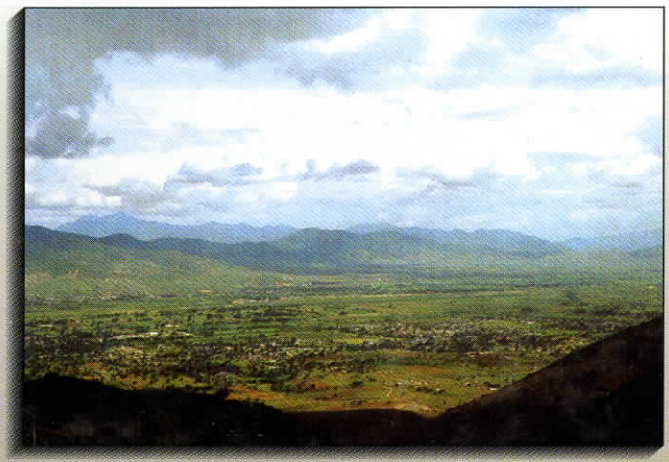
Memoria del X Foro Nacional de Estadística y II Congreso Iberoamericano de Estadística



Oaxaca, México. Septiembre 25-28, 1995

INEGI
INSTITUTO NACIONAL DE ESTADISTICA GEOGRAFIA E INFORMATICA

AME

Memoria del X Foro Nacional de Estadística
y
II Congreso Iberoamericano de Estadística

Oaxaca, México. Septiembre 25-28, 1995

Resúmenes *in extenso*

Editado por:

Ma. de Lourdes de la Fuente D. - ITAM
Eduardo Gutiérrez Peña - IIMAS, UNAM
Jorge Olguín - IIMAS, UNAM

Memoria del X Foro Nacional de Estadística y II Congreso Iberoamericano de Estadística

Impreso en México
ISBN 970-13-1404-2

Presentación

Después de nueve años de celebrarse sin interrupción, en 1995 llegó el turno a la décima versión del *Foro Nacional de Estadística* que organiza nuestra asociación. En esta ocasión ya de por sí festiva tuvimos el privilegio además de disfrutar, como anfitriones, de la compañía de nuestros colegas de Iberoamérica que acudieron al *II Congreso Iberoamericano de Estadística*. Este congreso se fundió con nuestro *Foro* en una sola reunión académica que no sólo cumplió con el objetivo de congregar a un numeroso grupo de estadísticos de la región sino que refrendó el alto nivel científico de que ya había dejado constancia la reunión inaugural celebrada en Cáceres, España en 1992.

El esfuerzo de planear, organizar y coordinar los *Foros*, que cada año la *Asociación Mexicana de Estadística* delega en algunos de sus miembros, ha logrado que estas reuniones se hayan establecido ya como una tradición que la comunidad aguarda año con año para conocer el avance del trabajo de investigación, aplicación y docencia de la Estadística en nuestro país. Si bien este es un compromiso que la *AME* asume sin reservas, es indispensable reconocer y valorar el inmenso apoyo que, no sólo de muchos individuos, sino también de una variedad de instituciones recibe para que estos eventos sean una realidad. En el caso del *X Foro* y el *II Iberoamericano* la lista es afortunadamente extensa y pretender una enumeración completa seguramente nos conduciría a cometer alguna omisión. Sin embargo, no podemos dejar de mencionar el extraordinario apoyo que fué brindado al Comité Organizador por el Gobierno del Estado de Oaxaca, en cuya hermosa ciudad capital se realizó la reunión. Sin su participación, la reunión no hubiese sido posible. De la misma manera, la generosa aportación que en lo tocante a infraestructura, recursos materiales y soporte en una variedad de gestiones, fué recibida de parte del Instituto Tecnológico Autónomo de México resultó invaluable. Del Insituto de Investigaciones en Matemáticas Aplicadas y en Sistemas de la UNAM, como siempre, se contó con apoyo irrestricto. En el nivel operativo local la colaboración del Instituto Tecnológico de Oaxaca fué simplemente decisivo. Por otra parte, la selección y organización tanto de las conferencias invitadas como de las contribuciones libres que se presentaron en el *Foro* fué posible gracias al trabajo de los colegas que desde distintos puntos de la región iberoamericana formaron parte del Comité de Programa. También debe señalarse el enorme esfuerzo realizado por el Comité Organizador y en especial por el Dr. Manuel Mendoza como Presidente de dicho Comité.

La suma de todos estos, y muchos otros, apoyos y voluntades hizo posible la realización del *II Congreso Iberoamericano de Estadística / X Foro Nacional de Estadística* en la ciudad de Oaxaca en los días 25 a 28 de Septiembre de 1995. Con el objeto de dar cuenta de los trabajos científicos que ahí se presentaron se ha elaborado esta memoria de resúmenes *in extenso* para cuya producción hemos contado con el insustituible soporte del Instituto Nacional de Estadística, Geografía e Informática. Finalmente, agradecemos el dedicado apoyo de Angélica Torres en la transcripción de los resúmenes que aquí se presentan.

El Comité Editorial

M. de L. de la Fuente D.
E. Gutiérrez Peña
J. Olguín Uribe

CONTENIDO

CONFERENCIAS

INVITADAS

Cálculo de Precisión Bajo Estructuras de Covarianza Permutabl

GUIDO E. DEL PINO

Universidad Católica de Chile, Chile

RESUMEN

La varianza de cualquier combinación lineal de las observaciones es, en principio, calculable a partir de los coeficientes de la combinación y de la matriz de covarianza de las observaciones. En la práctica, sin embargo, este enfoque resulta excesivamente laborioso y es, además, poco atractivo para obtener expresiones analíticas. En diseños experimentales y muestrales, así como en los modelos usuales de efectos aleatorios, las covarianzas son invariantes bajo permutaciones de los niveles de un factor, cuando se mantienen fijos los niveles de los demás.

Se propone acá un enfoque alternativo, que descansa en la tabla ANOVA teórica asociada a una estructura de covarianza de este tipo, la que no es sino una traducción estadística de la descomposición de la matriz de covarianza en sus vectores y valores propios.

La varianza buscada es una suma de productos, entre ciertos cuadrados medios esperados y sumas de cuadrados asociadas a un vector artificial de datos, formado por los coeficientes de la combinación lineal original.

1. INTRODUCCIÓN

La comparación de las respuestas medias que se producen en diversas situaciones, es un problema frecuente, que se formaliza definiendo una funcional paramétrica adecuada, construyendo luego un estimador de ésta, y evaluando la varianza de este estimador.

Denotando por $\boldsymbol{Y}$ al vector de observaciones, sólo se consideran acá estimadores lineales de la forma $\boldsymbol{a'Y}$. Si $V = \text{Var } \boldsymbol{Y}$, $\omega = \text{Var } \boldsymbol{a'Y} = \boldsymbol{a'Va}$, que es una forma cuadrática en $\boldsymbol{a}$. En la práctica V no se conoce, pero se supone que pertenece a un conjunto $\mathcal{V}$, que habitualmente admite una representación paramétrica ($\mathbf{V}_\tau$, $\tau \in \boldsymbol{T}$). Luego $\omega = g(\boldsymbol{a}, \tau) = \boldsymbol{a'V}_\tau \boldsymbol{a}$. En el modelo lineal mixto (Harville, 1976, 1977; Searle, Casella y Mc Culloch, 1992), $\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{Z}\gamma$, la media $\mu = \boldsymbol{X\beta}$ yace en el espacio columna M de la matriz X, mientras que $V = Z \text{ Var } \gamma Z'$. En este caso τ contiene las componentes paramétricas que se requieran para determinar Var γ.

Aunque la elección de $\boldsymbol{a}$ depende generalmente del modelo supuesto para las medias, lo que se representa en el modelo lineal mixto por X o M, una vez hecha esta elección $g(\boldsymbol{a}, \tau)$ puede evaluarse sin referencia alguna a las medias. Por esta razón, supondremos en lo sucesivo que $\mu = \mathbf{0}$. Por otra parte, en los casos tratados en este trabajo, el estimador lineal insesgado de varianza mínima coincide con el de cuadrados mínimos, de modo que $\mathcal{V}$ no interviene en la elección del estimador. No se trata acá el importante problema de estimar $g(\boldsymbol{a}, \tau)$.

La evaluación directa la forma cuadrática, resulta excesivamente laboriosa, tanto por la gran cantidad de productos y de sumas involucradas, como por la organización de los cálculos. El propósito de este trabajo es mostrar un procedimiento indirecto para una clase $\mathcal{V}$ que puede caracterizarse por restricciones de igualdad entre las componentes de V, las

que, a su vez, se justifican apelando a ciertos supuestos de intercambiabilidad, como veremos más adelante.

En lo que sigue utilizamos la terminología de factores y niveles, muy popular en la literatura de diseño experimental, pero ella se emplea acá como una representación concreta de conceptos más generales. Lo que importa es que cada componente de $\boldsymbol{Y}$ esté asociada con una combinación específica $z = (z_1, z_2, \ldots, z_q)$, donde z_j se interpreta como el nivel del factor F_j, $j \in I = \{1, 2, \ldots, q\}$, y la escribimos como $Y(z)$.

Supongamos ahora que hay invarianza con respecto a permutaciones de los niveles de los factores, hechas separadamente para cada uno de ellos. Para determinar la covarianza entre dos observaciones basta examinar, para cada factor por separado, si los niveles de ambas observaciones coinciden o no. Esto implica que $r \leq 2^q$, en contraste con la cota general $r \leq n(n + 1)/2$. Los modelos equilibrados de efectos aleatorios satisfacen esta condición de simetría y constituyen un ejemplo fundamental. Siendo ellos un caso especial del modelo lineal mixto, no permiten modelar correlaciones negativas entre las observaciones, y por tanto, no se pueden aplicar ni al muestreo ni al análisis de diseños experimentales desde el punto de vista de la aleatorización. Para incluir estos dos casos, enfatizamos la estructura de covarianza, tal como lo hacen Nelder (1965ab, 1977), Dawid (1977, 1988), Tjur (1984), Speed (1987), y Speed y Bailey (1987). Especial mención corresponde a Nelder, quien visualiza a los efectos aleatorios como una herramienta útil para representar la estructura de covarianza.

El modelo con efectos aleatorios trae consigo una tabla ANOVA, cuyos cuadrados medios esperados se pueden usar para calcular la varianza de combinaciones lineales especiales. En este trabajo se extiende esto a una combinación lineal arbitraria y, más aún, al caso en que el modelo de efectos aleatorios no es válido.

Sea $T = S(z, z')$ el conjunto de todos los j para los cuales F_j comparte un nivel común para ambas observaciones. Para 5 factores, A,B,C,D, y E, el par (Y_{28947}, Y_{38447}) tiene asignado el conjunto $T = \{2,3,5\}$, o usando nombres de factores, el conjunto $\{B,C,E\}$. Se verifica fácilmente que $S(z, z') = S(w, w')$ implica que la distribuciones de $(Y(z), Y(z'))$ y $(Y(t),Y(t'))$ coinciden, y por tanto lo mismo ocurrirá con la función de covarianza K definida sobre $\overline{\Omega}^2$ por $K(z, z') = \mathrm{Cov}(Y(z), Y(z'))$. Para calcular las varianzas de las combinaciones lineales, el supuesto esencial es

$$S(z, z') = S(w, w') \Rightarrow K(z, z') = K(w, w'), \qquad (1)$$

Entonces, $K(z, z')$ está determinado por T, lo cual justifica escribirlo como γ_T. Así $\mathrm{Cov}(Y_{28497}, Y_{38447}) = \gamma_{2,3,5} = \gamma_{BCE}$. Llamamos a (1) *Supuesto Básico de Intercambiabilidad* (BEC), y significa que la covarianza entre dos observaciones cualesquiera puede ser determinada examinando, *por separado para cada factor*, si los niveles correspondientes coinciden o no.

Consideremos, por ejemplo, a dos factores $A(2)$ y $B(3)$, donde entre paréntesis se indica el número de niveles del factor Ordenando las observaciones en la forma $\boldsymbol{Y} = (Y_{11}, Y_{12}, Y_{13}, Y_{21}, Y_{22}, Y_{23})'$, la matriz de covarianza más general que satisface BEC es

$$V = \begin{bmatrix} \gamma_{AB} & \gamma_A & \gamma_A & \gamma_B & \gamma_1 & \gamma_1 \\ \gamma_A & \gamma_{AB} & \gamma_A & \gamma_1 & \gamma_B & \gamma_1 \\ \gamma_A & \gamma_A & \gamma_{AB} & \gamma_1 & \gamma_1 & \gamma_B \\ \gamma_B & \gamma_1 & \gamma_1 & \gamma_{AB} & \gamma_A & \gamma_A \\ \gamma_1 & \gamma_B & \gamma_1 & \gamma_A & \gamma_{AB} & \gamma_A \\ \gamma_1 & \gamma_1 & \gamma_B & \gamma_A & \gamma_A & \gamma_{AB} \end{bmatrix}$$

donde $\gamma_1 = \gamma_\phi$, i.e. la covarianza de dos observaciones sin niveles comunes.

Bajo BEC $\boldsymbol{a'Va}$ es la combinación lineal de los γ_T, donde T recorre los 2^q subconjuntos de I. En el ejemplo anterior, ésta es una combinación lineal de γ_1, γ_A, γ_B, y γ_{AB}. En particular, la varianza de la suma total Y_{++} es

$$12\gamma_1 + 12\gamma_A + 6\gamma_B + 6\gamma_{AB}, \tag{2}$$

donde cada coeficiente es simplemente el número de veces que aparece el término correspondiente en la matriz de covarianza.

Un modelo de efectos aleatorios con matriz de covarianza de este tipo es $Y_{ij} = Z + a_i + b_j + (ab)_{ij}$ con Var $Z = \sigma_1^2$, Var(a_i) = σ_A^2, Var $b_j = \sigma_B^2$, , y Var($(ab)_{ij}$)= σ_{AB}^2. La varianza de la suma total es ahora

$$Var\ Y_{++} = 6\left[\sigma_1^2 + 3\sigma_A^2 + 2\sigma_B^2 + \sigma_{AB}^2\right]. \tag{3}$$

Despejando los términos σ^2 en $\gamma_1 = \sigma_1^2$, $\gamma_A = \sigma_1^2 + \sigma_A^2$, $\gamma_B = \sigma_1^2 + \sigma_B^2$, $\gamma_{AB} = \sigma_1^2 + \sigma_A^2 + \sigma_B^2 + \sigma_{AB}^2$, y substituyendo en (3) lleva a (2), pero esta deducción es sólo válida para modelos de efectos aleatorios. Por otra parte, el término en corchetes se reconoce como el cuadrado medio esperado del término constante. Este ejemplo ilustra el hecho que, para un modelo de efectos aleatorios, Var $\boldsymbol{a'Y}$ puede ser escrita como combinación lineal de, ya sea las covarianzas, las componentes de varianza o los cuadrados medios esperados, habiendo una correspondencia uno a uno entre las tres representaciones.

BEC es, típicamente, una consecuencia de supuestos de intercambiabilidad sobre la distribución conjunta de los $Y(z)$, bajo permutaciones de los niveles de un factor, en los que se mantiene constantes a los niveles de los factores restantes. Cuando los niveles de A y B se representan por las filas y columnas de una tabla, BEC significa que el intercambio de filas o de columnas no altera la distribución conjunta de las observaciones asociadas a esta tabla.

Este trabajo propone un procedimiento conveniente para encontrar expresiones analíticas como ésta, sin necesidad de escribir la matriz V explícitamente, y que son válidas para cualquier modelo que satisface BEC. Tal vez la ventaja principal de concentrarse en los modelos de dispersión es que las mismas respuestas son válidas para modelos estadísticos muy diferentes, siempre que admitan una estructura de covarianza. De esta forma, podemos eludir muchos temas controvertidos sobre la interpretación de los efectos aleatorios y la clase de inferencias que pueden hacerse. Una idea de la controversia se puede obtener, por ejemplo, leyendo a Samuels, Casella, y Mc Cabe (1991).

La aleatorización es, tal vez, la única manera de asegurarse de que se cumplan los supuestos de intercambiabilidad. Los desarrollos teóricos de Nelder (1965ab) y Bailey

(1981, 1991) en diseño experimental se realizan en este marco, así como también muchos trabajos de Fisher.

Ericson (1969ab) estudia distribuciones predictivas en muestreo aleatorio, mientras que Cornfield y Tukey (1956) desarrollan sus bien conocidas reglas para calcular cuadrados medios esperados, considerando a los niveles de un factor como una muestra aleatoria de una población finita o infinita.

La formulación del problema, en términos del modelo lineal mixto estándar, donde cada observación es escrita como una suma de efectos fijos y aleatorios, es una forma poderosa de deducir resultados, aunque no puede aplicarse cuando las correlaciones son negativas, como es el caso con aleatorización o muestreo.

Las ideas de intercambiabilidad son también provechosas en ciertos análisis Bayesianos. Para el modelo lineal normal, contribuciones importantes son Lindley y Smith (1972) y Dawid (1977). La clave es que si la verosimilitud y la distribución a priori son normales multivariadas y satisfacen BEC, lo mismo se cumple para la distribución a posteriori y para la predictiva.

Dada su popularidad, parece atractivo deducir las fórmulas de varianza, usando un modelo (artificial) de efectos aleatorios traducir luego los resultados en el marco de otro modelo que satisface BEC y que admite estructura de covarianza semejante. Esta idea está relacionada con las de Nelder (1965ab, 1977), y Dawid (1977, 1988), pero se entrega acá una justificación más rigurosa. Smith y Murray (1984) discuten también este tema para modelos de clasificación de una y dos vías.

La Sección 2 discute la intercambiabilidad y las estructura de covarianza en mayor detalle. Los resultados principales se entregan en la Sección 3. El Teorema 3.1 entrega un resultado muy general, que es luego aplicado a un caso particular, donde se cumple BEC, para así obtener el Teorema 3.3, que es el resultado principal. La Sección 4 contiene varios ejemplos, que ilustran el uso de estos teoremas. Finalmente, la Sección 5 muestra como se aplica el resultado general a problemas de muestreo.

2. INTERCAMBIABILIDAD Y ESTRUCTURA DE COVARIANZA

Denotemos por 2^I la colección formada por los 2^q subconjuntos de I. Bajo BEC, sea $\tau = \left(\gamma_T, T \in 2^I\right)$ y sea $\Gamma = \{\tau \mid V\tau \text{ es n.n.d.}\}$. El modelo de covarianza maximal, bajo BEC, es $V\tau$, $\tau \in \Gamma$. Para un modelo de efectos aleatorios equilibrado, con todos los factores cruzados, la estructura de covarianza es $V\tau$, $\tau \in \Gamma_0$, donde Γ_0 tiene interior no vacío. Las simplificaciones que permite el conocimiento de la situación física se traducen en la eliminación de términos en el modelo de efectos aleatorios o en la imposición de igualdades entre algunas componentes de τ. En diseño experimental, tanto la eliminación de términos, como las igualdades puede ser deducidas a partir de supuestos sobre la anidación de ciertos factores en otros. Tjur (1984), Speed (1987) y Speed y Bailey (1987) sugieren que las igualdades de covarianza sean generadas implícitamente, a partir de un reticulado $\mathcal{P} \subseteq 2^I$ especificado, i.e., para conjuntos S, T $\in \mathcal{P}$ cualesquiera, su unión y su intersección pertenecen a $\mathcal{P}$. Para poner esto en un lenguaje estadístico más familiar, identificaremos el reticulado con una fórmula simbólica del modelo. Los términos faltantes en esta fórmula indican cuáles igualdades de covarianza deben ser impuestas. La regla es simple:

Identificar el valor γ asociado a un término faltante, con el correspondiente al mayor término de la fórmula que esté contenido en él.

Denotemos por BEC($\mathcal{P}$) al modelo general que satisface BEC y la igualdades de covarianza generadas por $\mathcal{P}$. La fórmula asociada a $\mathcal{P}$, genera inmediatamente un modelo de efectos aleatorios que satisface BEC($\mathcal{P}$), al cual denotamos por REF($\mathcal{P}$), y uno de efectos fijos, al que denotamos por FIX($\mathcal{P}$). Asociado a este último, hay una tabla ANOVA, incluyendo fórmulas para las sumas de cuadrados y los grados de libertad.

Consideremos tres factores $F_1 = A$, $F_2 = B$, $F_3 = C$, y el reticulado $\mathcal{P} = \{\phi, \{1\}, \{2\}, \{1,2\}, \{1,3\}, \{1,2,3\}\}$. Este se convierte en $1+A+B+AB+AC+ABC$, siendo $\gamma_c = \gamma_1$ y $\gamma_{BC} = \gamma_B$ las igualdades generadas. Por otra parte, REF($\mathcal{P}$) es

$$Y_{ijk} = Z + A_i + B_j + (AB)_{ij} + (AC)_{ik} + \varepsilon_{ijk},$$

donde todos los términos son no correlacionados y aquellos del mismo tipo tienen igual varianza, e.g. σ^2_A, σ^2_{AB}. La regla formal

$$\gamma_S = \sum_{T \leq S} \sigma^2_T, \quad S \in \mathcal{P}, \tag{4}$$

donde $T \leq S$ significa $T \subseteq S$ y $T \in \mathcal{P}$, significa que un término γ es la suma de los términos σ^2 de igual o menor orden, e.g. $\gamma_{AC} = \sigma^2_1 + \sigma^2_A + \sigma^2_{AC}$. Para una familia de covarianzas (γ_s, S $\in \mathcal{P}$) dada, Nelder (1977) define recursivamente *los componentes de exceso de covarianza* como la única solución al sistema de ecuaciones lineales (4).

Para todo conjunto $D \subseteq I$ denotamos por D^c al complemento de D y definimos n_D como

$$n_D = \prod_{i \in D} n_i$$

Para REF($\mathcal{P}$) se sabe que cada EMS es un valor acumulado, *desde arriba*, de los productos $n_{T^c}\sigma^2_T$ (ver e.g. Searle et al, (1992; p. 116). En nuestra notación

$$\lambda_S = \sum_{T \geq S} n_{T^c}\sigma^2_T, \tag{5}$$

donde $T \geq S$ significa $T \supseteq S$ y $T \in \mathcal{P}$.

Los cuadrados medios esperados, los cuales dependen del verdadero comportamiento probabilístico de la observaciones, están dados por (5) para REF($\mathcal{P}$). Un hecho clave, que usamos sin demostración, es que estas expresiones son válidas para cualquier modelo BEC($\mathcal{P}$).

En un diseño experimental, la información sobre el experimento es, a menudo, expresada en términos de relaciones de anidamiento entre los factores. Hay diferentes maneras de interpretar la condición de que un factor sea anidado dentro de otros factores, pero, a veces, la más simple es que ella representa una manera concisa de resumir una familia de igualdades de covarianza. De hecho, dada la información sobre el cruce y el anidamiento, información, hay un procedimiento automático para generar la fórmula del modelo: *eliminar todo término que incluya un factor, pero que no sea alguno de los factores que lo anidan.* Por ejemplo, el modelo analizado más arriba es inmediatamente deducido del supuesto de que A y B son cruzados y C está anidado en A.

3. TEOREMAS

Suponemos aquí que Y satisface un modelo BEC($\mathcal{P}$). Por falta de espacio se omitirá la demostración de los teoremas.

Teorema 3.1 Sea $(E_1, ..., E_r)$ una descomposición ortogonal de R^n. Sea $SS_j(t)$ la j-ésima suma de cuadrados en la tabla ANOVA asociada con esta descomposición y un vector de datos arbitrario t. Supongamos que la matriz de covarianza V de $\mathbf{Y}$ tiene subespacios propios E_j, $j = 1, ..., r$ y denotemos por α_j los valores propios correspondientes. Entonces:

$$\text{Var } (\boldsymbol{a}'\mathbf{Y}) = \sum_{j=1}^{r} SS_j(a)\alpha_j$$

El valor propio α_j puede ser identificado con el j-ésimo cuadrado medio esperado, digamos EMS_j, asociado al vector de datos centrados $(\mathbf{Y} - \boldsymbol{\mu})$.

$$\text{Var } (\boldsymbol{a}'\mathbf{Y}) = \sum_{j=1}^{r} SS_j(a)EMS_j$$

Corolario 3.1 Bajo los supuestos del Teorema 3.1

$$\text{Var } (\boldsymbol{a}'\mathbf{Y}) = \|a\|^2 EMS_j, \quad \text{si } \boldsymbol{a} \in E_j,$$

Teorema 3.2 Todas las matrices de covarianza V que satisfacen BEC ($\mathcal{P}$) admiten una familia común de subespacios propios (E_S, S $\in \mathcal{P}$). Más aún, los subespacios son idénticos con aquellos que surgen en la descomposición ANOVA de $FEF(\mathcal{P})$.

