geografía
Av. Héroe de Nacozari Sur Núm. 2301, Puerta 11, Nivel Acceso
Fracc. Jardines del Parque, CP 20276
Aguascalientes, Ags.
México