Teorema 3.3 Si la distribución de $\mathbf{Y}$ satisface BEC($\mathcal{P}$), entonces, Var $(\boldsymbol{a}'\ \mathbf{Y})$ puede calcularse como sigue:

(a) Calcule las EMS en la tabla ANOVA para REF($\mathcal{P}$). Denótelos como (λ_T, $T \in \mathcal{P}$).

(b) Calcule la suma de cuadrados en la tabla ANOVA para FEF($\mathcal{P}$), *usando como vector de datos a los coeficientes de la combinación lineal.* Denótelos por ($SS_T(ba)$, $T \in \mathcal{P}$).

(c) Calcule Var $(\boldsymbol{a}'\mathbf{Y})$ multiplicando los resultados correspondientes en (a) y (b) y sumando todos los productos, i.e.

$$\text{Var } (\boldsymbol{a}'\mathbf{Y}) = \sum_{T \in P} SS_T(a)\lambda_T$$

Corolario 3.2 Bajo los supuestos del Teorema 3.3

$$\text{Var } (\boldsymbol{a}'\mathbf{Y}) = \|a\|^2 \lambda_T \quad \text{si } \boldsymbol{a} \in E_T,$$

4. ALGUNOS EJEMPLOS

Una aplicación directa del Corolario 3.2 prueba (3). Acá $\boldsymbol{a}$ = (1,1,1,1,1,1)' pertenece al subespacio E_1 asociado con el término constante. El coeficiente 6 es la longitud al cuadrado de este vector y el término en corchetes es la EMS para la constante, dada por (2).

El caso más simple $\mathcal{P} = \{I\} \leftrightarrow A$, corresponde a todas las observaciones no correlacionadas y con la misma varianza, siendo una condición suficiente Y_i, $i = 1, ..., n$, i.i.d

con media 0 y varianza σ^2. Otra aplicación del Corolario 3.2, con un único subespacio propio R^n y EMS = σ^2, da Var $(\boldsymbol{a'Y})$ = $\sigma^2 \sum a_i^2$. Esto está de acuerdo con el resultado obtenido por cálculo directo.

El siguiente caso, en orden de complejidad, es $\mathcal{P} = \{\phi, I\} \leftrightarrow 1+A$ que corresponde a la hipótesis de simetría compuesta (varianza constante y equicorrelación), siendo una condición suficiente la intercambiabilidad completa de todas las variables. Hay ahora dos subespacios propios, correspondientes a la constante y al error en una tabla ANOVA con sólo el efecto constante. Del Corolario 3.2 Var $(\boldsymbol{a'Y}) = \sigma_A^2 \sum a_i^2$ para todo contraste y

$$\text{Var } \overline{Y} = \frac{I}{n}\left(n\sigma_Z^2 + \sigma_A^2\right)$$

$$= \sigma_Z^2 + \frac{\sigma_A^2}{n}$$

Esto coincide con el resultado obtenido directamente del modelo

$$Y_i = Z + a_i,\ i = 1, \ldots, n,$$

donde ε_i, $i = 1, \ldots, n$, iid $(0, \sigma_A^2)$ e independiente de Z, a través de $\overline{Y} = Z + \overline{a}$.

El reticulado $\mathcal{P} = \{\phi, \{1\}, \{1,2\}\} \leftrightarrow 1+A+AB$ representa la situación de I conglomerados intercambiables y J unidades intercambiables dentro de cada uno. Cada conglomerado se identifica por un nivel de A, y la unidad dentro de él por un nivel de B. Hay tres covarianzas: $\gamma_\phi = \gamma_1$ es la covarianza para dos unidades en conglomerados diferentes; $\gamma_{\{1\}} = \gamma_A$ es la covarianza para dos unidades en el mismo conglomerado; $\gamma_{\{1,2\}} = \gamma_{\{AB\}}$ es la varianza común de todas las observaciones. El Teorema 3.3 da una fórmula para

$$\text{Var} \sum_{i=1}^{I} \sum_{j=1}^{J} d_{ij} Y_{ij},$$

pudiendo aplicarse el Corolario 3.2 en casos particulares. El modelo de efectos aleatorios asociado es $Y_{ij} = Z + a_i + (ab)_{ij}$, $i = 1, \ldots, I, j = 1, \ldots, J$ donde las variables Z, a_i, $(ab)_{ij}$ son no correlacionadas, con Var $(Z) = \sigma_1^2$, Var $(a_i) = \sigma_A^2$, Var $((ab)_{ij}) = \sigma_{AB}^2$, . Entonces,

$$\text{Var } \boldsymbol{a'Y} = SS_1(\boldsymbol{d})\ \text{EMS}_1 + SS_A(\boldsymbol{d})\ \text{EMS}_A + SS_{AB}(\boldsymbol{d})\ \text{EMS}_{AB}.$$

Las sumas de cuadrados en la tabla ANOVA, con d_{ij} como la observaciones, son $SS_1 = R_1$ $SS_A = R_A - R_1$ y $SS_{AB} = R_{AB} - R_A$, donde

$$R_1 = d_{++}^2 / IJ, \qquad \boldsymbol{R}_A = \sum_{i=1}^{I} d_{i+}^2 / J, \qquad \boldsymbol{R}_{AB} = \sum_{i=1}^{I} \sum_{j=1}^{J} d_{ij}^2.$$

Por otra parte,

$$EMS_1 = IJ\sigma_1^2 + J\sigma_A^2 + \sigma_{AB}^2, \qquad EMS_A = J\sigma_A^2 + \sigma_{AB}^2, \qquad EMS_{AB} = \sigma_{AB}^2.$$

El Corolario 3.2 da Var $c\sum_{i=1}^{I}\sum_{j=1}^{J} Y_{ij} = IJc^2 EMS_1$ Un efecto principal para A corresponde a $d_{ij} = \lambda_i / J$, donde $\Sigma\ \lambda_i = 0$. Como $R_1 = 0$, $R_A = R_{AB}$, la única suma de cuadrados no nula es $SS_A = R_A$, y, por tanto,

$$\text{Var} \sum_{i=1}^{I} \lambda_i \bar{Y}_{i\cdot} = \frac{1}{J}\sum_{i=1}^{I} \lambda_i^2 EMS_A = \sum_{i=1}^{I} \lambda_i^2 \left(\sigma_A^2 + \sigma_{AB}^2 / J\right).$$

5. UNA APLICACIÓN A MUESTREO ALEATORIO

Supongamos que los elementos de la población pueden ser indexados por un arreglo $\mathbf{z} = (z_1, \ldots, z_q) \in \Omega$ donde z_i toma valores entre 1 y N_i, para $i = 1,2, \ldots, q$. Existen muchos esquemas para elegir al azar $n = n_1\, n_2 \ldots\ n_k$ elementos de la población de tamaño $N_1\, N_2 \ldots N_k$. Para $q = 2$ podemos imaginar una gran tabla de N_1 filas y N_2 columnas. Identifiquemos las filas y columnas con los factores A y B respectivamente, y consideremos los siguientes tres esquemas:

i) Elegir $n_1\, n_2$ celdas al azar, sin reposición. Esto es muestreo aleatorio simple y las $n_1\, n_2$ variables son intercambiables. El modelo asociado es $1+AB$.

ii) Elegir n_1 filas al azar. Separadamente de cada una de las filas en la muestra, elegir n_2 elementos al azar. Este es un diseño muestral bietápico. Las filas (conglomerados) son intercambiables, como lo son las celdas dentro de las filas en la muestra. Esto es análogo a B anidado en A y el modelo asociado es $1+A+AB$.

iii) Elegir n_1 filas al azar y n_2 columnas al azar. Las $n_1\, n_2$ celdas así determinadas constituyen la muestra. Esto es análogo a B cruzado con A, y el modelo asociado es $1+A+B+AB$.

Ilustramos ahora como aplicar el Teorema 3.3 para el estudio de muestreo aleatorio simple. Esto corresponde al modelo $1+A$, para el cual hay sólo dos covarianzas distintas, γ_A y γ_1. Representando a la variable por una función h, que asigna el número real $h(z)$ a cada z en la población, γ_A = Var $Y(z)$ está dado por:

$$\gamma_A = \text{Var}\ Y(z) = \frac{1}{N}\sum_{z\in\Omega} \left(h(z) - \bar{h}\right)^2, \quad \text{con} \quad \bar{h} = \frac{1}{N}\sum_{z\in\Omega} h(z).$$

Usando $Y(+) = 0$ and Var $Y(+) = N\gamma_A + N(N\text{-}1)\gamma_1$, da

$$\gamma_1 = -\frac{1}{N-1}\gamma_A$$

Por cálculo directo

$$\text{Var}\ \bar{Y} = \frac{1}{n^2}\left(n\gamma_A + n(n-1)\gamma_1\right)$$

$$= \frac{1}{n}\left(\gamma_A - \frac{n-1}{N-1}\gamma_A\right)$$

$$= \frac{1}{n}\frac{N-n}{N-1}\gamma_A.$$

La componente de exceso de covarianza asociada con A es

$$\sigma_A^2 = \frac{1}{N-1}\sum_{z\in\Omega}\left(h(z)-\bar{h}\right)^2,$$

que coincide con la definición usual de varianza para poblaciones finitas. Los cuadrados medios esperados son

$$EMS_1 = n\sigma_I^2 + \sigma_A^2 = \frac{N-n}{N-1}\gamma_A$$

$$EMS_A = \sigma_A^2 = \frac{N}{N-1}\gamma_A$$

Del Teorema 3.3

$$\text{Var}\,(\boldsymbol{a'Y}) = \left[n\bar{a}^2\frac{N-n}{N-1} + \sum(a_i-\bar{a})^2\frac{N}{N-1}\right]\gamma_A$$

Los siguientes casos particulares pueden obtenerse de (5.1) o del Corolario 3.2:

$$\text{Var}\,\bar{Y} = \frac{1}{n}\frac{N-n}{N-1}\gamma_A = \frac{1}{n}\frac{N-n}{N}\sigma_A^2$$

$$\text{Var}\,(\boldsymbol{a'Y}) = \frac{N}{N-1}\gamma_A\sum a_i^2 = \sigma_A^2\sum a_i^2, \text{ si } \sum a_i = 0$$

En este marco, el insesgamiento de la varianza muestral, como estimador de σ_A^2, se deduce del hecho trivial de que el cuadrado medio del error es un estimador insesgado de su valor esperado.

REFERENCIAS

Aigner, M. (1979). *Combinatorial Theory*. New York: Springer Verlag.

Bailey, R.A. (1981). A unified approach to design de experiments. *J. Roy. Statist. Soc.* A **144**, 214--223.

Bailey, R.A. (1991). Strata for randomized experiments. *J. Roy. Statist. Soc. B* **53**, 719-727.

Cornfield, J. y Tukey, J.W. (1956). Average values de mean squares in factorials. Ann. *Math. Statist.* **41**, 528-538.

Dawid, A.P. (1977). Invariant distributions and analysis of variance models. *Biometrika* **64**, 292-297.

Dawid, A.P. (1988). Symmetry models and hypothesis for estructured data layouts. *J.Roy.Statist.Soc.* B **50** 1-34.

Haberman, S.J. (1975). Direct products and linear models for complete factorial tables. *Ann. Statist* **3**, 314-333.

Lindley, D.V. and Smith, A.M.F. (1972). Bayes estimates for the linear modelo (with discussion). *J. Roy. Stat. Soc. Ser. B*, **34**, 1-41.

Nelder, J.A. (1965a). The analysis of randomized experiments with orthogonal block estructure I. *J. Roy. Stat. Soc.* A **283**, 143-162.

Nelder, J.A. (1965b). The analysis of randomized experiments with orthogonal block estructure II. *J. Roy. Stat. Soc. A* **283**, 163-178.

Nelder, J.A. (1977). A reformulation of linear models. *J. Roy Statist. Soc. A.* **140**, 48-63.

Samuels, M.L., Casella, G. and Mc Cabe, G.P. (1991). Interpreting blocks and random factors. *J. Amer. Stat. Assoc.* **86**, 798--821.

Searle, S.R., Casella, G., and Mc Culloch, C.E. (1992). *Variance Components*. New York: John Wiley

Smith, S.P. and Murray, L.W. (1984). An alternative to Eisenhart's Model II and mixed models in the case of negative variance estimates. *J. Amer. Stat. Assoc.* **79**, 145-151.

Speed, T.P. (1987). What is analysis of variance? *Ann. Statist.* **15**, 885-910.

Speed, T.P. and Bailey, R. A. (1987). Factorial dispersion models International *Statistical Review* **55**, 261-277.

Tjur, T. (1984). Analysis of variance models en orthogonal designs. *Int. Stat. Rev.* **52**, 33-81.

Algunas Aplicaciones de la Estimación de Densidades

RICARDO FRAIMAN

Centro de Matemática Univ. de la República, Montevideo, Uruguay

RESUMEN

En este trabajo, consideraremos algunas aplicaciones recientes de la estimación de densidades. Mas precisamente, consideraremos tres aplicaciones: L-estimadores multivariados, estimación del soporte de una distribución y aplicaciones a detección de conglomerados o "cluster analysis".

1. INTRODUCCIÓN

En este trabajo, consideraremos tres problemas que se resolverán utilizando la estimación de densidades.

1.1 Problema No. 1

Sea f una densidad multivariada en $\mathcal{R}^d$ con soporte compacto S. Queremos estimar S a partir de una muestra $X_1,...,X_n$ de vectores aleatorios independientes con distribución f. Dos aplicaciones son:

(a) Detección de un comportamiento anormal de un sistema. Este problema ha sido considerado por Devroye y Wise (1980).

Supongamos que tenemos un estimador S_n del soporte S, basado en una muestra de tamaño n (suficientemente grande). Si ahora tenemos una nueva observación X_{n+1}, un procedimiento rápido y sencillo para detectar posibles cambios en la distribución subyacente que genera los datos (en el sistema) es observar si $X_{n+1} \in S_n$ o no.

(b) Reconocimiento de formas (Pattern Analysis). En el caso que f sea una densidad uniforme en el conjunto S la estimación del soporte nos provee de la forma del conjunto S, y fue analizada por Grenander (1981) como un problema de reconocimiento de formas.

1.2 Problema No. 2

El problema de búsqueda de conglomerados o "cluster analysis" puede resumirse como sigue: dada una muestra $X_1, ..., X_n$ de variables aleatorias independientes en $\mathcal{R}^d$, descomponerla en "k grupos homogeneos". Podemos entonces pensar, por ejemplo que la densidad subyacente f de las variables aleatorias es una mezcla de k densidades desconocidas. Luego, dado $\alpha > 0$, consideremos

$$S_\alpha = \left\{x \in \mathcal{R}^d : f(x) > \alpha\right\} \tag{1}$$

el conjunto que llamaremos la parte α-significativa del soporte de f, que supondremos un conjunto acotado.

Si elegimos α convenientemente, el número de "clusters" o grupos será el número de componentes conexas del conjunto S_α y la "parte central" de cada cluster corresponderá a cada componente conexa de S_α. Un enfoque relacionado a éste a sido propuesto por

Hartigan (1975) y considerado por Wong (1982), Wong y Lane (1983) y por Cuevas, Febrero y Fraiman (1995). En este contexto, podemos resolver el problema de cluster estimando el conjunto S_α.

1.3 Problema No. 3

Dada una muestra $X_1, ..., X_n$ de vectores aleatorios en $\mathcal{R}^d$, con densidad f, definir "L-estimadores multivariados".

En dimensión mayor que uno, el concepto de estadísticos de orden no es claro, y más de una definición ha sido propuesta, como por ejemplo en Tukey (1975), Liu (1990), Oja (1983), Brown (1983) y Fraiman y Meloche (1995), entre otros. Todas ellas intentan ordenar las observaciones de acuerdo a su profundidad en la nube de puntos, definiendo la mediana como la observación "más profunda".

Más precisamente, el problema será, dada una muestra $X_1, ..., X_n$ de una distribución d-variada (con distribución F desconocida) y centro de simetría $\mu \in \mathcal{R}^d$ desconocido, estimar μ mediante un "L-estimador".

Por simetría consideraremos el concepto de simetría angular (uno de los mas débiles) que requiere que los vectores

$$\frac{(X_i - \mu)}{|X_i - \mu|} \qquad \frac{(\mu - X_i)}{|\mu - X_i|}$$

tengan la misma distribución.

Supongamos, para fijar ideas, que la densidad de f fuera elipsoidal. En este caso, una noción de profundidad que resulta adecuada es la introducida en Fraiman y Meloche (1995), en que profundidad en la nube de puntos corresponderá a alta verosimilitud. Esta correspondencia sugiere la siguiente noción de estadísticos de orden multivariados: sea $\hat{f}$ un estimador de la densidad f. Ordenemos las observaciones $X_1, ..., X_n$ de acuerdo a sus rangos entre sus verosimilitudes aproximadas $\hat{f}(X_1),...,\hat{f}(X_n)$, y definamos estadísticos de orden. Luego, por ejemplo, la mediana multivariada quedará definida a través de:

$$\text{mediana}(X_1,...,X_n) = M = X_i \qquad \text{si} \qquad \hat{f}(X_i) \geq \hat{f}(X_j) \qquad \forall j$$

Un L-estimador $\hat{\mu}_n$ se define como un promedio pesado

$$\hat{\mu}_n = \frac{\Sigma_{i=1}^n X_i \, \phi\left(\hat{f}(X_i)\right)}{\Sigma_{j=1}^n \phi\left(\hat{f}(X_j)\right)} \tag{2}$$

donde ϕ es una función de *score* a seleccionar. Bajo condiciones generales este estimador resultará consistente y asintóticamente normal. Si elegimos

$$\phi\left(\hat{f}(X_i)\right) = J\left(R\left(\hat{f}(X_i)\right)/(n+1)\right)$$

donde R denota el correspondiente rango de $\hat{f}(X_i)$ entre $\hat{f}(X_1),...,\hat{f}(X_n)$, tendremos que

$$\hat{\mu}_n = \frac{\Sigma_{i=1}^n X^{(i)} J\left(i/(n+1)\right)}{\Sigma_{j=1}^n J\left(j/(n+1)\right)} = \sum_{i=1}^n c_{n,i} X^{(i)}. \qquad (3)$$

En particular, si $\phi = I_{[\beta,+\infty)}$, donde I_A denota la función indicatriz del conjunto A

$$\hat{\mu}_n = \frac{\Sigma_{i=1}^n X_i \, I_{[\beta,+\infty)}\left(\hat{f}(X_i)\right)}{\Sigma_{j=1}^n I_{[\beta,+\infty)}\left(\hat{f}(X_j)\right)}$$

corresponde a las medias podadas multivariadas. La poda excluye aquellas observaciones en que la verosimilitud aproximada es muy pequeña y el parámetro β se puede elegir resolviendo la ecuación:

$$\frac{1}{n}\sum_{i=1}^n I_{[\beta,+\infty)}\left(\hat{f}(X_i)\right) = 1-\alpha$$

para un porcentaje de poda $0 < \alpha < 1$.

Mas generalmente, sea $D(x, F)$ una profundidad del punto $x \in \mathcal{R}^d$ respecto de la distribución F. Por ejemplo, (a) la profundidad simplicial, definida por Liu (1990)

$$D(x,F) = SD(x,F) = P_F\left(x \in S(X_1,\ldots,X_{d+1})\right)$$

donde X_1, ..., X_{d+1} son variables independientes, con distribución común F, y $S(X_1,\ldots,X_{d+1})$ denota al simplex de vértices X_1, ..., X_{d+1}. Si $d = 1$, $SD(x,F) = 2F(x)\left(1-F(x-)\right)$.

(b) la profundidad definida por Tukey (1975)

$$D(x, F) = TD(x, F) = \inf_H \{F(H), x \in H\},$$

donde H es un semiespacio cerrado que contiene a x. Si $d = 1$ entonces $TD(x, F) = \min\left(F(x),\left(1-F(x-)\right)\right)$.

(c) la profundidad de verosimilitud, definida por Fraiman y Meloche (1995)

$$D(x, F) = LD(x, F) = f_h(x) = K_h * f(x),$$

donde K es un nucleo no negativo y de integral uno, $K_h(x) = h^{-d}K(x/h)$ y h es una ventana fija.

Otros ejemplos de profundidades son "Majority depth" (Singh, 1991), "Mahalanobis depth", basada en la distancia de Mahalanobis, "Spatial depth" o "L1-depth" (Brown, 1993), etc.

$D_n(x) = D(x, F_n)$ denotará la versión empírica de la profundidad correspondiente y un L-estimador asociado estará entonces definido por:

$$\tilde{\mu}_n = \frac{\Sigma_{i=1}^n X_i \phi\left(D_n(X_i)\right)}{\Sigma_{j=1}^n \phi\left(D_n(X_j)\right)},$$

para una función de score ϕ. Ver, por ejemplo, Fraiman, Meloche y Perera (1996).

Volviendo al caso de la profundidad de verosimilitud, los "estadísticos de orden", y por tanto $\hat{\mu}_n$ serán afinmente equivariantes con tal que la verosimilitud aproximada que usemos sea afinmente invariante, o sea si

$$\hat{f}(AX_i + b; AX_1 + b, ..., AX_n + b) = \hat{f}(X_i; X_1, ..., X_n)$$

para todo $b \in \mathcal{R}^d$ y toda matriz no singular A. Luego, si $\hat{\Sigma}_X$ es un estimador equivariante de la matriz de dispersión, entonces la verosimilitud aproximada definida por

$$\hat{f}_h(X_i) = \frac{1}{n}\sum_{j \neq i}^{n} K_h\left(\left(X_i - X_j\right)^t \hat{\Sigma}_X^{-1}\left(X_i - X_j\right)\right) \tag{4}$$

donde K_h: $R \to [0,+\infty)$, es afinmente invariante cualquiera sea el núcleo K_h.

2. NORMALIDAD ASINTÓTICA Y CONSISTENCIA.

Consideremos primero el caso no afinmente equivariante, el estimador

$$\hat{\mu}_n = \frac{\Sigma_{i=1}^n X_i \phi\left(\hat{f}\left(X_i\right)\right)}{\Sigma_{j=1}^n \phi\left(\hat{f}\left(X_j\right)\right)}$$

con K_h: $R \to [0, +\infty)$ un núcleo acotado. No requeriremos ninguna condición a la ventana h ya que la misma será fija; para los resultados asintóticos no será necesario que h tienda a cero.

Teorema 1. Sean $X_1, ..., X_n$ vectores aleatorios independientes idénticamente distribuidos con densidad $f \in \mathcal{R}^d$ y simétricas alrededor de $\mu \in \mathcal{R}^d$, y supongamos que ϕ tiene dos derivadas acotadas. Entonces si ϕ es nula en un entorno del origen,

$$\sqrt{n}\left(\hat{\mu}_n - \mu\right) \to^w N(0, C),$$

donde

$$C = \frac{E\left(\upsilon(X_1)\,\upsilon(X_1)^t\right)}{E^2\left(\phi\left(f_h(X_1)\right)\right)},$$

$$f_h(x) = E\left(K_h\left(x - X_1\right)\right),$$

$$\upsilon(x) = \frac{1}{2}\left(\left(x - \mu\right)\phi\left(f_h(x)\right) + E\left(\left(X_2 - \mu\right)K_h\left(X_2 - \mu\right)\phi'\left(f_h\left(X_2\right)\right)\right)\right)$$

en que X_2 es una variable aleatoria con densidad f, independiente de X_1 y $\to^w$ denota convergencia débil.

Teorema 2. Bajo las condiciones del Teorema 1, si además $\hat{\Sigma}_X \to \Sigma_X$ en probabilidad, el estimador μ_n, afinmente equivariante, con $\hat{f}_h$ definida por (4), tiene la misma distribución asintótica que $\hat{\mu}_n$ (del Teorema 1).

Volvamos ahora al Problema 1 (Estimación del soporte de f). Devroye y Wise (1980) y Grenander (1981) consideran el estimador

$$S_n = \bigcup_{i=1}^{n} B(X_i, \varepsilon_n) \tag{5}$$

donde $B(x, a)$ es la bola cerrada de centro x y radio a, y $\{\varepsilon_n\}$ una sucesión de parámetros de suavizado, y prueban que si $\varepsilon_n \to 0$ suficientemente despacio (tal que $n\varepsilon_n^d \to +\infty$), S_n es un estimador consistente de S respecto de la distancia en medida

$$d_M(T,S) = \lambda\left(\left(T \cap S^c\right) \cup \left(S \cap T^c\right)\right)$$

o sea la medida de la diferencia simétrica, donde λ denota la medida de Lebesgue en $\mathcal{R}^d$.

Un criterio distinto de proximidad entre conjuntos está dado por la distancia de Haussdorf, definida por:

$$d_H(T,S) = \inf\left\{\varepsilon > 0 : T \subset S^{\varepsilon}, \quad S \subset T\varepsilon\right\}$$

donde

$$S^{\varepsilon} = \bigcup_{x \in S} B(x, \varepsilon). \tag{6}$$

La relación entre los estimadores del soporte y los de la función de densidad es fácil de describir. En efecto, la idea más directa sería considerar

$$\widetilde{S}_n = \{f_n > 0\}$$

donde f_n es un estimador de densidad, por ejemplo uno basado en núcleos

$$f_n(x) = \frac{1}{nh^d} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right).$$

Observemos que si $K(u) = I_{B(0,1)}(u)$ (el indicador de la bola unidad) obtenemos el estimador de Grenander.

Para otros núcleos tenemos que: (i) K debe tener soporte compacto (para evitar el estimador $\widetilde{S}_n = \mathcal{R}^d$) (ii) si K tiene soporte compacto, los estimadores definidos por (5) dan en general nuevamente una unión de bolas (si el núcleo es esférico), o bien unión de deformaciones de bolas (el soporte de K dilatado en h).

En Cuevas y Fraiman (1995) se considera una versión modificada de (5), que evita estos problemas al costo de introducir un parámetro adicional. Más concretamente consideremos el estimador de S dado por

$$S_n = \{f_n > \alpha_n\}, \tag{7}$$

donde α_n es una sucesión que converje a cero. El parámetro α_n nos dá más flexibilidad en la forma de S_n, y en particular este estimador tendrá típicamente un borde diferenciable.

3. DISTANCIA EN MEDIDA. CONSISTENCIA Y VELOCIDADES DE CONVERGENCIA

3.1 Resultados Universales (sin restricciones sobre la forma del conjunto S).

Teorema 3. Sea f una densidad en $\mathcal{R}^d$ con soporte compacto S. Dada una sucesión de estimadores de densidad $\{f_n : n \geq 1\}$ consideremos la sucesión de estimadores

$$S_n = \{f_n > \alpha_n\}, \quad \alpha_n \to 0$$

y supongamos que: $(i)\ \lambda(\{x \in S : f(x) = 0\}) = 0, \quad (ii)\ \alpha_n^{-1} \int |f_n - f| \to 0$ *c.s.*, entonces

$$d_M(S_n, S) \to 0 \text{ c.s.}$$

Observación. Un estudio de la velocidad de convergencia para el error L_1 de estimadores multivariados de densidad (basados en núcleos) se puede ver en Holmstron y Klemel (1992). Bajo condiciones sobre f y K, y una ventana del tipo $h_n = O\left(n^{1/(d+4)}\right)$, da una velocidad de convergencia del orden $O\left(n^{-2/(d+4)}\right)$ o sea que

$$n^{\frac{2}{d+4}} \int |f_n - f| \to 0 \text{ c.s.}$$

y por tanto bastará tomar $\alpha_n = O\left(n^{2/(d+4)}\right)$.

El siguiente resultado relaciona la velocidad de convergencia de los estimadores de densidad con los de S.

Teorema 4. Sea f una densidad en $\mathcal{R}^d$ con soporte compacto. Sea f_n una sucesión de estimadores tales que

$$\rho_n \int |f_n - f| \to 0 \text{ en probabilidad}$$

para una sucesión $\rho_n \to +\infty$. Luego, dada una sucesión α_n, el estimador S_n es consistente a S con respecto a la métrica d_M, con una tasa de convergencia

$$\beta_n = \frac{1}{a_n + (\rho_n\ \alpha_n)^{-1}}$$

donde $a_n = \lambda(\{x \in S : f(x) \leq 2\alpha_n\})$.

Observaciones. (1) Una buena estimación de f (correspondiente a una velocidad rápida ρ_n) conduce a una buena estimación de S. (2) α_n debe converger a cero suficientemente despacio, dependiendo de ρ_n. Por otro lado, como $a_n = \lambda(\{x \in S : f(x) \leq 2\alpha_n\})$, a_n convergerá a cero más rápido cuando α_n converge a cero rápido. Este comportamiento es análogo al compromiso entre sesgo y varianza en la estimación de densidades. (3) La sucesión a_n depende directamente de la forma en que f "se acerca al piso". La situación más favorable es cuando $f > a > 0$ en su soporte. En este caso $a_n = 0$.

3.2 Velocidades de convergencia bajo restricciones de "forma".

La idea general es que cuanto "mejor" sea el conjunto S, más rápido puede ser estimado. Daremos algunas definiciones de la complejidad de un conjunto.

Entropia Métrica. (Kolmogorov y Tikhomirov). Dado un conjunto no vacío S, $S \subset \mathcal{R}^d$, definimos para todo $h > 0$ la entropía métrica como

$$R(S,h) = R(h) = \min\left\{ n \,;\, \exists\ z_1,\ldots,z_n \in S, \quad S \subset \bigcup_{i=1}^{n} B(z_i,h) \right\}.$$

Función de Dilatación (blowing-up volume function). (Cuevas y Fraiman). Dado S compacto no vacío, definimos para todo $h > 0$ la función de dilatación:

$$BUV(S;h) = \lambda\left(S^h\right) - \lambda(S).$$

Un conjunto acotado $S \subset \mathcal{R}^d$ se dice *estándar* si existen $\delta > 0$, $\eta > 0$ tales que

$$\lambda\left(S \cap B(x,\varepsilon)\right) \geq \delta\ \lambda\left(B(x,\varepsilon)\right) \qquad \forall x \in S,\ 0 < \varepsilon < \eta.$$

Una caracterización de las velocidades de convergencia obtenibles en términos de la entropía y de la función de dilatación se da en el siguiente teorema. Como consecuencia podemos concluir que los conjuntos "regulares" se pueden estimar a una velocidad $n^{-1/d}$

Teorema 5. Sea $S_n = \{f_n > \alpha_n\}$ donde el núcleo K utilizado para f_n tiene soporte compacto. Entonces, si $f > a > 0$ en S, tenemos que $\forall n \geq n_0$

$$E\left(d_M(S_n,S)\right) \leq \lambda_1\left(\frac{h}{2}\right)^d R(S,h/2)\,\exp\left(-anh^d\right) + BUV(S,h)$$

donde $\lambda_1 = \lambda(B(0,1))$, si $\alpha_n \to 0$ suficientemente rápido. En particular, si $R(S,h) = O\left(h^{-d}\right)$ y $BUV(S,h) = O(h)$,

$$E\left(d_M(S_n,S)\right) = O\left(h + \exp\left(-anh^d\right)\right).$$

Las condiciones $R(S;h) = O\left(h^{-d}\right)$ y $BUV(S,h) = O(h)$ son satisfechas por los poliedros, bolas y uniones finitas de ellos. Más aún, por la desigualdad isoperimétrica (ver, por ejemplo, Bhattacharya y Ranga Rao, 1976), $BUV(S,h) = O(h)$ para toda unión finita de convexos. Como además $R(S,h) = O\left(h^{-d}\right)$ vale para todo conjunto acotado, ambas condiciones se verifican. Resultados respecto de la métrica de Haussdorf se pueden ver, por ejemplo en Cuevas y Fraiman (1995).

REFERENCIAS

Bhattacharya, R.N. and Ranga Rao, R. (1976). *Normal approximations and asymptotic expansions.* Wiley, New York.

Brown, M. (1983). Statistical Uses of the Spatial Median. *J. R. Statist. Soc. B*, **45**, 25-30.

Cuevas, A. and Fraiman, R. (1995). Support estimation: an aplication of density estimates to pattern analysis. *Unpublished manuscript.*

Cuevas, A. , Febrero, M. and Fraiman, R. (1995). Cluster analysis: a further approach based on density estimation. *Unpublished manuscript*

Devroye, L. and Wise, G. (1990). Detection of abnormal behavior via nonparametric estimation of the support. *SIAM J. Appl. Math.* **38**, 480-488.

Fraiman, R. and Meloche, J. (1995). Multivariate L-estimation. *Unpublished manuscript.*

Grenander, U. (1981). *Abstract Inference*. New York, Wiley.

Hartigan, J.A. (1975). *Clustering Algorithms*. New York, Wiley.

Holmstron, L. and Klemel, J. (1992).Asymptotic bounds for the expected L_1 error of a multivariate kernel density estimator. *J. Multiv. Anal.* **42**, 245-266.

Liu, R. (1990). On a Notion of Data Depth Based on Random Simplices. *Ann. Statist.* **18**, 405-414.

Oja, H. (1983). Descriptive Statistics for Multivariate Distributions. *Stat. and Prob. Letters*, **1**, 327-332.

Tukey, J.W. (1975). Mathematics and Picturing Data. *Proceedings of the International Congress of Mathematics*, Vancouver, **2**, 523-531.

Wong, M. A. (1982). A hybrid clustering method for identifying high-density clusters. *J. Amer. Statist. Assoc.* 77, 841-847.

Wong, M.A.and Lane, T. (1983). A k*th* nearest neighbour clustering procedure. *J. R. Statist. Soc. B*, **45**, 362-368.

Using High Frequency Data and Time Series Models to Improve Yield Management

JOSE RAMON CANCELO y ANTONI ESPASA

Univ. de La Coruña, España *Univ. Carlos III España*

1. INTRODUCTION

The implementation of time series models (from now on TSM) for efficient analysis of high frequency data on activity variables is one of the most promissing fields in applied economics. There exists a large set of variables (consumption of electricity, water, gas or petrol, withdrawals of funds from financial institutions, commuters using public transports, traffic levels, production levels, sales, etc.) which are observed weekly, daily or even hourly: time series of thousands of observations, with valuable information on the characteristics of economic phenomena, are available to the analyst; and the problem consists of handling this huge amount of information for efficient decision making.

Forecasting systems development has occurred in response to new capabilities in data accumulation, among other factors (Murdick and Georgoff, 1993). The purpose of this paper is to highlight the usefulness of TSM in analysing high frequency data; we focus on short term forecasting because this is the main concern when dealing with high frequency information, although some other applications are also reviewed. Throughout the paper the expression 'high frequency data' will refer to data observed at least twice per month: the sampling interval may be one week, one day, one hour or any other that meets this condition; and 'short-term forecasting' refers to the specific problem of forecasting this type of data.

The application to yield management and specifically to the optimal sale of perishable products is straightforward: a good model provides an adequate representation of the data generating process, which can be used to reduce the uncertainty of demand forecasts; supply can adjust accordingly, and costs lower.

The paper is organized as follows: general pros and cons of TSM are discussed in section two, and they are compared with some competing procedures; an application to daily electricity consumption is discussed in section three; and the main conclusions are summarized in section four.

2. FORECASTING TECHNIQUES FOR HIGH FREQUENCY DATA

For the purposes of this paper forecasting techniques can be broadly separated into subjective forecasts and model-based methods. Although the former are usually referred to as judgmental forecasting, we think that judgment plays a central role in the forecasting and planning process, no matter the specific forecasting technique involved: as Hogarth and Makridakis (1981) point out, decisions on the specification of goals, choices concerning data sources, forecasting methodologies, adjustments to basic forecasts and the assessment of implementation strategies make the forecasting task, taken in the broad sense, a matter of judgment.

2.1 Subjective Forecasts

In most organizations short-term forecasting is charged to qualified experts that produce realiable forecasts because of their experience (Goodwin and Wright, 1994). Moreover, sometimes non-systematic, heterogeneous information is available: it is hard to introduce this type of information into a quantitative model, but it may be processed and incorporated by an expert into the final forecasts (Brown, 1988).

However, purely subjective forecasting is not the best choice for setting up a forecasting system for high frequency data, because:

1) It is very difficult to transmit to other people the way information is processed to produce forecasts; as a consequence, the whole planning process is highly dependent on the permanence of particular persons in the organization.

2) It is an expensive forecast, as it takes a significant part of the working hours of qualifed personnel. Detailed analysis by an expert is justified in specific moments where complex conditions prevail, buto not in normal days.

3) Socioeconomic conditions change, so that the variables we are interested in react to changes in the explanatory variables in a more sophisticated way; subjective learning becomes more difficult and new types of tools must be considered.

4) Recent research has focused on the inconsistencies of human judgment, even for very qualified experts: see Goodwin and Wright (1993, 1994) and references therein; see also Ashouri (1993).

2.2 Quantitative Methods

Weekly, daily or hourly activity series display the same characteristics than monthly or quarterly series, even though their short term components (seasonal, irregular, calendar, effects, outliers, etc) are much more complex; most components are induced by stable patterns of behavior of the economic agents, so that a model -i.e., an explicit representation of the data generation process (DGP)- may be built and used to get informative forecasts. However, because of this complexity simple smoothing techniques do not provide an adequate approximation to the DGP: true TSM are needed if data are to be processed in an efficient way to produce optimal forecasts.

The main purposes of a TSM are:

1) To generate reliable forecasts with no need of suplementary evaluation by an expert.

2) To become an operative tool within the organization: management support, user involvement, personal stake and the implementation strategy are as relevant to forecast success as accuracy (Schultz, 1992).

3) To produce an adequate anchor in the presence of very complex conditions. Goodwin and Wright (1994) point out that it seems that a process of anchoring and adjustment is used in judgmental extrapolation: although the joint ocurrence of anomalous events may require the forecast of the model being adjusted by an expert, this forecast is still the best starting point for the subjective adjustment.

4) To help the organizations become better learning systems: organizational learning is defined as the capacity within an organization to maintain or improve performance based on experience (Nevis et al, 1995). A TSM is not just a tool for acquisition of knowledge, it has to do with its transfer: knowledge becomes institutionally available, as opposed to being the property of selected individuals.

5) To quantify the influence of explanatory variables with a double purpose: a) better forecasts may be produced if good predictions for explanatory variables are available; and b) simulation exercises may be carried out.

6) To extract a more reliable signal by eliminating from the observed series the effect added noise, in order to use it in the decision making process.

The biggest objection to TSM is the amount of resources needed to build and maintain them. Makridakis et al (1983) consider four elements of cost in a forecasting method: development costs, data storage costs, maintenance costs and the costs of repeated applications. Speaking in relative terms with respect to the total amount, development costs are by far the most important. Maintenance costs are important too, as they include adjusting the model whenever changes in the basic pattern are detected. On the contrary, once a TSM is implemented storage and repeated applications costs are almost negligible.

In any case, to build a model for a typical high frequency series is a hard job, and a detailed analysis has to be carried out to decide which variables will have their own TSM. Although each case deserves specific consideration, a good rule is to determine the monetary loss as a function of the prediction error for all variables; next the loss functions are compared to the actual total cost of a TSM, so that a model is built only when a substantial saving is expected (but this is not so simple as it seems: see for instance Remus (1991) on the consequences of the criterion being a nonlinear function of the variable).

3. AN APPLICATION TO DAILY CONSUMPTION OF ELECTRICITY

3.1 The Problem

Electricity consumption is a typical example of the problem we are considering in this paper: long series of hourly and daily data are available; it is a perishable good, because overproduction (the difference between total production and instantaneous consumption) is wasted, so a very accurate forecast of the demand is needed. Short-term electricity consumption forecasting deserves a remarkable place in the literature of high frequency data analysis: see, inter alia, Bogard et al (1982), Bunn and Farmer (1985), Gross and Galiana (1987), Adams et al (1991) or Engle et al (1992). References on very related problems also provide valuable guidance: see for instance Ashouri (1993) on gas demand.

When we were charged to build a forecasting system for spanish daily consumption, which could also help in setting weekly and hourly production schedules, we approached the job in the following way:

1) To begin with, a several year, homogeneous series of daily data was collected. We had to define consumption in an operative way, in order to separate actual demand from final destination of overproduction Homogeneous time series for the explanatory variables were also prepared.

2) The second step consisted of determining the main characteristics of the resulting series.

3) Next we built a complex nonlinear transfer function model to explain these characteristics.

4) From this model daily forecasts are obtained automatically. Weekly forecasts result from aggregating daily ones; hourly forecasts can be produced by identifying, typical load curves and distributing daily forecasts accordingly.

5) The model is also used to improve our knowledge on the influence of explanatory variables, and to extract a more reliable signal of electricity consumption.

These five stages are related to what Murdick and Georgoff (1993) call the central components of a forecasting system: the input data (point 1), the output we would like (points 4 and 5) the assumptions about the behavior of the variables (point 2 and the extrasample information used in point 3) and the process relating dependent to independent variables (the model that results from point 3).

In the next sections a more detailed description is given. However, in doing so our purpose is just to use this application as an example of the potential use of TSM; as a consequence some relevant results concerning the specific problem of modelling electricity consumption will be ommitted. A complete exposition can be found in Cancelo and Espasa (1991a), available from the authors upon request.

3.2 The Data

The variable to model is the net demand for electrical energy, defined as total production from all sources plus international interchanges balance less intermediate autoconsumption and pumping consumption. The only available data referred to the peninsular part of the spanish territory taken as a whole, almost half a million squared kilometers with more than 36 million inhabitants.

The original model was built for the sample 1983-1989. The series displays: a growing trend; annual and weekly seasonal oscillations; complex calendar effects, related to changes in the usual pattern of working conditions (holidays, vacation periods, Easter); some anomalous values, caused by strikes, elections, and the like. Moreover, weather conditions are known to have a significant influence.

3.3 Overview of The Model

From section 3.2 it follows that observed consumption in day t(C_t) can be expressed as:

$$C_t = TC_t * SC_t * CE_t * IA_t * CMV_t * IC_t$$

where:

TC_t: trend consumption, related to socioeconomic factors;
SC_t: seasonal consumption;
CE_t: calendar effect;
IA_t: intervention analysis to treat anomalous observations which deserve specific consideration;
CMV_t: contribution of meteorological variables;
IC_t: irregular consumption, which captures all transitory disturbances which are not included in previous components.

The model is completely multiplicative, so that all components are assumed to increase in size in direct proportion to the trend level: it seems to be the general rule in activity series, no matter the frequency of observation of the data (Bogard et al, 1982). Taking logarithms

$$\ln C_t = \ln TC_t + \ln SC_t + \ln CE_t + \ln IA_t + \ln CMV_t + \ln IC_t \qquad (1)$$

From (1) a basic consumption (BC_t) can be defined

$$ln\,BC_t = ln\,C_t - ln\,CE_t - ln\,IA_t - ln\,CMV_t = ln\,TC_t + ln\,SC_t + ln\,IC_t \qquad (2)$$

Basic consumption displays a smooth evolution and may be explained in a satisfactory way from its past values:

$$ln\,BC_t = b_1\,ln\,BC_{t-1} + \ldots + b_p\,ln\,BC_{t-p} + residual_t \qquad (3)$$

Calendar effects and intervention analysis can be expressed as:

$$ln\,CE_t + ln\,IA_t = F_1 DV_{1,t} + F_2 DV_{2,t} + \ldots + F_m DV_{m,t} \qquad (4)$$

where $DV_{i,t}$ denotes a dummy variable that indicates whether a specific calendar effect or an anomaly happens in t; F_i denotes its dynamic filter, that simplifies into a single coefficient if $DV_{i,t}$ has no dynamic effect.

As for the contribution of meteorological variables,

$$ln\,CMV_t = G_1 MV_{1,t} + G_2 MV_{2,t} + \ldots + G_n MV_{n,t} \qquad (5)$$

where $MV_{i,t}$ stands for a meteorological variable and G_i for its dynamic filter, which need not be a linear one.

By combining (2) and (5) the general formulation of the model results:

$$\begin{aligned} ln\,C_t = {} & b_1\,ln\,C_{t-1} + \ldots + b_p\,ln\,C_{t-p} + F_1^* DV_{1,t} + F_2^* DV_{2,t} + \ldots + \\ & + F_m^* DV_{m,t} + G_1^* MV_{1,t} + G_2^* MV_{2,t} + \ldots + G_n^* MV_{n,t} + residual_t \end{aligned} \qquad (6)$$

In (6) all observations are handled in a single, general model, which captures all potential changes in electricity consumption due to changes in the explanatory variables. Database management is heavily simplified and forecasts are easy obtained in an automatic way, two major conditions for the system being really useful.

3.4 ON MODELLING THE COMPONENTS

3.4.1 Basic consumption

Trying to model trend and seasonality by including explanatory variables is unfeasible in most cases, because good data observed with the required sampling interval is seldom available. However, their contribution to the present observed value can be approximated quite well by using previous values of electricity consumption, due to the fact that the underlying factors change slowly. Relating trend and seasonal to the past history of the variable allows the resulting estimates to adapt to recent observations (Box et al 1987, Mills 1990), this flexibility being one of the main determinants of the success of modern time series analysis.

3.4.2 Calendar Effects

Although the literature has focused mainly on calendar effects in monthly series, they also exist in higher frequency series (Cleveland and Grupe, 1982). In fact, the smaller the sampling interval the more important the influence of the calendar.

Take for instance a holiday. In most cases it will have a minor influence in monthly data; but in a daily series its presence distorts the whole usual weekly pattern. As a consequence, if its effect is not specifically considered then bad forecasts for the day of the holiday and for the following days will result. The trouble is more serious in latin countries like Spain than in the U.S.: in America a given holiday usually falls on the same day of the week, but in latin countries the general rule is to fix the day of the month, so that it may fall on any of the seven days of the week; and this mobility increases the distorting effect.

A TSM makes possible to analyze in full detail the influence of the calendar: a large sample is carefully screened, stable patterns of behavior for each type of effect are detected and general rules for forecasting are stated.

As an example, table 1 summarizes the estimated effects of holidays on spanish electricity consumption: an estimated coefficient of 30, for instance, means that observed consumption would be 30% higher if the holiday did not exist. From table 1 it can be seen that in our series: a) the distortion varies according to the day of the week on which the holiday falls; and b) there is a dynamic effect, so that a holiday falling on t alters the consumption of two or more days.

Table 1

Estimated Effects of Holidays on Spanish Daily Electricity Consumption

Day of the holiday	Effect on					
	MON	TUE	WED	THU	FRI	SAT
MON	30.9	4.1				
TUE	10.7	35.1	3.2			
WED			30.4	3.8		
THU				29.3	11.6	2.6
FRI					28.8	9.1
SAT						8.2

3.4.3 Meteorological Variables

Among meteorological variables temperature is the most important. Our measure of temperature is a weighted mean of maximum daily temperatures registered in ten selected observatories throughout the whole territory. To model the relationship between this indicator and electricity consumption the following extrasample information must be taken into account:

1) The relationship is U-shaped: there are two bounds of temperature, T* and T**, that define a neutral zone so that temperatures within this interval do not influence consumption. Below T* we enter into the cold zone, and above T** in the hot zone. In both zones the response function is also expected to be nonlinear. We have estimated that in our series T* = 20C (68F) and T** = 24C (75.2F).

2) In daily data a dynamic response is expected, as consumption in day t depends on observed temperatures in t, t-1, ..., t-h.

3) Exhaustion effects may exist: when temperature is so low (high) that every heating (cooling) system is operating at full capacity, then additional decreases (increases) of temperature will have no effect on observed consumption.

4) The influence of a given temperature may be different for a working day than for a non-working day, or vary according to the season of the year, etc.

5) If the sample is several years long, the stock of appliances may increase and shifts in the response function along the sample are to be looked for.

All these effects have been tested and modelled in our application, so that we got a deep knowledge of the relationship between consumption and temperature in our problem.

The effect of other meteorological phenomena of lesser importance (which Ashouri (1993) calls misery factors) are harder to model: homogeneous series for the whole sample are not available, and the forecasts provided by the weather center are not good enough. As a consequence the model does not take them into account, although the experts may adjust the forecasts of the model for their influence in the presence of very extreme conditions.

3.5 ON USING THE MODEL

3.5.1 Forecasting

The final model has a residual standard error equal to 0.0130, which entails a 90% confidence interval for the one period forecast equal to plus/minus 2.13% times the point forecast. It represents a major improvement with respect to previous holistic forecasts. In fact, the actual improvement was greater: the model explains sudden changes in consumption caused by unexpected changes in weather conditions, and large errors are much more uncommon; for this series the extracost caused by a bad forecast is a convex function of the prediction error, so that a remarkable saving is achieved by eliminating big errors.

The model can also be used to obtain provisional forecasts of the consumption with a higher level of time aggregation, and these forecasts may enter as inputs in models explaining other variables. Bodo et al (1991) use daily data on electricity consumption to forecast monthly consumption, and the latter to forecast the monthly industrial production index: while the official figure of the production index for month M is available by the end of M+2, with their proposal a quite reliable forecast can be advanced once the first fortnight electricity consumption of M is known.

3.5.2 Simulation

Given that calendar effects and weather variables are explicitly introduced in the model, the behavior of electricity consumption under different scenarios may be simulated. As an example, assume maximum daily temperature has been constant in 16C (60.8F) during the

whole month of January. On Monday, February 1, it suddenly falls to 11C (51.8F), remaining there for a week; then on Monday, February 8, it returns to 16C and keeps unchanged onwards. Assume also that on Wednesday, February 3, a successful 24-hour general strike takes place.

For the purposes of our analysis we may consider that on January 31 we were in equilibrium; the sudden fall in temperature and the general strike are exogenous transitory distrubances, and on the long run consumption will return to equilibrium. However, it is interesting to know how consumption reacts in the short run: figure 1 displays the estimated effects from Saturday, January 30, to Tuesday, February 16.

Equilibrium is characterized by a temperature in the cold zone, so there is overconsumption with respect to the situation where temperature has no influence. The solid line shows overconsumption in equilibrium expressed as the relative increase with respect to normal consumption with no temperature effects. Notice the peaks at weekends, due to the fact that in our series the effect of a given temperature is higher on non-working days.

The dashed line refers to the proposed scenario: overconsumption (removing the influence of the strike) is higher than in equilibrium, because temperature is lower. The dynamic effect of temperature is easy to see: the distortion caused by the cold wave lasts until February 14, although temperature has returned to equilibrium on February 8. The contribution of the strike is also very clear: in order to estimate its effect we treated it as if it were a holiday falling on Wednesday.

The gap between both lines measures the influence of transitory disturbances, and the total effect of the cold wave plus the strike results from aggregating daily gaps.

3.5.3 Signal Extraction

The previous example has shown that our variable is heavily influenced by disturbances that distort time comparisons, up to the point of making them uniformative. With a TSM this type of effects can be eliminated, so that a more informative signal results: see Cancelo and Espasa (1991b), whose main results are summarized in figure 2.

The dashed line displays the relative change in the observed series with respect to the same month of the previous year. The solid line is computed from a daily series of corrected consumption: we first eliminate from the observed series the effect of every type of disturbance that may distort time comparisons (see the original paper for details); then daily data are aggregated to form a monthly series, and relative growth is computed.

It can be seen in figure 2 that the corrected series of growth displays a smoother evolution, so that most of the peaks and troughs of the observed series of growth are caused by short term disturbances. As a consequence, it seems rather inaccurate to base decision making on original growths.

4. CONCLUSION

In this paper we tried to show the potential contribution of time series models to the analysis of high frequency data of economic activity. Although most people consider them just a forecasting tool, we remark their central role in the acquisition, sharing and utilization of knowledge within an organization. This view agrees with recent developments in management science, which favour organizational memory and a publicly documented body

FIGURE 1.- SIMULATING THE INFLUENCE OF CALENDAR EFFECTS AND TEMPERATURE ON THE CONSUMPTION OF ELECTRICITY: AN EXAMPLE

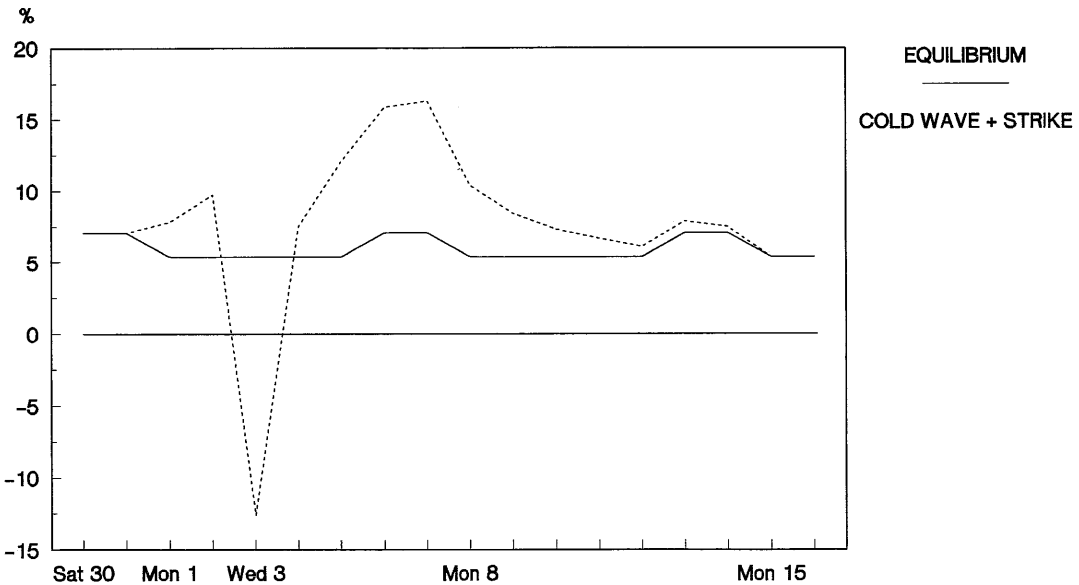
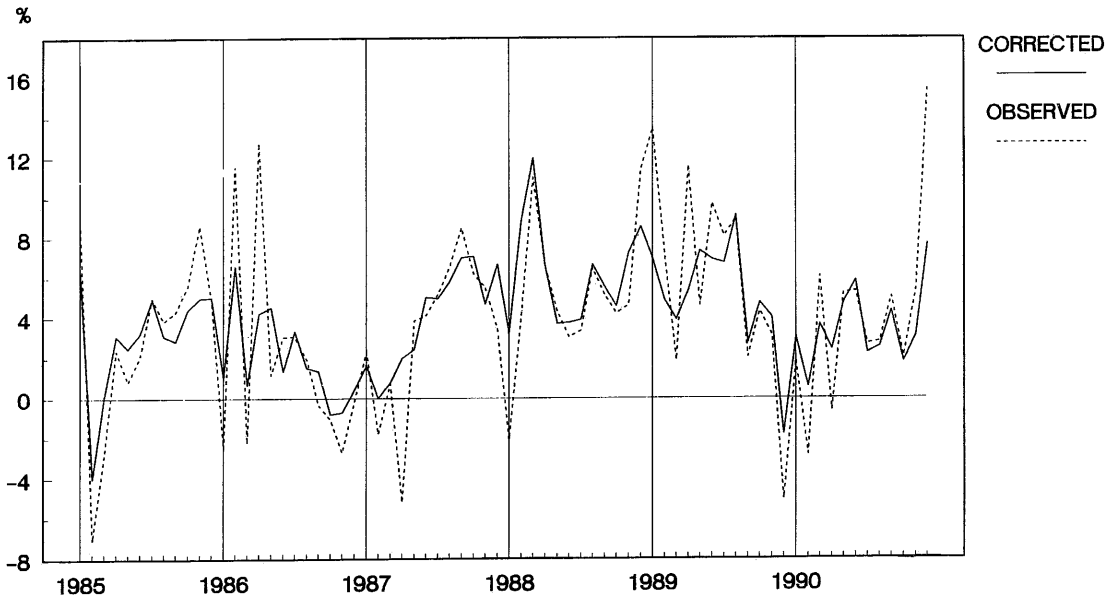


FIGURE 2.- OBSERVED AND CORRECTED GROWTHS OF THE MONTHLY CONSUMPTION OF ELECTRICITY IN SPAIN



NOTE: Growth is computed as relative change with respect to the same month of the previous year

of knowledge as opposed to personal knowledge that is lost when a long-time employee leaves the organization (Nevis et al, 1995).

We have argued that forecasting is not the only application of TSM. There are other by-products that may become as important as direct extrapolation from the observed history. Simulations and signal extraction provide valuable information for the planning process: the former goes one step further in analysing the environment, as new scenarios can be defined and their influence quantified; the latter, because high frequency (daily, weekly), free-from-noise signals may be aggregated into lower frequency signals (monthly, quarterly, yearly), providing managers a much better perception of the underlying trends in the observed data.

REFERENCES

Adams, G., P.G. Allen and B.J. Morzuch (1991). 'Probability distributions of short-term electricity peak load forecasts", *International Journal of Forecasting*, **7**, 283-297.

Ashouri, F. (1993): An expert system for predicting gas demand: a case study, *Omega*, **21**, 307-317.

Bodo, G., A. Cividini and L.F. Signorini (1991). Forecasting the italian industrial production index in real time, *Journal of Forecasting*, **10**, 285-299.

Bogard, C., G. George, G.M. Jenkins and G. McLeod (1982). *Analysing a large number of energy time series for a utility company*, chapter 5 in G.M. Jenkins and G. McLeod (eds., 1982), Case studies in time series analysis, Gwilim Jenkins & Partners Ltd., Lancaster.

Box, G.E.P., D.A. Pierce and P. Newbold (1987). Estimating trend and growth rates in seasonal time series, *Journal of the American Statistical Association*, **82**, 276-282.

Brown, L.D. (1988). Editorial: comparing judgmental to extrapolative forecasts: it's time to ask why and when, *International Journal of Forecasting*, **4**, 171-173.

Bunn, D.W. and E.D. Farmer (eds., 1985). *Comparative models for electrical load forecasting*, John Wiley & Sons, New York.

Cancelo, J.R. and A. Espasa (1991a). Forecasting daily demand for electricity with multiple-input nonlinear transfer function models: a case study, Working **Paper 9121**, *Economics Department, University Carlos III of Madrid.*

Cancelo, J.R. and A. Espasa (1991b). New weekly and monthly indicators of activity based on electricity consumption (in spanish), **Documento de Trabajo 9106**, *Economics Department, University Carlos III of Madrid.*

Cleveland, W.P. and M.R. Grupe (1982). *Modelling time series when calendar effects are present* (with discussion), in A. Zellner (de., 1982*), Applied time series analysis of economic data*, Bureau of the Census, Washington.

Engle, R.F., C. Mustafa and J. Rice (1992). Modelling peak electricity demand, *Journal of Forecasting*, **11**, 241-251.

Goodwin, P. and G. Wright (1993). Improving judgmental series forecasting: a review of the guidance provided by research, *International Journal of Forecasting*, **9**, 147-161.

Goodwin, P. and G. Wright (1994). Heuristics, biases and improvement strategies in judgmental time series forecasting, *Omega*, **22**, 553-568.

Gross, G., and F.D. Galiana (1987). Short-term load forecasting, *Proceedings of the IEEE*, **75**, 1558-1573.

Hogarth, R.M. and S. Makridakis (1981). Forecasting and planning: an evaluation, *Management Science*, **27**, 115-138.

Makridakis, S., S.C. Wheelwright and V.E. McGee (1983). *Forecasting: methods and applications*, John Wiley & Sons, New York.

Mills, T.C. (1990). *Time series techniques for economists*, Cambridge University Press, Cambridge:

Murdick, R.C. and D.M. Georgoff (1993). forecasting: a systems approach, *Technological Forecasting and Social Change*, **44**, 1-16.

Nevis, E.C. , A.J. DiBella and J.M. Gould (1995). Understanding organizations as learning systems, *Sloan Management Review*, **36**, 73-85.

Remus, W.E. (1991). Criterion-referenced judgmental forecasting models, *Journal of Forecasting*, **10**, 415-423.

Schultz, R.L. (1992). Fundamental aspects of forecasting in organizations, *International Journal of Forecasting*, **7**, 409-411.

A General Method for Approximating to the Distribution of Some Statistics

GAUSS M. CORDEIRO y SILVIA L.P. FERRARI

Univ. Fed. de Pernambuco, Brazil *Univ. de Sao Paulo, Brazil*

SUMMARY

The object of this paper is to show that for any statistic satisfying fairly general conditions, we can construct an adjusted statistic having the same distribution of an arbitrary first-order approximating distribution to order $O(n^{-1/2})$ or even $O(n^{-1})$. We prove that the multiplication of the statistic by a suitable stochastic correction improves the first order approximation to its distribution. This paper extends the results of the closely related paper by Cordeiro and Ferrari (1991) to cope with several other statistical tests. The resulting expression for the adjustment factor requires knowledge of the Edgeworth-type expansion to order $O(n^{-1})$ for the distribution of the unmodified statistic. In practice its functional form involves some derivatives of the first-order approximating distribution, certain differences between the cumulants of appropriate order in n of the unmodified statistic and those of its first order approximation, and the unmodified statistic itself. Some applications are discussed.

SOME KEY WORDS: Bartlett correction; Chi-squared distribution; Edgeworth-type expansion; Generalized Bartlett correction; Maximum likelihood estimate; Signed likelihood ratio statistic.

1. INTRODUCTION

In the past 15 years or so there has been a renewed interest in Bartlett corrections leading to better approximations of the null distribution of the likelihood ratio statistic by a chi-squared distribution. Computation of Bartlett corrections has been discussed by Lawley (1956), Barndorff-Nielsen and Cox (1984) and Cordeiro (1993a). General formulae for Bartlett corrections have been obtained explicitly in several regression models by Cordeiro (1983, 1985, 1987), Cordeiro and Paula (1989), Cordeiro, Paula and Botter (1994) and authors cited therein. The numerical benefits of Bartlett corrections have been demonstrated by Møller (1986) and Cordeiro (1993b, 1995) among others. The main goal of this paper is to show that Bartlett's technique can be carried out for general continuous statistics.

Several statistical tests rely in some way on first-order approximations derived from distributions other than chi-squared. We are often interested in computing significance levels or confidence intervals based on these first-order approximations. However, it is also well known that these approximations may not work well for small or moderate-sized samples.

The problem of developing a correction similar to the Bartlett correction to other test statistics was posed by Cox (1988) and solved three years later for statistics which converge to chi-squared by Cordeiro and Ferrari (1991). We now generalize this result to general continuous statistics.

Modified statistics have been widely used to obtain good approximations for classes of statistics associated with the normal and chi-squared distributions. The principal advantage of our main result is that it applies in full generality in a number of senses. First, for rather

general parametric models, we can easily improve any continuous statistic by multipling it by an adequate adjustment factor. Second, our technique includes as special cases, some commonly used adjusted statistics, which enables us to study them within the same framewok, rather than as an unrelated collection of adjusted statistics. Our arguments will be informal without explicit attention to regularity conditions, these being essentially those required for the expansions needed for maximum likelihood theory in regular estimation problems.

Let S be a continuous statistic whose distribution function has a known first-order approximating distribution. A natural question is then: Can we find a better approximation to the distribution of the statistic in use? The purpose of our paper is to answer this question to some extent. We propose a new statistic S^* whose distribution function agrees with the first-order approximating distribution to order $O(n^{-1/2})$ or even $O(n^{-1})$ where n is the sample size. Thus, S^* is better approximated by this first-order distribution than S. The key idea for deriving the new statistic is to know the Edgeworth-type expansion for the distribution function of S in terms of the first-order approximating distribution to the required order. We assume the same conditions for the validity of the Edgeworth-type expansions (Feller, 1971; Skovgaard, 1981a, b, 1986). The infinite series expansion for the distribution of S can sometimes be divergent and we need to impose some further restrictions on the cumulants of S for using a truncated expansion to $O(n^{-1})$.

For most applications the ith cumulants k_i's of S satisfy $k_i = k_{i0} + O(n^{-1/2})$ if $i \leq 2$ and $k_i = k_{i0} + O(n^{1-i/2})$ if $i \geq 3$, as $n \to \infty$, where the cumulants k_{i0}'s refer to the first-order approximation and do not depend on n. Under these assumptions, we can obtain the Edgeworth-type expansion of the distribution of S corrected to order $O(n^{-1})$ using a truncated series with a few terms. The truncated series gives in general significant improvement over the first-order approximation.

In Section 2 we define a new statistic S^* whose distribution function is identical to the first order approximating distribution ignoring terms of order less than $n^{-1/2}$ or even n^{-1}. We show that whenever a truncated Edgeworth-type expansion for the distribution function $F_S(x)$ of S to $O(n^{-1})$ is available, a new statistic S^* can be worked out in such a way that it will generally improve the original inference. The modified statistic is determined by a simple multiplicative adjustment to the statistic S which makes the terms of order $n^{-1/2}$ and, in some cases, the terms of order n^{-1} in the asymptotic expansion of the distribution function of the modified statistic S^* vanish This scaling factor extends Bartlett's idea of correcting likelihood ratio statistics to several other types of statistics. It is called "generalized Bartlett correction" and can be given as a function of some derivatives of the first-order approximating distribution, the differences $(k_i - k_{i0})$'s between the cumulants of order greater than $O(n^{-3/2})$ and the unmodified statistic itself. Finally, in Section 3, we show through several applications that our method has the potential to be a very useful contribution to statistical literature since it comprises a very wide spectrum of improved statistics widely used to test hypotheses of interest including Bartlett-type corrected statistics which converge to chi-squared, the Cornish-Fisher polynomial transformation to normality and improved maximum likelihood estimates.

2. AN ADJUSTED STATISTIC

When testing a statistical hypothesis or estimating unknown parameters, it is often convenient to use an asymptotic approximation to the distribution of a statistic. Large sample assumptions are then commonly used in statistics since exact results are not always available. In such cases, inferences rely on what is called first-order asymptotics, i.e., they employ the quantiles of a known limiting distribution, but they may be inaccurate for small or moderate sample sizes. This section addresses the issue of obtaining a new statistic S^* which is better approximated by the first-order limiting distribution.

Let S be a given one-dimensional continuous statistic with cumulative distribution function $F_S(x)$ and density function $f_S(x)$. It is assumed that S is obtained from a sample Y of n independent observations having densities that depend on the vector parameter θ. Let Z be a scalar random variable with cumulative distribution function $F_Z(x)$, assumed to be absolutely continuous with density function $f_Z(x)$. Now suppose that $F_Z(x)$ is free from n and then $Z = O_p(1)$. Further, the distribution function $F_Z(x)$ will be assumed to be arbitrarily differentiable and that $f_Z(x) > 0$ for all x in the support of S.

Let the cumulants of S and Z be $\{k_i\}$ and $\{k_{i0}\}$ respectively. The cumulants are assumed to be known at least up to some order. In regular problems, the statistic S has a cumulative distribution function that admits an expansion in terms of the initial approximating distribution $F_Z(x)$ of Z, of the form

$$F_S(x) = F_z(x) + \sum_{i=1}^{m} (-1)^i \eta_i \frac{D^i F_z(x)}{i!}, \qquad (1)$$

where $D^i F_Z(x) = d^i F_Z(x)/dx^i$, as m $\to \infty$ (see eq. 5.6 in McCullagh, 1987). This formal Edgeworth-type expansion for the distribution function of S is given in different notation by Hill and Davis (1968) for arbitrary analytic $F_Z(x)$. In equation (1) the η_i's are "formal moments" obtained by treating the differences $(k_i - k_{i0})$'s as "formal cumulants". These "formal moments" are generally used to obtain a formal Edgeworth expansion (1) for the distribution function of S. It would be valid provided that suitable regularity conditions hold. The truncated series approximation (1) is continuous. If S is a discrete random variable, $F_S(x)$ is discontinuous with jumps of order $n^{-1/2}$ at the support points of S. For this reason, no continuous series could approximate $F_S(x)$ with uniform accuracy in any non-trivial interval of R.

In many statistical applications, we can group the terms of (1) according to their order in n. Then, sucessive terms in the re-grouped series can decrease (monotonically) in half-powers of n. Fortunately, many statistics can be given by sums of independent identically distributed random variables, and this approach can be achieved after suitable standardization of the statistic S and by choosing $F_z(x)$ adequately, for example, as the limiting distribution function of S. We can therefore write $F_s(x)$ in the form

$$F_S(x) = F_z(x) + A_1(x) + A_2(x) + O(n^{-3/2}), \qquad (2)$$

where $A_1(x)$ and $A_2(x)$ are terms of orders $O(n^{-1/2})$ and $O(n^{-1})$, respectively, which depend on some differences $(k_i - k_{i0})$'s of the cumulants of S and Z. The terms $A_1(x)$ and $A_2(x)$ may be polynomials in x but this is not always the case.

Essentially, the idea behind our procedure of modifying S is based on the fact that the distribution function $F_S(x)$ may be formally expanded as in equation (2). We shall restrict ourselves to series expansions up to order n^{-1} leaving in (2) an error that is of order $O(n^{-3/2})$. We now prove that quite generally the statistic S can be modified by suitable functions $b_1(S)$ and $b_2(S)$ of the statistic S itself of orders $n^{-1/2}$ and n^{-1} to produce an adjusted statistic S^* which has the same distribution of Z to $O(n^{-1/2})$ or, in some cases, to order $O(n^{-1})$. The form (2) of the distribution function of S suggests the use of a modified statistic defined by

$$S^* = S - b_1(S) - b_2(S), \tag{3}$$

where $b_i(S) = O_p(n^{-i/2})$, for $i = 1,2$, are additive stochastic correction terms as functions of the statistic S.

The functions $b_1(S)$ and $b_2(S)$ are now determined to make the distribution of S^* to order n^{-1}, $F_{S^*}(x)$ say, identical to $F_Z(x)$. The formula (1) of Cox and Reid (1987) is used in conjunction with (3) to derive an expansion for the distribution function of interest $F_{S^*}(x)$. Applying Cox and Reid's (1987) formula (1) to equation (3) (see also equation (3.67) in Barndorff-Nielsen and Cox, 1989), under appropriate conditions, we find to $O(n^{-1})$

$$F_{S^*}(x) = F_S(x) - E\{-b_1(S)|S=x\}f_S(x) - E\{-b_2(S)|S=x\}f_S(x) + \frac{1}{2}\frac{d}{dx}\left[E\{b_1(S)^2|S=x\}f_S(x)\right]. \tag{4}$$

Some conditions are necessary to bound the remainder term in equation (4) (see Cox and Reid, 1987). It is now straightforward to conclude from equations (2) and (4) that the equality $F_{S^*}(x) = F_Z(x)$ holds to order n-1 if and only if

$$A_1(x) + A_2(x) + b_1(x)f_s(x) + b_2(x)f_s(x) + \frac{1}{2}\frac{d}{dx}\{b_1(x)^2 f_s(x)\} = 0.$$

Using (2) and collecting terms of orders $n^{-1/2}$ and n-1 in the last equation yields

$$A_1(x) + b_1(x)f_z(x) = 0$$

and

$$A_2(x) + b_1(x)A_1'(x) + b_2(x)f_z(x) + b_1(x)b_1'(x)f_z(x) + \frac{1}{2}b_1(x)^2 f_z'(x) = 0$$

where primes denote derivatives with respect to x. It is easy to verify that these equations have at least one solution given by $b_1(x) = -A_1(x)/f_Z(x)$ and $b_2(x) = -A_2(x)/f_Z(x) + A_1(x)^2 f'_Z(x)/\{2 f_Z(x)^3\}$, provided that $f_Z(x)$ is non-zero in the support of S.
Consequently, the modified statistic S^* whose distribution function is $F_Z(x)$ to order n^{-1} is given by

$$S^* = S\left[1+\frac{A_1(S)}{f_z(S)S}+\frac{1}{S}\left\{\frac{A_2(S)}{f_z(S)}-\frac{A_1(S)^2 f_z'(S)}{2f_z(S)^3}\right\}\right]. \qquad (5)$$

The method that leads to (5) is formally correct provided only that the distribution of S has a valid Edgeworth expansion (1) up to and including the $O(n^{-1})$ term. The bracketed multiplying factor in equation (5) is a kind of stochastic adjustment involving the $n^{-1/2}$ and n^{-1} functions $A_1(x)$ and $A_2(x)$ of expansion (2), the density $f_Z(x)$ with its first derivative $f_z'(x)$ and the statistic S itself. Clearly, the terms $A_1(x)$ and $A_2(x)$ are functions themselves of certain differences between the cumulants of S and Z and of some derivatives of the distribution function, a fact that may be seen from equations (1) and (2). In general, the stochastic multiplying factor in (5) may be written as $1+b(S,\eta_i, D^i F_Z)$, where the notation emphasizes the dependence of the derivatives of the distribution function $F_Z(x)$ and "formal moments" η_i's and the unmodified statistic S. Given its similarity with the Bartlett-type correction for a class of chi-squared statistics (Cordeiro and Ferrari, 1991), the adjustment factor $1+b(S, \eta_i, D^i F_Z)$ will be called "generalized Bartlett correction".

This is a very general result which can be used to improve many important tests in statistics and econometrics.

Instead of modifying S, an alternative approach is to modify the quantiles of the reference distribution in order to make better inferences based on S. From formula (2) of Cox and Reid (1987) (see also expression (3.68) in Barndorff-Nielsen and Cox, 1989) and using the fact that S^* in (3) has distribution function $F_Z(x)$ to order n^{-1}, it follows that, to this order, $F_S(x^*)= F_Z(x)$, where $x^* = x+b_1(x) + B_2(x)$ with $B_2(x) = b_2(x) + b_1(x)b_1'(x)$. Then,

$$x^* = x\left[1-\frac{A_1(x)}{xf_z(x)}-\frac{1}{x}\left\{-\frac{A_1(x)A_1'(x)}{f_z(x)^2}+\frac{A_2(x)}{f_z(x)}+\frac{A_1(x)^2 f_z'(x)}{2f_z(x)^3}\right\}\right] \qquad (6)$$

Therefore, improved inferences can be achieved from two distinct viewpoints, which are equivalent to $O(n^{-1})$. First, we can construct a new statistic in (5) which is better approximated by the first-order approximating distribution $F_Z(x)$. Second, we can obtain a new distribution based on the modified upper percentile point (6) of our statistic S which is closer to the true distribution of S than this first-order approximating distribution. It is clear that the functional forms of the multiplicative corrections to improve the upper tail of S and to improve the statistic S itself are not in general the same, unless the $n^{-1/2}$ term $A_1(x)$ is zero.

The η_is may be function of the unknown parameters and one should use the statistic $\hat{S}^* = \left\{1+b\left(S,\hat{\eta}_i, D^i F_z\right)\right\}$ with the parameters η_i's replaced by consistent estimates $\hat{\eta}_i$'s. When $A_1(x)$ is zero, and this is the case for asymptotically χ^2 statistics, the correction term $b(S, \eta_i, D^i F_Z)$ is of order $O_p(n^{-1})$ and such replacement induces a change of order $n^{-3/2}$. Therefore, from the equivalence of formulae (3c) and (4c) of Cox and Reid (1987), it can be shown that $\hat{S}^*$ has the same distribution as Z to order n^{-1}. However, it is important to notice that if $A_1(x)$ does not vanish, the replacement of the unknown parameters by consistent estimates induces a change of order $O_p(n^{-1})$. In such case, $\hat{S}^*$ has the same distribution as Z to order $n^{-1/2}$ only.

3. SPECIAL CASES

In this section we shall consider some special cases of equation (5) in order to show its importance and usefulness to produce more accurate approximations to the distributions of statistics. Examples include Cornish-Fisher's formula for the polynomial transformation to normality, accurate formula for correcting statistics which converge to a chi-squared distribution (Cordeiro and Ferrari, 1991), development of corrections to signed likelihood ratio statistics and corrected maximum likelihood estimates. Several other special cases could also be easily obtained because of the generality of this technique for correcting statistics.

3.1 Statistics Having a Limiting Normal Distribution

Let T be a statistic whose distribution depends on parameters n and θ. Assume that there exists $\mu = \mu(\theta)$ and $\sigma = \sigma(\theta)$ such that the standardized statistic $S = n^{1/2}(T-\mu)/\sigma$ has mean zero and unit variance and higher-order cumulants of the form $k_r(S) = \rho_r n^{1-r/2}$, for $r \geq 3$, where the coefficients $\rho_r's$ depend on the cumulants of the population distribution. Further, we assume that S converges in distribution to a standard normal random variable.

It is appropriate here to consider the basic limiting distribution of S in order to guarantee an asymptotic expansion for the distribution of S in decreasing powers of $n^{-1/2}$. The Edgeworth expansion for the distribution function of S to $O(n^{-1})$ is derived in a straightforward way from (1) (see McCullagh, 1987, equation (5.12))

$$F_s(x;\rho) = \Phi(x) - \phi(x)\left[\frac{\rho_3 h_2(x)}{6\sqrt{n}} + \frac{1}{n}\left\{\frac{\rho_4 h_3(x)}{24} + \frac{\rho_3^2 h_5(x)}{72}\right\}\right] \tag{7}$$

where $\phi(x)$ and $\Phi(x)$ are the standard normal density and distribution functions, respectively. The polynomials appearing in (7) are the Hermite polynomials. They are given by $h_2(x) = x^2-1$, $h_3(x) = x_3-3x$ and $h_5(x) = x_5-10x^3 + 15x$. The remaining terms of order $O(n^{-r/2})$, for $r \geq 3$, can be found in Niki and Konishi (1986). Combining equations (2) and (7), we can see immediately that $A_1(x) = -\rho_3\,\phi(x)\,h_2(x)/\left(6\sqrt{n}\right)$ and

$$A_2(x) = -\phi(x)\left\{\rho_4 h_3(x)/24 + \rho_3^2 h_5(x)/72\right\}/n.$$

By substituting these results in equations (5) and (6), we find

$$S^* = S - \frac{\rho_3}{6\sqrt{n}}\left(S^2-1\right) + \frac{1}{12n}\left\{\frac{\rho_3^2\left(4S^3-7S\right)}{3} - \frac{\rho_4\left(S^3-3S\right)}{2}\right\}, \tag{8}$$

$$x^* = x + \frac{\rho_3}{6\sqrt{n}}\left(x^2-1\right) - \frac{1}{12n}\left\{\frac{\rho_3^2\left(2x^3-5x\right)}{3} - \frac{\rho_4\left(x^3-3x\right)}{2}\right\}, \tag{9}$$

Equation (8) is just the classical Cornish-Fisher polynomial transformation to normality when stochastic quantities of order $O_p(n^{-3/2})$ and smaller are neglected, i.e., $S^* \sim N(0,1) + O_p(n^{-3/2})$ (see McCullagh, 1987, p. 166). Also, equation (9) gives the approximate

percentage points of S expressed in terms of the standard normal percentage points ignoring quantities of order $O(n^{-3/2})$ and smaller (see McCullagh, p. 171). This special case can be regarded as a partial check of the validity of equations (5) and (6).

Many extensively used statistics can be expressed as sums of independent and identically distributed random variables $Y_1,...,Y_n$ having finite cumulants δ_r to some order. Other statistics can be accurately approximated this way. In these cases, classical results due to the central limit theory show that, under fairly general conditions, the standardized sum $S = n^{1/2}\left(\sum Y_i - n\delta_1\right)/\left(n\delta_2\right)$ converges in distribution to a standard normal random variable. This means that equations (8) and (9) hold with the two constants ρ_3 and ρ_4 being the standardized cumulants corresponding to δ_3 and δ_4, namely $\rho_3 = \delta_3 / \delta_2^{3/2}$ and $\rho_4 = \delta_4 / \delta_2^2$. Stronger forms of the central limit theorem that are validy under substantially weaker conditions than those assumed here are available to apply equations (8) and (9). The assumption that the Y_i's are independent and identically distributed random variables is not essential.

Another simple example of (8) involves the standardized random variable $S = \left(\chi_n^2 - n\right)/\sqrt{2n}$ which is asymptotically normally distributed with zero mean and unit variance. The third and fourth cumulants of S yield $\rho_3 = 2\sqrt{2}$ and $\rho_4 = 12$. Thus the adjusted random variable S^* follows from (8) as $S^* = S - \sqrt{2}\left(S^2 - 1\right)/\left(3\sqrt{n}\right) + \left(7S^3 - S\right)/\left(18n\right)$, which is asymptotically $N(0,1)$ with error $O(n^{-3/2})$. We emphasize that the Wilson-Hilferty transformation $S_1 = \left(9n/2\right)^{1/2}\left\{\left(\chi_n^2\right)^{1/3} - 1\right\}$ is not asymptotically standard normal even to order $O(n^{-1/2})$, although Cox and Reid's (1987) modification $S_2 = S_1 n^{-1/3} + \left(\sqrt{2/3}\right)n^{-5/6}$ is asymptotically $N(0,1)$ with error $O(n^{-1})$. Clearly, the form S^* is superior to S_1 and S_2 in terms of normal approximations.

3.2 Corrected Test Statistics Whose Asymptotic Distributions are χ^2

We now apply the results of Section 3 by considering a class of statistics for testing simple or composite null hypotheses whose null asymptotic distributions are central chi-squareds. This is an important class of statistics since it includes some of the most used tests, such as the likelihood ratio, Lagrange multiplier and Wald tests.

For any statistic S whose null asymptotic distribution is central chi-squared with q degrees of freedom, under mild regularity conditions, we can write its distribution function to $O(n^{-1})$ as (Chandra, 1985)

$$F_s(x) = F_q(x) + \sum_{i=0}^{k} a_i F_{q+2i}(x), \qquad (10)$$

where the a_i's of order n^{-1} are functions of the unknown parameters and $F_q(x)$ is the distribution function of χ_q^2. In addition to (10), the condition $\sum a_i = 0$ is necessary to produce a distribution function to $O(n^{-1})$. Combining formulae (2) and (10) gives $A_1(x)$ 0 and $A_2(x)$ $\sum_{i=1}^{k} a_i F_{q+2i}(x)$. From (5) and using the recurrence relation

$F_{r+2}(x) = F_r(x) - (2x/r)dF_r(x)/dx$, one can verify that the multiplying factor $1 + b(S, \eta_i, D^i F_z)$ reduces to a polynomial in S of degree at most k-1. Hence,

$$S^* = S\left\{1 - 2\sum_{i=1}^{k}\left(\sum_{l=i}^{k} a_l\right)\mu_i'^{-1} S^{i-1}\right\}, \tag{11}$$

where $\mu_i' = E\left\{\left(\chi_q^2\right)^i\right\}$. This result was first given by Cordeiro and Ferrari (1991, equation (16)). Formula (11) can be used to improve many important tests in statistics and econometrics (Cordeiro, Ferrari and Paula, 1993; Cribari-Neto and Ferrari, 1995a, b, c; Ferrari and Cordeiro, 1996). An alternative way of obtaining an improved test is to consider the unmodified statistic S together with the modified percentage points x^* given in (6). It can be easily seen that

$$x^* = x\left\{1 + 2\sum_{i=1}^{k}\left(\sum_{l=i}^{k} a_l\right)\mu_i'^{-1} x^{i-1}\right\}.$$

The usual Bartlett correction to improve the likelihood ratio statistic ω comes from (11) with $k = 1$ by noting that $a_0 = -a_1 = -b/2$, where b is the n^{-1} term in $E(\omega)$. Improved score and Wald statistics are special cases of (11) for $k=3$.

3.3 Signed Roots of Likelihood Ratio Statistics

Consider continuous random variables having density function that depends on an unknown scalar parameter θ. Let ω be the usual likelihood ratio statistic $\omega = 2\left\{l(\hat{\theta}) - l(\theta)\right\}$, where $l(\theta)$ is the total log-likelihood function. In recent years there has been considerable interest in the signed root of the likelihood ratio statistic $S = \mathrm{sgn}(\hat{\theta} - \theta)\omega^{1/2}$. The standard normal approximation to the distribution of S can be used to construct approximate confidence limits for θ having coverage error of order $n^{-1/2}$. DiCiccio (1984), Jensen (1986) and Barndorff-Nielsen (1986, 1990, 1991) among others have worked with adjustments to S that improve the accuracy of the standard normal approximation.

The most commonly adjusted statistic of S is the signed likelihood ratio standardized with respect to its mean and variance given by

$$S_1 = \frac{S - a_1/\sqrt{n}}{\left(1 + (a_2 - a_1^2)/n\right)^{1/2}} \tag{12}$$

where the quantities a_1 and a_2 are obtained from $E(S) = a_1/\sqrt{n} + O(n^{-1})$ and $E(\omega) = 1 + a_2/n + O(n^{-2})$. Thus, a_1 is the coefficient of the $n^{-1/2}$ term in the mean of S and a_2 is the n^{-1} term in the Bartlett correction for improving the distribution of ω. The standardized statistic (12) has limiting normal distribution correct to order $n^{-3/2}$. It is also possible to construct a kind of score statistic U (Barndorff-Nielsen, 1986, 1990) of the form

$U = S + O_p\left(n^{-1/2}\right)$, such that the distribution of $S_2 = S + S^{-1}\, log(U/S)$ also follows a $N(0,1)$ distribution with relative error $O(n^{-3/2})$. However, the calculation of U could be very difficult in practice. An alternative statistic S_3 to overcome such difficulties was proposed by DiCiccio and Martin (1991), although it is not as accurate as S_1 and S_2. They constructed an auxiliary statistic $T = l(\theta)J\left(\hat{\theta}\right)^{-1/2}\left\{K\left(\hat{\theta}\right)/K(0)\right\}^{1/2}$, where $J(0)$ and $K(0)$ are the observed and expected informations for θ and showed that the distribution of $S_3 = S + S^{-1}\log(T/S)$ is asymptotically $N(0,1)$ but with higher error of order n^{-1}. In general, T and U are parameterization invariants and $T = U + O_p\left(n^{-1}\right)$.

We now give a fourth corrected signed likelihood ratio statistic as an alternative to S_1 and S_2 which is easily calculated from the asymptotic expansion for the distribution of S (Jensen, 1986)

$$F_s(x) = \Phi(x) - \phi(x)\left\{\frac{a_1}{\sqrt{n}} + \frac{a_2 x}{2n}\right\} + O\left(n^{-3/2}\right).$$

This result and equation (5) yield $S^* = S\left\{1 + \left(a_1^2 - a_2\right)/(2n)\right\} - a_1/\sqrt{n}$. The statistics S_1, S_2 and S^* are equivalent to order n^{-1}. Our next project is to compare them through Monte Carlo simulations.

3.4. Modification of Standardized Maximum Likelihood Estimates

We seek a statistic that is a function of the maximum likelihood estimate and whose distribution is normal excluding terms of order $O(n^{-3/2})$ and smaller. Let Y be the data vector of length n with total likelihood function $L(\theta) = L(\theta;Y)$ depending on a scalar parameter θ. We assume that the region of the sample space for which $L(\theta;Y) > 0$ does not depend on θ and that some conditions concerning smoothness of $L(\theta;Y)$ and its derivatives with respect to θ hold. The derivatives of the log-likelihood function $l(\theta) = log\, L(\theta)$ are denoted by $U_\theta - dl(\theta)/d\theta, U_{\theta\theta} = d^2 l(\theta)/d\theta^2$, etc. The standard notation will be adopted for the cumulants of log-likelihood derivatives (Lawley, 1956): $k_{\theta\theta} = E\left(U_{\theta\theta}\right), k_{\theta\theta\theta} = E\left(U_{\theta\theta\theta}\right), k_{\theta,\theta} = E\left(U_\theta^2\right), k_{\theta,\theta\theta} = E\left(U_\theta U_{\theta\theta}\right)$, etc. We define the derivatives of the cumulants by $k_{\theta\theta}^{(\theta)} = dk_{\theta\theta}/d\theta$, etc. All k's refer to a total over the components of Y and are, in general, of order $O(n)$. Let $\hat{\theta}$ be the maximum likelihood estimate of θ assumed unique for large n.

Under regularity conditions on $L(\theta;Y)$ (Cox and Hinkley 1974, Section 9.1), it follows quite generally that the score function U_θ is asymptotically $N\left(0, k_{\theta,\theta}\right)$, so $\hat{\theta}$ satisfies $U_{\hat{\theta}} = 0$ at least for large n. Also, the asymptotic distribution of $\hat{\theta}$ is $N\left(\theta, k_{\theta,\theta}^{-1}\right)$, with error apparently $O(n^{-1/2})$. These limiting results apply directly to situations in which the components of Y, while independent, need not be identically distributed. However, they still hold to dependent data under various conditions on the type of dependence.

We shall work with the standardized statistic $S=\left(\hat{\theta}-\theta\right)k_{\theta,\theta}^{1/2}$ as a pivot function for θ which is frequently used to test the null hypothesis H_0: $\theta = \theta_0$, or to construct confidence limits for θ. The normal approximation for S is unsatisfactory in one important respect: the exact and approximate distributions of S differ by an $O(n^{-1/2})$ term. It is then desirable to improve on this result by adjusting S to have more nearly a standard normal distribution. On this basis, $Z \sim N(0,1)$ with cumulants $k_{rZ} = 0$ for $r \geq 3$ and we can obtain the formal moments $\eta's$ after some algebra:

$$\eta_1 = k_{\theta,\theta}^{1/2} b_1(\theta) + O\left(n^{-3/2}\right), \qquad \eta_2 = k_{\theta,\theta}\left\{v_2(\theta) + b_1(\theta)^2\right\} + O\left(n^{-2}\right),$$

$$\eta_3 = \rho_{3\hat{\theta}} + O\left(n^{-3/2}\right), \qquad \eta_4 = \rho_{4\hat{\theta}} + 4\rho_{3\hat{\theta}} k_{\theta,\theta}^{1/2} b_1(\theta) + O\left(n^{-2}\right).$$

Here, $b_1(\theta)$ and $v_2(\theta)$ are the n^{-1} and the n^{-2} terms in the bias and variance of $\hat{\theta}$, respectively. Also, $\rho_{3\hat{\theta}}$ and $\rho_{4\hat{\theta}}$ are the third and fourth cumulants of $\hat{\theta}$ of orders $n^{-1/2}$ and n^{-1}, respectively. Formulae for $b_1(\theta)$, $v_2(\theta)$, $\rho_{3\hat{\theta}}$ and $\rho_{4\hat{\theta}}$ are given by Shenton and Bowman (1977, Sections 2.7.6 e 2.7.). We can also show that η_r is of order smaller than n^{-1} for $r \geq 5$. The distribution function of S to $O(n^{-1})$ follows from equation (1) as

$$F_S(x;k) = \Phi(x) - \phi(x)\left\{\frac{6\eta_1 + \eta_3 h_2(x)}{6} + \frac{12\eta_2 h_1(x) + \eta_4 h_3(x)}{24}\right\},$$

from which we obtain $A_1(x)$ and $A_2(x)$. Substituting these functions into (5) and simplifying, we find

$$\begin{aligned} S^* = S &+ \frac{\eta_3 - 6\eta_1}{6} + \frac{1}{72}\left\{(\eta_3 - \eta_1)^2 + 9(\eta_4 - 4\eta_2)\right\} S \\ &- \frac{\eta_3}{6} S^2 + \frac{1}{72}\left\{2\eta_3(\eta_1 - \eta_3) - 3\eta_4\right\} S^3 + \frac{\eta_3^2}{72} S^5, \end{aligned} \tag{13}$$

where $S = \left(\hat{\theta} - \theta\right)k_{\theta,\theta}^{1/2}$. The adjusted pivotal quantity (13) is therefore a polynomial of 5th degree in the maximum likelihood estimate itself. Let $S^* = S + \sum_{i=0}^{5} \alpha_i S^i$ be this polynomial, where

$$\alpha_0 = (\eta_3 - 6\eta_1)/6, \qquad \alpha_1 = \left\{(\eta_3 - \eta_1)^2 + 9(\eta_4 - 4\eta_2)\right\}/72, \qquad \alpha_2 = -\eta_3/6,$$

$$\alpha_3 = \left\{2\eta_3(\eta_1 - \eta_3) - 3\eta_4\right\}/72, \qquad \alpha_4 = 0, \qquad \alpha_5 = \eta_3^2/72.$$

Using general formulae for $b_1(\theta), v_2(\theta)$, $\rho_{3\hat{\theta}}$ and $\rho_{4\hat{\theta}}$ given by Shenton and Bowman (1977, equations (2.30a, b), (2.31a, b)) and some Bartlett identities, which usually facilitate the computation of the *k's*, we can obtain after some algebra

$$\alpha_0 = \left(4k_{00}^{(\theta)} - k_{000}\right)/\left(12k_{\theta,\theta}^{3/2}\right),$$

$$\alpha_1 = \left(k_{\theta\theta\theta\theta} - k_{\theta\theta,\theta\theta} - 2k_{\theta\theta}^{(\theta\theta)}\right)/\left(8k_{\theta,\theta}^2\right) + \left(109k_{\theta\theta\theta}^2 - 368k_{\theta\theta}^{(\theta)}k_{\theta\theta\theta} + 448k_{\theta\theta}^{(\theta)2}\right)/\left(288k_{\theta,\theta}^3\right)$$

$$\alpha_2 = \left(k_{\theta\theta\theta} - 3k_{\theta\theta}^{(\theta)}\right)/\left(6k_{\theta,\theta}^{3/2}\right)$$

$$\alpha_3 = -\left(k_{\theta\theta\theta\theta} - 4k_{\theta\theta\theta}^{(\theta)} + 6k_{\theta\theta}^{(\theta\theta)} + 3k_{\theta\theta,\theta\theta}\right)/\left(24k_{\theta,\theta}^2\right) + \left(73k_{\theta\theta}^{(\theta)}k_{\theta\theta\theta} - 7k_{\theta\theta\theta}^2 - 120k_{\theta\theta}^{(\theta)2}\right)/\left(72k_{\theta,\theta}^3\right),$$

$$\alpha_5 = \left(k_{\theta\theta\theta} - 3k_{\theta\theta}^{(\theta)}\right)^2/\left(72k_{\theta,\theta}^3\right)$$

Notice that α_0 and α_2 are $O(n^{-1/2})$ while α_1, α_3 and α_5 are $O(n^{-1})$. Computing the α_i s from equations (14) for the model under consideration, the improved statistic S^* for the pivot $S = \left(\hat{\theta} - \theta\right)k_{\theta,\theta}^{1/2}$ follows immediately. Then, in wide generality, the new pivotal quantity S^* is asymptotically standard normal distributed to a high degree of approximation, the relative error being typically $O(n^{-3/2})$. The statistical behavior of S^* and S can be quite different in finite samples.

Formula for S^* provides the basis for obtaining the corrected version of the maximum likelihood estimate $\hat{\theta}$. It is easy to check that $\hat{\theta}^* = \hat{\theta} + \Sigma_{i=0}^{5}\alpha_i\left(\hat{\theta} - \theta\right)^i k_{\theta,\theta}^{(i-1)/2}$ follows a $N\left(\theta, k_{\theta,\theta}^{-1}\right)$ distribution and typically the error of approximation is $O(n^{-2})$. The above polynomial transformation for $\hat{\theta}^*$ looks very much like a truncated power series in the pivot $\hat{\theta} - \theta$. The statement that $S^* = \left(\hat{\theta}^* - \theta\right)k_{\theta,\theta}^{1/2} \sim \mathrm{N}(0,1) + O_p\left(n^{-3/2}\right)$ implies that $\theta = \hat{\theta}^* \pm z k_{\theta,\theta}^{-1/2}$ is an improved set of approximate confidence intervals for θ, where z is a normal upper point, i.e., values of θ outside this set are incompatible with the data.

ACKNOWLEDGEMENTS

We wish to thank Francisco Cribari-Neto for helpful comments. The financial support of CNP Brazil is also gratefully acknowledged.

REFERENCES

Barndorff-Nielsen, O.E., (1986). Inference on full or partial parameters, based on the standardized signed log likelihood ratio. *Biometrika*, **73**, 307-322.

Barndorff-Nielsen, O.E., (1990). Approximate interval probabilities. *Journal of the Royal Statistical Society* B, **52**, 485-496.

Barndorff-Nielsen, O.E., (1991). Modified signed log likelihood ratio. *Biometrika*, **78**, 557-563.

Barndorff-Nielsen, O.E. & Cox, D.R., (1984). Bartlett adjustments to the likelihood ratio statistic and the distribution of the maximum likelihood estimator. *Journal of the Royal Statistical Society* B, **46**, 484-495.

Barndorff-Nielsen, O.E. & Cox, D.R., (1989). *Asymptotic Techniques for Use in Statistics*. Chapman and Hall, London.

Chandra, T.K. (1985). Asymptotic expansions of perturbed chi-square variables. *Sankhya*, **47**, 100-110.

Cordeiro, G.M., (1983). Improved likelihood ratio statistics for generalized linear models. *Journal of the Royal Statistical Society* B, **45**, 404-413.

Cordeiro, G.M., (1985). The null expected deviance for an extended class of generalized linear models. *Lecture Notes in Statistics*, **32**, 27-34.

Cordeiro, G.M., (1987). On the corrections to the likelihood ratio statistics. *Biometrika*, **74**, 265-274.

Cordeiro, G.M., (1993a). General matrix formulae for computing Bartlett corrections. *Statistics and Probability Letters*, **16**, 11-18.

Cordeiro, G.M., (1993b). Bartlett corrections and bias correction for two heteroscedastic regression models. *Communications in Statistics-Theory and Methods*, **22**, 169-188.

Cordeiro, G.M., (1995). Performance of a Bartlett-type modification for the deviance. *Journal of Statistical Computation and Simulation*, **51**, 385-403.

Cordeiro, G.M. & Ferrari, S.L.P., (1991). A modified score test statistic having chi-squared distribution to order n-1. *Biometrika*, **78**, 573-582.

Cordeiro, G.M. & Paula, G.A., (1989). Improved likelihood ratio statistics for exponential family nonlinear models. *Biometrika*, **76**, 93-100.

Cordeiro, G.M., Ferrari, S.L.P. & Paula, G.A., (1993). Improved score tests for generalized linear models. *Journal of the Royal Statistical Society* **55**, 661-674.

Cordeiro, G.M., Paula, G.A. & Botter, D.A., (1994). Improved likelihood ratio tests for dispersion models. *International Statistical Review*, **62**, 257-276.

Cox, D.R., (1988). Some aspects of conditional and asymptotic inference: a review. *Sankhya A*, **50**, 314-337.

Cox, D.R. & Hinkley, D.V., (1974). *Theoretical Statistics*. London: Chapman and Hall.

Cox, D.R. & Reid, N., (1987). Approximations to noncentral distributions. *Canadian Journal of Statistics*, **15**, 105-114.

Cribari-Neto, F. & Ferrari, S.L.P., (1995a). Bartlett-corrected tests for heteroskedastic linear models. *Economics Letters*, **48**, 113-118.

Cribari-Neto, F. & Ferrari, S.L.P., (1995b). Second order asymptotics for score tests in generalised linear models. *Biometrika*, **82**, 426-432.

Cribari-Neto, F. & Ferrari, S.L.P., (1995c). An improved Lagrange multiplier test of heteroskedasticity. *Communications in Statistics---Simulation and Computation*, **24**, 31-44.

DiCiccio, T.J., (1984). On parameter transformations and interval estimation. *Biometrika*, **71**, 477-485.

DiCiccio, T.J. & Martin, M.A. (1991). Approximation to marginal tail probabilities for a class of smooth functions with application to Bayesian and conditional inference. *Biometrika*, **78**, 891-902.

Feller, W. (1971). *An Introduction to Probability Theory and Its Applications* 2, 2nd ed. New York: Wiley.

Ferrari, S.L.P. & Cordeiro, G.M. (1996). Corrected score tests for exponential family nonlinear models. *Statistics and Probability Letters*. To appear.

Hill, G.W. & Davis, A.W., (1968). Generalized asymptotic expansions of Cornish-Fisher type. *Annals of Mathematical Statistics*, **39**, 1264-1273.

Jensen, J.L., (1986). Similar tests and the standardized log-likelihood statistic. *Biometrika*, **73**, 567-572.

Lawley, D.N., (1956). A general method for approximating to the distribution of likelihood ratio criteria. *Biometrika*, **71**, 233-244.

McCullagh, P., (1987). *Tensor Methods in Statistics*. London: Chapman Hall.

Moller, J., (1986). Bartlett adjustments for structured covariances. *Scandinavian Journal of Statistics*, **13**, 1-15.

Niki, N. & Konishi, S. (1986). Effects of transformation in higher order asymptotic expansions. *Annals of the Institute of Statistical Mathematics*, A **38**, 371-383.

Shenton, L.R. & Bowman, K.O. (1977). *Maximum Likelihood Estimation in Small Samples*. London: Griffin.

Skovgaard, I.M. (1981a). Transformation of an Edgeworth expansion by a sequence of smooth functions. *Scandinavian Journal of Statistics*, **8**, 207-217.

Skovgaard, I.M. (1981b). Edgeworth expansions of the distributions of maximum likelihood estimators in the general (non i.i.d.) case. *Scandinavian Journal of Statistics*, **8**, 227-236.

Skovgaard, I.M. (1986). On multivariate Edgeworth expansions. *International Statistical Review*, **54**, 169-186.

CONTRIBUCIONES

LIBRES

Nuevas Gráficas de Control para Monitorear la Variabilidad

CESAR A. ACOSTA y JOSEPH J. PIGNATIELLO, Jr.

ITAM, México *Texas, A&M Univ., U.S.A.*

1. RESUMEN

El Control Estadístico del Proceso es comúnmente utilizado para monitorear la localización y la dispersión de los procesos de manufactura. Cuando se monitorea un proceso, es importante el detectar incrementos o reducciones tan pronto como sea posible. Las gráficas para monitorear la variabilidad que se utilizan actualmente tienen serias limitaciones. Algunas no son capaces de detectar reducciones en la variabilidad del proceso, otras detectan los cambios con mucho retraso. Aún no se ha analizado completamente el desempeño de las gráficas de control de variabilidad basado en el número de muestras promedio (ARL) requeridos para detectar cambios en la dispersión. En este trabajo presentamos nuevas gráficas de control para monitorear la variabilidad de procesos. Estas gráficas son del tipo CUSUM (Suma acumulativa). Demostramos que estas nuevas gráficas de control reaccionan muy rápidamente tanto a incrementos como a reducciones en la variabilidad del proceso.

2. INTRODUCCIÓN

Para monitorear la media de un proceso normal, las gráficas de control del tipo CUSUM (suma acumulativa) o del tipo EWMA (promedio móvil ponderado exponencialmente) son utilizadas debido a su excelente desempeño para detectar cambios bruscos pequeños o moderados en la media del proceso. Las gráficas del tipo CUSUM fueron propuestas por Page (1954) y las gráficas del tipo EWMA fueron presentadas por primera vez por Roberts (1959). Muchos autores han analizado y comparado el desempeño de este tipo de gráficas, sin embargo no ha sucedido lo mismo con las gráficas para monitorear la variabilidad.

De entre los trabajos que han analizado el desempeño de las gráficas para la dispersión está el realizado por Page (1963) quien propuso el uso de la gráfica CUSUM basada en rangos para monitorear incrementos en la dispersión. Tuprah y Ncube (1987) evaluaron el desempeño de una CUSUM basada en las desviaciones estándar. Crowder y Hamilton (1992) han sugerido el uso de una gráfica del tipo EWMA basada en el logaritmo natural de la varianza muestral. Recientemente Chang y Gan (1995) han propuesto el utilizar una gráfica tipo CUSUM también basada en el logaritmo natural de la varianza muestral. Sin embargo todos estos trabajos analizan exclusivamente el desempeño de las gráficas para monitorear incrementos en la dispersión de los procesos.

En este trabajo, presentamos diferentes gráficas de control CUSUM para el monitoreo de incrementos y de reducciones en la variabilidad de los procesos. Presentamos también una estimación del desempeño de estas gráficas así como el desempeño de las gráficas que existen y son usadas en la actualidad para estos propósitos.

3. LA GRÁFICA CUSUM PARA MONITOREAR LOCALIZACIÓN

Sean μ_0 y σ_0, los valores estándar establecidos para la media y desviación estándar del proceso. Aquí asumimos que estos valores son conocidos. Si definimos $Z_t = \left(\overline{X}_t - \mu_0\right) / \left(\sigma_0 / \sqrt{n}\right)$ entonces las gráficas CUSUM estándar para monitorear incrementos o reducciones en la media de un proceso normal están dadas para t = 1,2,..., por

$$SU_t = \max\left\{0, Z_t - k_u + SU_{t-1}\right\} \qquad SL_t = \max\left\{0, -Z_t - k_l + SL_{t-1}\right\} \tag{1}$$

donde $SU_0 = SL_0 = 0$. Las constantes k_u y k_l son llamados los valores de referencia. Usualmente $k_l = k_u = k$ y puesto que k optimiza de desempeño de la gráfica al detectar cambios en la media de magnitud $2k\sigma_0 / \sqrt{n}$ generalmente se le asigna un valor igual a 1/2. Los estadísticos SU_t y SL_t se comparan contra los límites de control h_u y h_l, respectivamente. Siempre que SU_t o SL_t excedan su respectivo límite la gráfica CUSUM indica la existencia de un cambio en la media del proceso.

4. NUEVAS GRÁFICAS CUSUM PARA MONITOREAR DISPERSIÓN

En esta sección presentamos tres nuevas gráficas que pueden ser usadas para monitorear la dispersión de procesos normales. Estas gráficas pueden monitorear tanto incrementos como decrementos en la dispersión. Una gráfica CUSUM es derivada del problema del punto de cambio para varianzas de procesos normales. Las otras se basan en transformaciones que normalizan la distribución de variables relacionadas con la varianza muestral.

4.1 La Gráfica CPP CUSUM

Una CUSUM para monitorear la dispersión de procesos normales puede ser derivada del test de máxima verosimilitud para el punto de cambio del proceso. A esta la llamamos gráfica CPP (punto de cambio del proceso) CUSUM.

Si $Y_{t,j}$ representa la j-ésima observación en el subgrupo t, donde j = 1, 2, ..., n, t = 1, 2, ..., T. T representa el subgrupo actual. Consideremos el modelo siguiente. Para los subgrupos $0 < t \leq \psi$, asumimos que las $Y_{t,j}$ son independientes $N\left(\mu, \sigma_0^2\right)$, y para los subgrupos $\psi < t \leq T$, asumimos que las $Y_{t,j}$ son independientes $N\left(\mu, \sigma_a^2\right)$, donde $\sigma_a \neq \sigma_0$.. Denotamos como ψ al parámetro desconocido que representa el último subgrupo antes que la varianza del proceso cambie. Si $\psi \geq T$, entonces no ha existido cambio alguno en la varianza del proceso. Para probar la hipótesis de que la varianza del proceso no ha cambiado ($\psi \geq T$), contra la hipótesis alterna de que ha existido un cambio en la varianza ($\psi < T$) consideramos

$$\begin{aligned} H_o &: \quad \sigma_1^2 = \sigma_2^2 = \ldots = \sigma_T^2 = \sigma_0^2 \\ H_a &: \quad \sigma_1^2 = \sigma_2^2 = \ldots = \sigma_\psi^2 = \sigma_0^2 \\ &\qquad \sigma_{\psi+1} = \ldots\ldots = \sigma_T^2 = \sigma_a^2 \qquad\qquad 0 \quad \psi \leq T. \end{aligned}$$

Denotemos como $\delta^+ = \sigma_{a+}^2 / \sigma_0^2$ y $\delta^- = \sigma_{a-}^2 / \sigma_0^2$ el aumento y disminución en la varianza del proceso que uno desea detectar rápidamente, respectivamente. Denotemos también a $Z_{t,j} = (Y_{t,j} - \mu) / \sigma_0$ como la j-ésima observación estandarizada del subgrupo t y por último sea $Z_t^2 = \sum_{j=1}^{n} \left[\left(Y_{t,j} - \mu \right) / \sigma_0 \right]^2$. La CPP CUSUM indicará un cambio en la varianza el proceso en el subgrupo t si

$$\max_{0 < \psi \le T} \left\{ \sum_{t=\psi+1}^{T} \left[\sum_{j=1}^{n} Z_{t,j}^2 - nk^+ \right] \right\} \ge h_u \quad \text{para} \quad \delta^+ > 1$$

$$\max_{0 < \psi \le T} \left\{ \sum_{t=\psi+1}^{T} \left[\sum_{j=1}^{n} \left(-Z_{t,j}^2 \right) + nk^- \right] \right\} \ge h_l \quad \text{para} \quad \delta^- \le 1$$

donde $k^{\pm} = \ln \delta^{\pm} / \left(1 - \left(1 / \delta^{\pm} \right) \right)$. Así la gráfica puede expresarse como

$$SU_t = max\left\{ 0, Z_t^2 - nk^+ + SU_{t-1} \right\} \qquad \mathrm{SL}_t = max\left\{ 0, -Z_t^2 + nk^- + SL_{t-1} \right\}. \tag{2}$$

4.2 La Gráfica Q_σ CUSUM

Esta gráfica se basa en la transformación normal inversa propuesta por Quesenberry (1991). Si S_t^2 representa la varianza muestral de un subgrupo de tamaño n observado al tiempo t tomada de un proceso normal con desviación normal estándar σ, y si $\sigma = \sigma_0$, entonces $Z_t = \Phi^{-1} \left\{ F_{X_{n-1}^2} \left((n-1) S_t^2 / \sigma_0^2 \right) \right\}$ tiene una distribución normal estándar. Denotamos $F_{X_{n-1}^2}(y)$ como la función de distribución X^2 con n - 1 grados de libertad, y $\Phi(z)$ la distribución normal estándar.

4.3 La Gráfica χ-CUSUM

Otra transformación que puede utilizarse para normalizar la distribución de una variable aleatoria relacionada con la varianza muestral de procesos normales fue propuesta por Wilson y Hilferty (1931). Ellos demostraron que $\sqrt[3]{X_n^2 / n}$ es aproximadamente normal con media 1 - 2/ (9n) y varianza 2/(9n). Por tanto, para un proceso normal,

$$Z_t = \frac{\left(S_t^2 / \sigma_0^2 \right)^{1/3} - \left(1 - \dfrac{2}{9(n-1)} \right)}{\sqrt{\dfrac{2}{9(n-1)}}}$$

tendrá una distribución aproximadamente normal estándar $\sigma = \sigma_0$.

Para utilizar estas dos últimas gráficas se reemplaza Z_t en (1). Se requiere además conocer los valores de referencia k_u y k_l, y los límites de control h_u y h_l. Estos pueden obtenerse por medio de simulación o por búsqueda numérica. Desde el punto de vista del

usuario, la gráfica χ-CUSUM es más fácil de implementarse que la gráfica Q_σ CUSUM puesto que la transformación es más sencilla de calcularse.

5. ESTIMACIÓN DE VALORES ARL

Page (1954) introdujo la gráfica CUSUM y derivó una ecuación integral para su desempeño en base a valores ARL. Su solución puede aproximarse por métodos numéricos (ver Kantrovich y Krylov, (1958)). Las ecuaciones integrales para estimar los valores ARL de cada gráfica CUSUM unilateral definidas por las ecuaciones (2) se puede demostrar que son

$$L^{+}(x) = 1 + L^{+}(0)F(nk^{+} - x) + \int_0^{h_u} L^{+}(u)f(nk^{+} - x + u)du \tag{3}$$

$$L^{-}(x) = 1 + L^{-}(0)\left[1 - F(nk^{-} + x)\right] + \int_0^{h_l} L^{-}(u)f(nk^{-} + x - u)du \tag{4}$$

donde $L^{\pm}(x)$ es el valor del ARL para cada gráfica cuyo valor inicial es $x \in [0, h]$. Las funciones $F(\cdot)$ y $f(\cdot)$ representan las funciones de distribución y de densidad del estadístico Z_t^2. Los parámetros $k^{\pm}$, h_u y h_l deben seleccionarse de tal forma que brinden un desempeño deseable tanto cuando el proceso está bajo control estadístico como cuando no lo está.

Para la gráfica de control CPP CUSUM el estadístico muestral está dado por $Z_t^2 = \delta\sum_{j=1}^{n}\left[Y_{t,i} - \mu/\sigma^2\right]$, donde μ es la media del proceso, σ_0 es el valor estándar para la desviación estándar del proceso, $Y_{t,i}$ es la observación i-ésima en el subgrupo t y $\delta = (\sigma / \sigma_0)2$. Si asumimos que la desviación estándar del proceso es σ, el estadístico $\sum_{j=1}^{n}\left((Y_{t,j} - \mu)/\sigma\right)^2$ tiene distribución χ_n^2 o equivalentemente, Gama(n/2 , 2). Por tanto, Z_t^2 tiene distribución Gama(n/2, 2 δ).

Para las gráficas de control Q_σ CUSUM y χ-CUSUM los estadísticos muestrales están dados respectivamente por

$$Z_t = \Phi^{-1}\left\{F_{X_{n-1}^2}\left(\delta X_{n-1}^2\right)\right\} \tag{4}$$

$$Z_t = \left[\sqrt[3]{\frac{\delta\chi_{n-1}^2}{n-1}} - \left(1 - \frac{2}{9(n-1)}\right)\right] \Big/ \sqrt{\frac{2}{9(n-1)}} \tag{5}$$

Para estas gráficas los valores ARL se aproximaron por una cadena de markov de acuerdo al procedimiento sugerido por Brook y Evans (1972).

6. COMPARACIÓN DEL DESEMPEÑO

En esta sección comparamos el desempeño de las nuevas gráficas de control con el de otras gráficas de control para la dispersión. Entre éstas están las gráficas de Shewhart R y S la gráfica de Shewhart S con línea de advertencia como fue sugerida por Page (1963), la gráfica CUSUM de rangos, la gráfica CUSUM de desviaciones estándar, y las gráficas CUSUM y EWMA del logaritmo natural de la varianza muestral, sugerida por Chang y Gan

(1995) y Crowder y Hamilton (1992), respectivamente. Para efectuar la comparación decidimos hacer que el desempeño de las gráficas de control sea independiente del valor particular de σ_0. Para esto, los estadísticos CUSUM y los valores de referencia k fueron expresados en unidades estándar. En el caso de la CUSUM de rangos propuesta por Page (1963) se expreso como

$$SU_t = max\left\{0, \left(R_t / \sigma_0\right) - k^+ + SU_{t-1}\right\} \qquad \mathrm{SL}_t = max\left\{0, k^- - \left(S_t / \sigma_0\right) + SL_{t-1}\right\} \tag{7}$$

con $k^{\pm} = d_2\ (1 \pm (\sigma_d/\sigma_0))/2$. La gráfica CUSUM de S la expresamos como

$$SU_t = max\left\{0, \left(S_t / \sigma_0\right) - k^+ + SU_{t-1}\right\} \qquad \mathrm{SL}_t = max\left\{0, k^- - \left(R_t / \sigma_0\right) + SL_{t-1}\right\} \tag{8}$$

con $k^{\pm} = c_4\ (1 \pm (\sigma_d/\sigma_0))/2$, y donde c_4 y d_2 son los factores de las gráficas de Shewhart.

Las tablas 1 y 2 adjuntas presentan los valores de ARL para la comparación del desempeño de estas gráficas. Para el caso en que exista un incremento súbito en la desviación estándar del proceso del 20% los parámetros de todas las gráficas fueron establecidos de tal forma que este incremento se detectase tan rápido como se pueda. Además todos los límites de control fueron establecidos en un nivel tal que el ARL en control tenga un valor aproximado de 200 subgrupos, y en todos los casos se consideraron subgrupos de tamaño 5.

TABLA 1

Incrementos en la desviación estándar

% inc. σ	Shew. R UCL= 2.96	Shew. S UCL= 2.89	Shew[1] S h1 = 1.53 h2 = 2.03	EWMA $\ln S^2$ k=1.06 λ=0.05	CUSUM $\ln S^2$ k=0.068 h=2.66	CUSUM R k=2.56 h=4.88	X CUSUM k=0.38 h=4.28	Q_σ CUSUM k=0.38 h=4.28	CUSUM S k=1.034 h=1.90	CCP CUSUM k=1.193 h=18.45
0	200.18	200.0	200.0	200.0	199.93	201.80	200.70	201.10	200.60	200.76
10	68.75	58.9	58.9	43.0	42.94	40.40	41.04	41.04	38.80	34.60
20	30.72	24.6	24.6	18.0	18.07	17.60	17.17	17.15	16.85	14.14
30	16.55	13.0	13.0	11.0	10.75	10.82	10.23	10.21	10.36	8.42
40	10.20	8.1	8.1	7.6	7.63	7.81	7.26	7.24	7.50	5.93
50	6.96	5.7	5.7	6.0	5.98	6.13	5.66	5.65	5.85	4.58
100	2.40	2.2	2.2	3.2	3.18	3.13	2.90	2.98	3.01	2.20

En las tablas también puede verse que las gráficas χ-CUSUM, y la Q_σ CUSUM así como la CUSUM de R tienen un mejor desempeño que las gráficas EWMA de $\ln S^2$ y CUSUM de $\ln S^2$. Además el uso de una línea de advertencia en la gráfica de Shewhart S no mejora mucho el desempeño de ésta, en términos de valores ARL. La gráfica CPP CUSUM tiene el mejor desempeño de entre todas las gráficas comparadas. Para incrementos moderados y pequeños en la desviación estándar los valores de ARL de la gráfica CPP CUSUM son hasta la mitad de aquellos que corresponden a la gráfica de Shewhart para rangos. Esto implicaría que el tiempo en detectar un cambio de esa magnitud en la dispersión pudiera reducirse a la mitad.

[1] Gráfica tipo Shemhart con límite de advertencia h_1 y límite de control h_2

TABLA 2

Reducción en la desviación estándar

% dec. σ	Shew. R UCL= 2.05	Shew. S UCL= 2.09	CUSUM lnS2 k=.43 h=5.49	Q_σ CUSUM k=.23 h=5.76	X CUSUM k=.23 h=5.75	CUSUM R k=2.093 h=4.34	CUSUM S k=.846 h=1.70	CCP CUSUM k=.793 h=11.66
0	200.28	200.01	200.01	201.10	201.20	200.95	200.15	199.64
10	133.61	133.34	47.47	44.69	44.35	45.25	44.63	38.38
20	85.37	85.37	18.96	17.58	17.41	17.41	17.01	14.15
30	51.75	51.65	10.78	10.14	10.05	9.95	9.70	8.24
40	29.41	29.24	7.17	6.94	6.92	6.88	6.70	5.96

La tabla también compara el desempeño para monitorear reducciones en la dispersión. La gráfica EWMA de ln S^2 propuesta por Crowder y Hamilton no se incluye pues no puede detectar reducciones en dispersión. En este caso los parámetros fueron establecidos para detectar tan rápido como sea posible reducciones de 20% en la desviación estándar del proceso. El desempeño de tanto la gráfica χ-CUSUM así como de la gráfica Q_σ CUSUM son similares al de la gráfica CUSUM de S, y al de la CUSUM de R. Aquí también ocurre que el desempeño de la CPP CUSUM es el mejor que el de todas las otras gráficas. Además la mejora en términos de valores ARL al usar la gráfica CPP CUSUM es mayor al detectar reducciones en la variabilidad que al detectar incrementos.

7. CONCLUSIONES

En este trabajo se han propuesto tres nuevas gráficas bilaterales del tipo CUSUM para monitorear la dispersión de procesos. Todas estas nuevas gráficas pueden ser utilizadas para monitorear tanto incrementos como reducciones en la dispersión. Estas nuevas gráficas fueron comparadas con las tradicionales gráficas de dispersión como las gráficas de Shewhart de rangos y de desviaciones estándar. También se las comparó con otras gráficas que han sido sugeridas como alternativas a las gráficas de Shewhart.

Todas las nuevas gráficas son capaces de monitorear decrementos en la variabilidad de procesos. Las gráficas χ-CUSUM y Q_σ CUSUM mostraron un desempeño comparable con las CUSUM de rangos y de desviaciones estándar. Todas estas gráficas CUSUM tienen mucho mejor desempeño que las gráficas de Shewhart de rangos y de desviaciones estándar. Sin embargo la gráfica CPP CUSUM mostró tener uniformemente el mejor desempeño siendo capaz de detectar en promedio tanto incrementos como reducciones en la dispersión mas rápido que las otras gráficas.

REFERENCIAS

Brook D. and Evans, D. A. (1972). An Approach to the Probability Distribution of Run Length. *Biometrika* **59**, pp. 539-549.

Crowder, S. V. and Hamilton, M. D. (1992). An EWMA for Monitoring a Process Standard Deviation *Journal of Quality Technology* **24**, pp. 12-21.

Chang, T. C. and Gan, F. F. (1995). A Cumulative Sum Control Chart for Monitoring Process Variance. *Journal of Quality Technology* **27**, pp 109-119.

Kantrovich, L. V. and Krylov, V. I. (1958). *Approximate Methods of Higher Analysis*, Interscience Publishers, New York, NY.

Page, E. S. (1954). Continuous Inspection Schemes *Biometrika* **41**, pp. 100-114.

Page, E. S. (1963). Controlling the Standard Deviation by CUSUM and Warning Lines. *Technometrics* **5**, pp. 307-315.

Quesenberry C.P. (1991) SPC Q Charts for Start-up Processes and Short or Long Runs. *Journal of Quality Technology* **23**, pp. 213-224.

Roberts, S. W. (1959). Control Chart Tests Based on Geometric Moving Averages. *Technometrics* **1**, pp. 239-250.

Tuprah, K. and Ncube, M. (1987). A comparison of Dispersion Quality Control Charts. *Sequential Analysis* **6(2)**, pp. 155-163.

Wilson, E. P. and Hilferty, M. M. (1931). The Distribution of Chi-square. *Proceedings of the National Academy of Science* **17**, pp. 684-688.

Análisis Fractal de Series de Tiempo

ALEJANDRO ALEGRÍA H.

ITAM, México

1. INTRODUCCIÓN

Cuando se analiza una serie de tiempo es común suponer que la estructura de autocorrelación se caracteriza por tener "memoria de corto plazo" (MCP), lo que significa que las autocorrelaciones son absolutamente sumables, $\Sigma_k |r_k| < \infty$; no obstante, existen varios fenómenos que presentan la característica de tener correlaciones seriales que decaen muy lentamente y tienen "memoria de largo plazo" (MLP), y para las cuales $\Sigma_{k=0} r_k = \infty$. Ejemplos de esto último se pueden encontrar en geofísica e hidrología (Hurst, 1951), astronomía (Pearson, 1902), agricultura (Whittle, 1956), economía (Porter-Hudak, 1990), finanzas (Peters, 1994).

De los primeros intentos para modelar procesos con MLP se pueden mencionar los trabajos de Hurst (1951), Mandelbrot y Wallis (1968) y Mandelbrot y Van Ness (1968). Los modelos que han surgido desde entonces han sido motivo de muchas y variadas investigaciones: estimación, intervalos de confianza, pruebas de hipótesis, predicción, bondad de ajuste, entre otras. Sobre este punto, el trabajo de Beran (1992) presenta una revisión de resultados inferenciales obtenidos en la última década. Por lo general, estos resultados se basan en el supuesto distribucional de normalidad, por lo que habría que tener alguna precaución en su aplicación cuando se sospecha que el proceso bajo estudio no se adapta al modelo Gaussiano, o bien no hay razón para suponer dicha distribución. En este sentido, el análisis de rangos escalados (rescaled range), denotado R/S, nos da una opción robusta de análisis que no requiere de algún supuesto distribucional.

2. ANÁLISIS R/S

El análisis R/S fue inicialmente propuesto por Hurst en 1951, y surgió de la necesidad de estudiar el comportamiento del flujo de agua en el río Nilo. Una de las preguntas que deseaba responder Hurst era si las observaciones que se tenían del río Nilo se podían considerar provenientes de un proceso totalmente aleatorio. A partir del trabajo de Einstein sobre el movimiento Browniano, Hurst propone estudiar el valor de una constante H definida a partir de la relación

$$(R/S)_n = k\, n^H. \tag{1}$$

La construcción de $(R/S)_n$ es como sigue. Sean $X_1, X_2, \ldots, X_n$, valores consecutivos de una serie de tiempo, y $\overline{X}_n$ la media correspondiente. En primer lugar se calculan las desviaciones $Z_i = X_i - \overline{X}_n$, $i = 1, 2, \ldots, n$. Posteriormente se construye una serie de valores acumulados dada por $Y_i = Z_1 + Z_2 + \ldots + Z_i$, $i = 1,2, \ldots, n$. Finalmente, se define el rango ajustado como

$$R_n = \max\{Y_1, Y_2, \ldots, Y_n\} - \min\{Y_1, Y_2, \ldots, Y_n\}.$$

Para evitar problemas con las unidades es conveniente dividir a R_n por la desviación estándar, S_n, de $X_1, X_2, \ldots, X_n$. $(R/S)_n$ está dado por el cociente R_n/S_n.

La expresión (1) significa que el rango donde toma valores la serie varía, en forma potencial, de acuerdo a la longitud del periodo de tiempo observado. Este tipo de comportamiento es el que se presenta en el estudio de fractales. De hecho, 1/H está relacionado con la dimensión fractal del espacio de probabilidad involucrado.

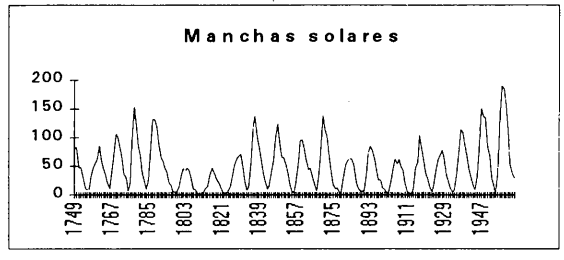
Las series de tiempo se pueden clasificar de acuerdo al valor del exponente H definido anteriormente. Cuando la serie es *aleatoria*, se tiene que $H = 0.5$. Si $0.5 < H \leq 1$, la serie se denomina *persistente*, y se caracteriza por tener una memoria que no se pierde. En este caso se tiene lo que se llama una serie fractal, que se puede describir como un movimiento Browniano fraccional. Si $0 \leq H < 0.5$, se trata de una serie *antipersistente* o ergódica.

El problema es ahora, ¿cómo estimar el valor del exponente H?. Al aplicar logaritmo a la ecuación (1), se obtiene

$$\ln(R/S)_n = \ln(k) + H \ln(n),$$

y entonces H puede estimarse por mínimos cuadrados, pues es la pendiente de un modelo de regresión lineal.

Los valores observados de $(R/S)_n$, para cada n, se obtienen al considerar todos los subperiodos contiguos de longitud n y promediar los valores de (R/S) de cada subperiodo. Aplicando el procedimiento anterior a la famosa serie de manchas solares que se muestra a continuación, se obtuvo un valor de H igual a 0.75.



Sunspot Index Data Centre, Observatoire Royal de Belgique

El valor de $H = 0.75$, parece indicar que la serie es persistente, no obstante, ¿qué tan significativo es este resultado? es decir, ¿podemos rechazar la hipótesis nula de que la serie es totalmente aleatoria? En la siguiente sección se presentan algunos resultados con respecto a este problema de inferencia sobre el parámetro H.

3. INFERENCIA

Utilizando medios bastante rústicos para simular una caminata aleatoria (lanzando una moneda) Hurst (1951) propone que $E\{(R/S)_n\}$ es igual a $(n\,\pi/2)^{0.5}$, lo cual es corroborado por Feller en 1951, al encontrar las siguientes fórmulas:

$$E\{(R/S)_n\} = (n\,\pi/2)^{0.5} \quad , \quad V\{(R/S)_n\} = (\pi^2/6 - \pi/2)\, n. \tag{2}$$

Con la ayuda de una computadora es posible validar empíricamente las anteriores expresiones. A partir de los valores obtenidos al simular un proceso gaussiano en donde hay independencia, se obtienen los siguientes valores de $\log(R/S_n)$,

número de obs.(n)	$\log(R/S)_n$ simulación	Hurst	Corrección
10	0.4577	0.5981	0.4582
20	0.6530	0.7486	0.6528
25	0.7123	0.7970	0.7120
40	0.8332	0.8991	0.8327
50	0.8891	0.9475	0.8885
100	1.0577	1.0981	1.0568
125	1.1097	1.1465	1.1097
200	1.2190	1.2486	1.2196
250	1.2710	1.2970	1.2711
500	1.4292	1.4475	1.4287
625	1.4801	1.4960	1.4792
1000	1.5869	1.5981	1.5849
1250	1.6351	1.6465	1.6348
2500	1.7839	1.7970	1.7888

Como se puede observar en la tabla anterior, los valores obtenidos al aplicar las fórmulas de Hurst dadas en (2), sobrestiman el valor obtenido por la simulación, sobre todo para valores pequeños de n. Una corrección empírica, mostrada en la misma tabla, mejora bastante la propuesta de Hurst. La corrección está dada por

$$E\{(R/S)_n\} = \{(n-0.5)\} / n)\ (n\,\pi / 2)^{-0.5} \sum_{k=1}^{n-1} \sqrt{(n-k)/k}$$

Con esta última expresión es posible generar valores esperados del exponente H. Como (R/S) se distribuye normalmente, es de esperar que H también tenga esa misma distribución. Es más, la varianza de H debería de ser $1/N$, con N el total de observaciones en la muestra. A partir de simulaciones también se corrobora esto último,

N	H, simulación	H teórico	σ_H, simulación	σ_H teórico
200	0.613	0.613	0.0704	0.0704
500	0.615	0.613	0.0451	0.0446
1000	0.615	0.613	0.0319	0.0315
5000	0.616	0.613	0.0138	0.0141
10000	0.614	0.613	0.0101	0.0100

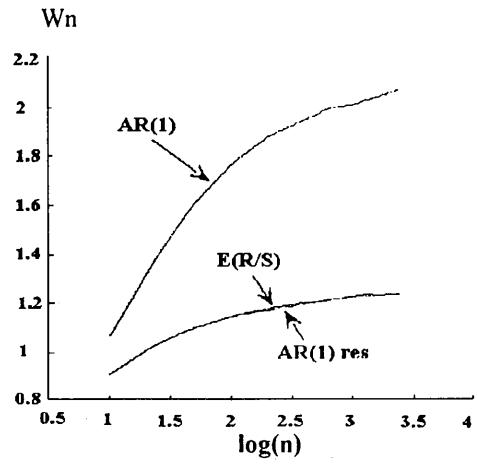
Veamos ahora qué ocurre al aplicar el análisis (R/S) a algunos procesos ARMA. Para esto es conveniente considerar la Estadística W_n definida como

$$W_n = (R/S)_n / n^{0.5}.$$

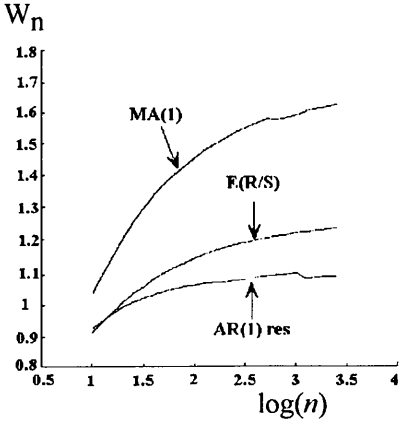
Refiriéndonos a la ecuación (1), un proceso aleatorio independiente es aquel en donde $(R/S)_n = k\, n^{0.5}$, y en este caso es claro que W_n tendría un valor constante. Por otra parte, si el proceso es persistente ($0.5 < H \leq 1$), la gráfica de W_n contra $\log(n)$ presentaría una pendiente

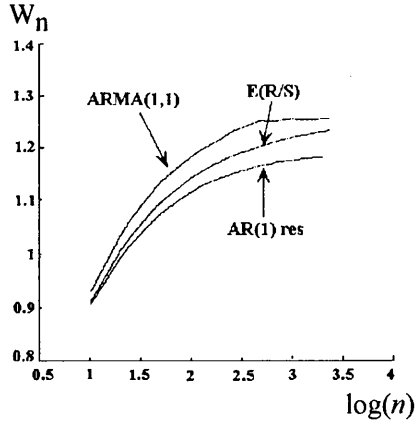
positiva. Inversamente, en un proceso antipersistente ($0 \leq H < 0.5$) la pendiente de esta última gráfica sería negativa.

Consideremos en primer lugar un proceso AR(1), dado por $Y_n = 0.5\ Y_{n-1} + e_n$. Para realizar un análisis (R/S) se generaron 5000 errores. Los resultados se presentan en la gráfica de la derecha. La Estadística W_n indica la presencia de un proceso persistente, lo cual era de esperarse porque en un proceso AR la memoria es infinita. En la misma gráfica se observa la Estadística W_n para los residuales de este modelo suponiendo que también se comportan en forma AR(1). Se presenta menor persistencia en los residuales, y de hecho ésta no es significativa.



El mismo análisis se presenta para un modelo MA(1) y otro ARMA(1,1), siendo las gráficas correspondientes,





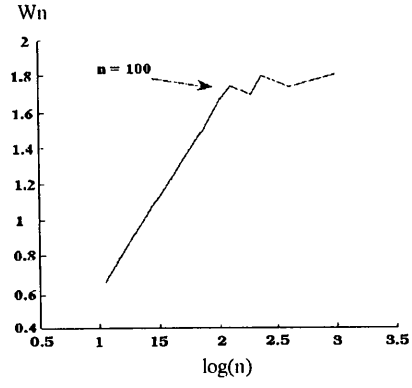
La siguiente tabla muestra los valores del exponente H para la serie original y para los residuales.

	H Serie original	*H* Resid. AR(1)	E(*H*) Residuales	*N*	*n*	repeticiones
AR(1)	0.669	0.574	0.576	5000	250	300
MA(1)	0.615	0.541	0.576	5000	250	300
ARMA(1,1)	0.669	0.568	0.576	5000	250	300

Una clase más amplia de modelos son los denominados ARIMA, los cuales permiten estudiar series no estacionarias. Un modelo ARIMA(p,d,q) considera una parte autorregresiva de orden p y otra parte de promedios móviles de orden q. Además, la serie original ha sido transformada aplicando el operador diferencia d veces, con el fin de lograr estacionariedad. Lo usual es suponer que d es un valor entero, pero si este supuesto se

elimina , se llega al concepto de diferencias fraccionales, lo que nos lleva a los modelos ARFIMA (autorregresivos, fraccionalmente integrados y de promedios móviles). En esta clase más amplia de modelos, coexisten procesos antipersistentes y persistentes, y esta es la razón por la que se piensa pueden dar mejores resultados que los modelos ARIMA. (Trabajos iniciales sobre estos modelos son los de Granger y Joyeux (1980), y Hosking (1981)). La relación entre el exponente H y el orden d de la diferencia está dada por $d = H - 1/2$.

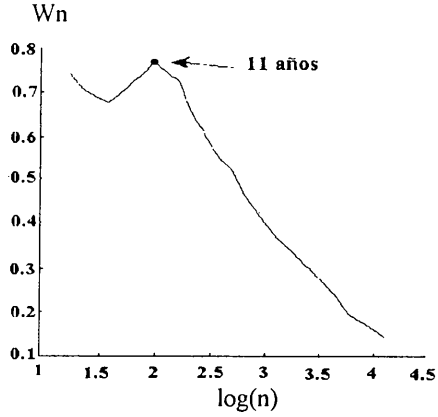
Regresando al análisis (R/S), éste también es de utilidad para detectar ciclos en una serie. Lo usual es suponer que si existe un componente cíclico en una serie, este ciclo es regular o periódico, pero, ¿cómo detectar ciclos no periódicos, es decir, aquellos ciclos que tienen en promedio cierta duración, pero se desconoce la duración de un ciclo futuro? El análisis (R/S) permite detectar ciclos periódicos y no periódicos. A manera de ejemplo consideremos el proceso Y_t definido como $Y_t = \text{sen}(t) + e_t$. Al generar observaciones de Y con un ciclo de longitud 100 y realizar un análisis (R/S), se obtiene la gráfica.



Aplicar un análisis espectral también permite detectar componentes periódicos, pero la forma en que lo hace el análisis (R/S) es lo importante. Esencialmente, una vez que el proceso ha cubierto un ciclo completo (para el ejemplo anterior el ciclo es de $n = 100$), su rango deja de crecer, porque ha alcanzado su máxima amplitud. Este comportamiento es el que se aprecia en la gráfica anterior. En procesos más complicados, el análisis (R/S) puede determinar no solamente el ciclo primario, sino también ciclos dentro de ciclos.

Los ciclos no periódicos carecen de una frecuencia absoluta, pero sí tienen una frecuencia promedio, lo que nos imposibilita saber cuál será la duración de un ciclo futuro. La existencia de este tipo de comportamiento cíclico no regular puede tener dos orígenes: (i) Existe persistencia en el sentido definido por Hurst; (ii) Es el resultado de un sistema dinámico no lineal, o caos determinístico.

En Peters (1994) se hace uso del análisis (R/S) para detectar ciclos no regulares en la serie mensual del número de manchas solares. Sus resultados muestran que efectivamente existe un ciclo no periódico de aproximadamente 11 años. La gráfica de la estadística W_n contra $\log(n)$ es



4. CONCLUSIONES

El análisis (*R/S*) nos proporciona una herramienta más para estudiar el comportamiento de procesos que presentan la característica de ser persistentes, es decir con memoria de largo plazo. El análisis es bastante robusto ante la presencia de ruido, lo cual permite estudiar sistemas dinámicos con ruido observacional o aditivo. Asimismo, no es necesario suponer normalidad para realizar un análisis (*R/S*). El problema de estimar el parámetro H requiere de una gran cantidad de operaciones y de datos, lo cual en principio no es problema si se cuenta con dichos datos y una computadora. A partir del trabajo de Hurst, varios estimadores del exponente H han sido propuestos. Suponiendo normalidad, es posible obtener estimadores máximo verosímiles, los cuales son asintóticamente normales (ver Dahlhaus, 1989). En este último caso, el estimador de Hurst es bastante menos eficiente que el máximo verosímil.

Varios son los fenómenos que tienen la característica de tener una memoria de largo plazo. Si este tipo de dependencia no es tomada en cuenta, se puede llegar a inferencias equivocadas. Más investigación es requerida en el desarrollo de teoría y métodos que permitan analizar eficazmente el fenómeno de persistencia. También queda mucho por hacer en el caso de series multivariadas y series espaciales.

REFERENCIAS

Beran, J. (1992). Statistical methods for data with long-range dependence. *Statistical Science*, **v.7**, **n.4**, 404-427.

Dahlhaus, R. (1989). Efficient parameter estimation for self-similar processes. *Ann. Statist.*, **17**, 1749-1766.

Granger, C.W.J. y Joyeux, R. (1980). An introduction to long-range time series models and fractional differencing. *J. Time Ser. Anal.*, **1**, 15-30.

Hosking, J.R.M. (1981). Fractional differencing. *Biometrika*, **68**, 165-176.

Hurst, H. E. (1951). Long-term storage capacity of reservoirs. *Transactions of the American Society of Civil Engineers*, **116**, 770-779.

Mandelbrot, B.B. y Walllis, J.R. (1968). Noah, Joseph and operational hydrology. *Water Resources Research*, **4**, 909-918.

Mandelbrot, B.B. y van Ness, J.W. (1968). Fractional Brownian motions, fractional noises and applications. *SIAM Rev.*, **10**, 422-437.

Pearson, K. (1902). On the mathematical theory of errors of judgement, with special reference to the personal equation. *Philos. Trans. Roy. Soc. Ser., A*, **198**, 235-299.

Peters, E. E. (1994). *Fractal Market Analysis*. John Wiley.

Porter-Hudak, S. (1990). An application of the seasonal fractionally differenced model to the monetary aggregates. *JASA*, **85**, 338-344.

Whittle, P. (1956). On the variation of yield variance with plot size. *Biometrika*, **43**, 337-343.

Importancia de la información Estadística en la Proposición de Modelos Cinéticos de Procesos Industriales : Análisis del Proceso FCC

JORGE ANCHEYTA-JUAREZ,
SGTI-IMP, México

H. FELIPE LÓPEZ-ISUNZA
DIPH-CBI UAM, México

y

ARMANDO NÚÑEZ-BETANCOURT
ESIQIE-IPN, México

1. INTRODUCCIÓN

En este trabajo se analiza la importancia de la información estadística en la proposición y validación de modelos cinéticos de procesos industriales, concretamente para el caso del proceso de desintegración catalítica (FCC) que es uno de los más importantes en los centros de refinación del petróleo por su alta capacidad de producción de materias primas para la elaboración de combustibles y petroquímicos. El análisis se efectúa mediante la realización de balances de materia y energía, cálculo de parámetros operacionales y determinación de rendimiento de productos, con el fin de seleccionar la información más representativa que pueda emplearse para el estudio de modelos cinéticos.

A nivel industrial las plantas de proceso trabajan bajo diversas condiciones de operación, lo cual se debe principalmente a cambios en las cargas de alimentación, necesidades de elaboración de ciertos productos, arranques de la unidad, fallas de servicios auxiliares o problemas operacionales. La información estadística que se genera durante estas etapas debe registrarse de manera adecuada y lo más completa posible para su análisis posterior, ya sea con el fin de evitar los errores que se hayan cometido o simplemente como un historial de la unidad.

En el área de investigación, un uso muy importante de esta información es en el desarrollo de modelos cinéticos que simulen los procesos industriales, ya que éstos requieren para su concepción, datos detallados sobre conversión, rendimientos de productos, condiciones de operación, características de cargas y productos, propiedades de catalizadores y aditivos, principalmente.

2. BREVE DESCRIPCIÓN DEL PROCESO FCC.

El incremento en el uso de combustibles ha requerido por parte de los refinadores una evolución acelerada en los esquemas de procesamiento del petróleo crudo, siendo uno de los objetivos principales la transformación de crudos pesados a fracciones ligeras de mayor valor comercial. Una opción para obtener gasolina a partir de fracciones pesadas es el proceso de desintegración catalítica (FCC : Fluid Catalytic Cracking). Desde su introducción, el proceso ha mantenido su posición dentro del esquema de refinación como la unidad de mayor conversión. Una de las principales razones para esto, ha sido la asombrosa capacidad del proceso para responder a los cambios en los requerimientos de producción. Además de que este proceso aporta grandes volúmenes de gasolina a la refinería y contribuye al aprovechamiento integral del petróleo crudo.

El proceso FCC convencional consiste en la combinación de dos lechos fluidizados en fase densa (sistema reactor-regenerador) y las respectivas líneas de transporte. Existen numerosos diseños de unidades FCC, generalmente se les puede dividir en tres secciones principales (Mongomery): a) Reactor-riser (línea de transferencia), b) Regenerador-manejo de gases de combustión y, c) Fraccionamiento.

Asociado con la unidad de desintegración, se encuentran las plantas de manejo de gases insaturados (olefinas) y la unidad de alquilación.

3. INFORMACIÓN ESTADÍSTICA A NIVEL INDUSTRIAL

Las principales condiciones de operación registradas en las unidades FCC se muestran en la tabla 1.

TABLA 1

Condiciones de operación

Día	1	2	3	4	5	6	7	8	9
Carga fresca, MB/D	38	38	35	35	35	35	37	37	37
Temp. de reacción, °C	512	516	518	518	517	519	519	521	518
Temp. de regeneración, °C	686	687	684	686	681	681	682	682	682
Temp. de prec. de carga, °C	311	304	279	284	287	286	292	294	299
Carbón en cat. reg., %peso	0.15	0.15	0.18	0.18	0.19	0.19	0.19	0.19	0.19
Vapor de dispersión, %peso	2.16	2.16	2.19	2.19	2.15	2.19	2.18	2.15	2.18
Vap. de agot., Kg/ton cat.	2.39	2.39	2.65	2.71	2.66	2.75	2.63	2.61	2.60

Bajo las condiciones anteriores, la desintegración catalítica se lleva a cabo para producir gases, gasolina, aceites y coque, los cuales se registran en forma volumétrica con excepción del coque que se obtiene mediante un balance de materia en peso. El gas seco se mide en m^3/D (metros cúbicos por día) y los demás productos en B/D (barriles por día). Los rendimientos de todos los productos (líquidos y gases) se determinan dividiendo su producción entre la carga alimentada y se reportan en % volumen. La conversión total se encuentra restando a cien la suma de la producción de aceites (que se consideran como lo no reaccionado de la carga), y la conversión total líquida, sumando la producción de líquidos y gases sin incluir el gas seco, estos parámetros se presentan en la tabla 2.

TABLA 2

Rendimientos globales de productos

Día	1	2	3	4	5	6	7	8	9
Gas Seco, m^3/B	7.47	7.34	8.26	7.71	7.85	7.67	6.87	7.06	6.88
Gas LP, %vol	20.53	20.79	22.14	22.14	22.00	22.29	21.35	21.00	21.08
Propanos, %vol	7.89	8.21	8.57	8.57	8.57	8.64	8.54	8.40	8.43
Butanos, %vol	12.64	12.58	13.57	13.57	13.43	13.65	12.81	12.60	12.65
Gasolina, %vol	60.03	60.05	61.14	60.57	60.34	60.57	60.88	60.54	60.00
Aceite Cíclico Ligero, %vol	20.00	20.00	20.00	19.89	20.57	20.91	20.43	21.41	20.76
Aceite Decantado, %vol	5.79	5.26	4.57	4.86	5.18	5.96	5.00	5.95	6.08
Conversión Total, %vol	74.21	74.74	75.43	75.26	74.25	73.13	74.57	72.65	73.16
Conversión Total líq., %vol	106.3	106.1	107.9	107.5	108.1	109.7	107.7	108.9	107.9

Las corrientes que a nivel industrial se denominan "gas seco", "propanos" y "butanos", en realidad no han sido fraccionadas completamente, por ejemplo, en el gas seco, que debe contener únicamente hidrógeno, ácido sulfhídrico, metano, etano y etileno, existen cantidades pequeñas de compuestos más pesados como se observa en la tabla 3, es por esto

que es necesario restar estas cantidades del gas seco e integrarlas a las que correspondan, lo mismo sucede con las corrientes de propanos y butanos.

TABLA 3

Compuestos pesados presentes en el gas seco

Día	1	2	3	4	5	6	7	8	9
Propanos, %vol	0.94	1.00	1.13	1.03	1.05	1.03	0.94	0.96	1.07
Propano, %vol	0.14	0.15	0.17	0.14	0.15	0.14	0.13	0.13	0.14
Propileno, %vol	0.80	0.85	0.96	0.89	0.90	0.89	0.81	0.83	0.93
Butanos, %vol	0.12	0.14	0.16	0.15	0.15	0.15	0.13	0.14	0.08
Total (Gas LP), %vol	1.06	1.14	1.29	1.18	1.20	1.18	1.07	1.10	1.15

Después de efectuar la separación de productos mencionada anteriormente, los rendimientos de gases quedan de la siguiente manera :

TABLA 4

Rendimientos volumétricos netos de gases

Día	1	2	3	4	5	6	7	8	9
Gas seco, m^3/B	7.00	6.83	7.68	7.18	7.31	7.14	6.39	6.57	6.36
Propanos, %vol	9.31	9.26	10.13	9.83	9.75	9.96	9.61	9.92	9.95
Butanos, %vol	12.33	12.71	13.35	13.54	13.50	13.57	12.85	12.23	12.33

Con esta información y con las condiciones de operación (tabla 1), pueden efectuarse los balances de materia y energía para obtener los rendimientos másicos de productos, así como la relación catalizador/aceite (C/O) que es una de las principales variables que indican la severidad del proceso (tabla 5).

TABLA 5

Rendimientos másicos de productos

Día	1	2	3	4	5	6	7	8	9
Relación C/O	6.19	6.41	6.94	7.05	7.25	7.30	6.75	6.76	6.51
Conversión total, %peso	72.42	73.05	73.87	73.65	72.55	71.27	72.91	70.78	71.29
Gas Seco, %peso	5.77	5.73	6.40	6.10	6.15	6.01	5.38	5.74	5.57
Gas LP, %peso	12.76	12.91	13.76	13.76	13.66	13.84	13.24	13.02	13.08
Propanos, %peso	4.52	4.70	4.90	4.90	4.90	4.94	4.88	4.80	4.82
Butanos, %peso	8.24	8.21	8.86	8.86	8.76	8.90	8.36	8.22	8.26
Gasolina, %peso	49.67	49.69	50.59	50.12	49.93	50.12	50.38	50.09	49.65
Aceite Cíclico Lig., %peso	20.71	20.71	20.71	20.59	21.30	21.66	21.16	22.17	21.50
Aceite Decantado, %peso	6.87	6.24	5.42	5.76	6.15	7.07	5.93	7.05	7.21
Coque, %peso	4.13	4.24	4.53	4.63	4.65	4.60	4.24	4.22	4.31
Total, %peso	99.91	99.52	101.41	100.96	101.84	103.30	100.33	102.29	101.32

Como se observa en la tabla anterior, la cantidad total de productos rebasa ligeramente el 100% y en algunos casos es menor, lo cual es común en procesos industriales, ya que los instrumentos de medición y el personal que toma las lecturas de los flujos, no registran variaciones muy pequeñas de los mismos, pero mientras el ajuste en el balance de materia sea de ± 3 %, éste puede considerarse bastante aceptable. Sin embargo, esto puede eliminarse mediante una normalización de rendimientos, que consiste básicamente en distribuir en forma proporcional esta diferencia en todos los productos (tabla 6).

TABLA 6

Rendimientos másicos normalizados de productos

Día	1	2	3	4	5	6	7	8	9
Conversión total, %peso	72.39	72.92	74.23	73.90	73.04	72.19	72.99	71.44	71.66
Gas Seco, %peso	5.78	5.76	6.31	6.04	6.04	5.82	5.36	5.61	5.50
Gas LP, %peso	12.77	12.97	13.57	13.63	13.41	13.40	13.19	12.73	12.91
Propanos, %peso	4.52	4.72	4.83	4.85	4.81	4.78	4.86	4.69	4.76
Butanos, %peso	8.25	8.25	8.74	8.78	8.60	8.62	8.33	8.04	8.15
Gasolina, %peso	49.71	49.93	49.89	49.64	49.03	48.52	50.22	48.96	49.00
Aceite Cíclico Lig., %peso	20.73	20.81	20.42	20.39	20.92	20.97	21.09	21.67	21.22
Aceite Decantado, %peso	6.88	6.27	5.35	5.71	6.04	6.84	5.92	6.89	7.12
Coque, %peso	4.13	4.26	4.46	4.59	4.56	4.45	4.22	4.14	4.25
Total, %peso	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0

4. MODELO CINÉTICO DEL PROCESO FCC.

Existen en la actualidad varios modelos cinéticos reportados en la literatura para representar las reacciones que se llevan a cabo en el proceso FCC (ver Weekman, 1970; Yen, 1989; Jacob, et al., 1976). Todos involucran, en base a su complejidad, cierto número de parámetros que es necesario estimar empleando información experimental generada para tal efecto, es decir, bajo condiciones de operación controladas y con unidades a nivel microrreactor Ancheyta y López (1995a).

Por ejemplo, el modelo cinético de 6 pseudocomponentes (lumps o familias) propuesto por Ancheyta y López (1995b), involucra en total 10 parámetros (9 constantes cinéticas y 1 constante de desactivación del catalizador), el manejo de los productos en forma de pseudocomponentes facilita la estimación de parámetros, por lo que en este trabajo, se han agrupado de la siguiente manera :

- Gas seco (ácido sulfhídrico, hidrógeno, metano, etano, etileno)
- Propanos (propano, propileno)
- Butanos (n-butano, isobutano, butilenos)
- Gasolina (34-216°C)
- Gasóleo (Aceite cíclico ligero 216-344°C + Aceite decantado 344°C$^+$)
- Coque

los pseudocomponentes anteriores se relacionan en las siguientes ecuaciones de velocidad de reacción :

$$\frac{dY_{A_1}}{dt} = -\left(k_1 + k_2 + k_3\right)Y_{A_1}^2\phi = -k_{A_1}Y_{A_1}^2\phi \tag{1}$$

$$\frac{dY_{A_2}}{dt} = \left[k_1 Y_{A_{11}}^2 - \left(k_4 + k_5 + k_6\right)Y_{A_2}\right]\phi = \left(k_1 Y_{A_1}^2 - k_{A_2}Y_{A_2}\right)\phi \tag{2}$$

$$\frac{dY_{A_3}}{dt} = \left[k_2 Y_{A_1}^2 + k_4 Y_{A_2} - \left(k_7 + k_8\right)Y_{A_3}\right]\phi = \left(k_2 Y_{A_1}^2 + k_4 Y_{A_2} - k_{A_3}Y_{A_3}\right)\phi \tag{3}$$

$$\frac{dY_{A_4}}{dt} = \left(k_5 Y_{A_2} + k_7 Y_{A_3} - k_9 Y_{A_4}\right)\phi \tag{4}$$

$$\frac{dY_{A_5}}{dt} = \left(k_8 Y_{A_3} + k_9 Y_{A_4}\right)\phi \tag{5}$$

$$\frac{dY_{A_6}}{dt} = \left(k_3 Y_{A_1}^2 + k_6 Y_{A_2}\right)\phi \tag{6}$$

$$\phi = e^{-\alpha t} \tag{7}$$

donde :

A_1	: Gasóleo	Y_i	:	Rendimiento del producto *i*
A_2	: Gasolina	k_i	:	Constante cinética
A_3	: Butanos	ϕ	:	Función de desactivación
A_4	: Propanos	α	:	Constante de desactivación
A_5	: Gas seco	t	:	Tiempo de reacción
A_6	: Coque			

La determinación de los parámetros cinéticos debe hacerse basándose en la información experimental mediante métodos de estimación no lineal como el propuesto por Marquardt.

5. INFORMACIÓN ESTADÍSTICA PARA LOS MODELOS CINÉTICOS

Una vez evaluados los parámetros cinéticos y la influencia de ciertas variables (p. ej. temperatura de reacción, propiedades de la carga, etc.), así como la integración de estos en un modelo matemático que simule el sistema reactor-regenerador del proceso FCC ya sea en forma dinámica o en estado estacionario López (1992), López, et al. (1994), es necesario validar el funcionamiento de dicho modelo empleando información obtenida a nivel industrial. Para el caso del modelo cinético explicado anteriormente, se requiere de los datos que se detallan en la tabla 7, los cuales fueron generados con los rendimientos normalizados presentados en la tabla 6.

TABLA 7

Información industrial sobre rendimientos de productos para validar el modelo matemático del proceso FCC

Día	1	2	3	4	5	6	7	8	9
Relación C/O	6.19	6.41	6.94	7.05	7.25	7.30	6.75	6.76	6.51
Conversión total, %peso	72.39	72.92	74.23	73.90	73.04	72.19	72.99	71.44	71.66
Gas Seco, %peso	5.78	5.76	6.31	6.04	6.04	5.82	5.36	5.61	5.50
Propanos, %peso	4.52	4.72	4.83	4.85	4.81	4.78	4.86	4.69	4.76
Butanos, %peso	8.25	8.25	8.74	8.78	8.60	8.62	8.33	8.04	8.15
Gasolina, %peso	49.71	49.93	49.89	49.64	49.03	48.52	50.22	48.96	49.00
Coque, %peso	4.13	4.26	4.46	4.59	4.56	4.45	4.22	4.14	4.25
Gasóleo, %peso	27.61	27.08	25.77	26.10	26.96	27.81	27.01	28.56	28.34

Para considerar aceptable el modelo propuesto, éste debe predecir razonablemente el comportamiento del sistema reactor-regenerador con el fin de determinar posteriormente, regiones óptimas de operación que maximicen el rendimiento de productos líquidos.

6. CONCLUSIONES

La información estadística generada a nivel industrial es indispensable para el desarrollo y validación de modelos cinéticos que representen el esquema de reacción y los modelos matemáticos que simulen el efecto de las variables de operación principalmente.

Es necesario que esta información, antes de ser empleada para el propósito ya mencionado, sea analizada detalladamente con el fin de verificar que no hallan existido situaciones anómalas, que puedan influir en la operación normal de las unidades, si esto ocurre, deben descartarse los datos del período correspondiente.

En el caso del proceso analizado (FCC), los datos que se registran directamente en las unidades industriales tienen que integrarse en los balances de materia y energía para determinar otros parámetros operacionales, y sobre todo para verificar que las pérdidas o ganancias de masa no rebasen un límite preestablecido (p.ej. ± 3 %). Si dichas variaciones son muy altas, esta información no debe considerarse y deberán verificarse que los medidores de flujo estén funcionando correctamente y bien calibrados.

El intervalo de condiciones de operación industriales seleccionado debe ser lo más amplio posible con el fin de estudiar sus efectos mediante el modelo matemático y así determinar su validez en cuanto a la variación de las mismas, para el caso en estudio, se tuvieron datos estadísticos con los siguientes intervalos de condiciones de operación : Temperatura de reacción de 512 a 521°C; Temperatura de regeneración de 681 a 687°C; Temperatura de precalentamiento de carga de 279 a 311°C; Relación catalizador/aceite de 6.19 a 7.25.

REFERENCIAS

Ancheyta, J.J. y López, I.F. (1995a). *Obtención de información experimental para la estimación de parámetros cinéticos en el proceso de desintegración catalítica FCC*; SPOCI-I, UAM-I, México.

Ancheyta, J.J. y López, I.F. (1995b). *Aspectos importantes en el modelado cinético del proceso de desintegración catalítica* FCC; XVI Enc. Nal. AMIDIQ; S.L.P.

Jacob, S.M., Gross, B. y Voltz, S.E. (1976). *AIChE J.*, **22**(4), pp. 701-713

López, I.F. (1992). Dynamic modelling of an industrial fluid catalytic cracking unit; *european Symposium on Computer Aided Process Eng.*-**1**, pp. s139-s148.

López, I.F., Esparza, Y.T., Aréchiga, V.U. y Ruiz, M.R. (1994). *Un modelo de estado estacionario para el reactor de cracking catalítico*; UAM-I, México.

Marquardt, D.W. Solution of non-linear squares estimation of non-linear parameters, *J. Soc. Ind. Appl. Math.*, **2**, 431-441.

Montgomery, J.A. The Evolution of the fluid catalytic cracking unit; *The Davison Chemical Guide to Catalytic Cracking*, pp. **9**-21.

PEMEX-Refinación. (1994). *Información estadística de la Unidad FCC*, Refinería de Salina Cruz, Oax., México.

Weekman, V.W. y Nace, D.M. (1970). Kinetics of catalytic cracking selectivity in fixed, moving and fluid bed reactors; *AIChE J.*, Vol. **16**, No. 3, pp. 397-404.

Yen, L.C. (1989). *Kinetic modeling of fluid catalytic cracking*; AIChE Spring National Meeting; Houston, Texas.

Metodología Experimental Taguchi Aplicada a Una Parte Automotriz de Inyección de Plástico

J.F. BURGUETE y T. KRAP

Univ. de las Americas, Puebla, México

1. INTRODUCCIÓN

En este trabajo de investigación aplicada, se pretende ilustrar el proceso de experimentación industrial a través de un experimento automotriz. Las cabeceras de los asientos para modelos A3 de VW tienen ciertas especificaciones 60 $\pm$ 20N para accionarse manualmente hacia arriba y hacia abajo. El esfuerzo real requerido era excesivo. El objetivo de este trabajo fue analizar las causas que provocan que el esfuerzo de extracción sea demasiado.

El esfuerzo requerido para su accionamiento es el resultado de la fricción entre un herraje metálico y 2 bujes plásticos. Se tiene evidencia que muestra que el diámetro del herraje y el ángulo de éste son correctos. Las condiciones de inyección del buje son muy variables, por lo que se busca maximizar la dimensión en un punto específico del buje en centésimas de milímetro , dada la contracción que éstos presentan.

2. FACTORES A ESTUDIAR

Con la colaboración de un grupo de trabajo y mediante un diagrama de Ishikawa se analizaron las posibles causas que producen la contracción del buje. Los cuales se redujeron a solamente siete. Se clasificaron de acuerdo a si eran controlables o no se deseaban controlar. Las variables controlables o factores de ajuste son:

A: temperatura (°C)
B: tiempo de enfriamiento (seg)
C: postpresión (%)
D: tiempo de postpresión (seg)
E: velocidad de inyección (%),

y las variables que no se deseaban controlar o factores de ruido son:

F: material reciclado (%)
G: lado de inyección (izq, der)

Adicionalmente, el grupo técnico considero conveniente estudiar las siguientes interacciones:

AB, AC, AD, BC, CD, y ABC.

3. METODOLOGÍA

Se propuso un diseño experimental producto (arreglo interno X arreglo externo) para generar unidades robustas. Esta metodología puede considerarse dentro de las propuestas por Taguchi. En la Tabla 1 se presenta su estructuración a través de un L16 X L 4.

Tabla 1. Arreglo L16 x L14

		L4	
		ARREGLO EXTERNO	
		Material reciclado	10% 0%
	L 16	Lado de inyección	izq. der.
ARREGLO INTERNO			
Temperatura	260°C 240°C		
Tiempo de Enfriamiento	35 seg 28 seg		
Postpresión	45% 15%		
Tiempo de Postpresión	5 seg 3 seg		
Velocidad de Inyección	40% 20%		

Para el arreglo interno se asignaron los factores principales y sus interacciones a un arreglo ortogonal L16, buscando obtener las interacciones de interés, incluso la triple, sin confundirlas con otras posibles interacciones significativas. La asignación se muestra en la figura 1.

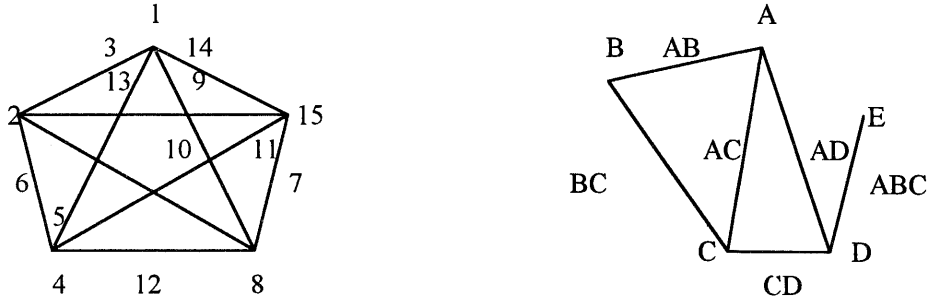


Figura 1. Asignación de los factores a las columnas del arreglo L16

Columna No. 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15
asignación A B AB C CA BC ABC D AD e e CD e e E

En el caso del arreglo externo, la asignación fue la siguiente:
Columna No. 1 2 3
Asignación: F G

4. RELACIÓN DEL ARREGLO TAGUCHI CON LA ESTRUCTURA ALIAS

En el subgrupo intrabloque del diseño factorial 2^{k-p}, se tienen k-p soluciones linealmente independientes.

Si se estructuran ecuaciones módulo 2 con las soluciones, se puede obtener una solución, ésta es la relación de definición en el diseño factorial fraccionado.

En este caso: 2^{5-1}, se tienen cuatro soluciones linealmente independientes.

Tabla 2.

Notación Taguchi (1,2)	Notación Clásica (0,1)
1 1 1 2 2	0 0 0 1 1
1 1 2 1 2	0 0 1 0 1
1 2 1 1 2	0 1 0 0 1
2 1 1 1 2	1 0 0 0 1

Tabla 3.

ECUACIONES:			D		+E	= 0	(mod 2)
			C		+E	= 0	(mod 2)
		B			+E	= 0	(mod 2)
	A				+E	= 0	(mod 2)
SOLUCION:	0	0	0	0	0		
	1	1	1	1	1		

Por tanto, la relación de definición es: I = A B C D E.

Nota: Obsérvese que se tenía interés en estudiar ABC. ABC es alias con DE, pero DE no existe técnicamente.

5. REALIZACIÓN DEL EXPERIMENTO

El experimento se llevó a cabo en una compañía inyectora de materiales plásticos de ingeniería, y las mediciones se llevaron a cabo en una máquina de medición por coordenadas de control numérico haciendo uso del paquete GEOPAK 2.0.

Los resultados se presentan en la Tabla 4.

Tabla 4. Dimensión del buje en mm.

CORRIDA	1	2	3	4
1	17.5875	17.5725	17.6250	17.6375
2	17.5975	17.6175	17.6250	17.6425
3	17.5800	17.6700	17.6275	17.6350
4	17.6075	17.6125	17.6425	17.6425
5	17.5925	17.5900	17.6450	17.6675
6	17.6375	17.6425	17.6500	17.6675
7	17.5200	17.6150	17.6550	17.6400
8	17.6375	17.6000	17.6775	17.6650
9	17.550	17.5525	17.5925	17.5700
10	17.5725	17.5725	17.6200	17.6150
11	17.5650	17.5650	17.6225	17.5925
12	17.5950	17.5950	17.6200	17.6450
13	17.5775	17.5775	17.6600	17.6575
14	17.6150	17.6150	17.6375	17.6525
15	17.59.50	17.5950	17.6200	17.5925
16	17.6100	17.6100	17.6575	17.6600

6. ANÁLISIS DE DATOS

6.1 Análisis del proceso:

Se analizó la capacidad de proceso obteniendo los siguientes resultados para el cálculo de la habilidad del proceso:

$$\bar{x} = 17.61$$
$$s = 0.034$$
$$Cpk = -1.80$$
$$Cp = 1.96$$

Obsérvese que la media está fuera de especificaciones por lo que se toma el Cpk en términos reales y no en valor absoluto. El proceso es hábil potencialmente, sin embargo no lo es realmente.

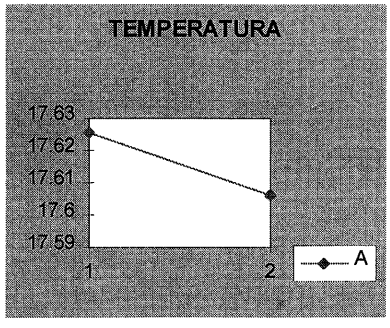
6.2 Análisis de los factores:

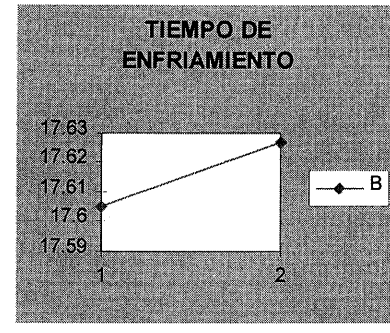
En la Tabla 5 se presenta el Análisis de Varianza

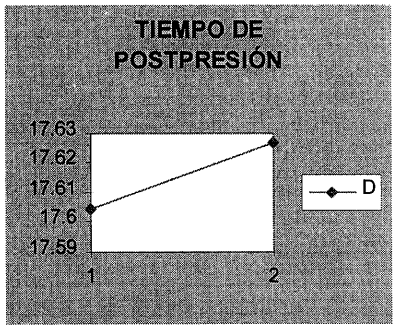
Tabla 5. Análisis de Varianza

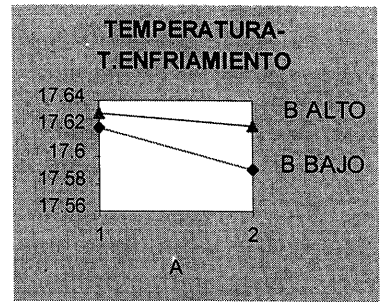
FV	GL	SC	CM	F	
A: TEMPERATURA	1	0.006300	0.006300	15.66	*
B: TIEMPO ENFIRAMIENTO	1	0.007656	0.007656	19.03	*
C: POSTPRESION	1	0.000127	0.000127	< 1	
D: TIEMPO POSTPRESION	1	0.007877	0.007877	19.58	*
E: VEL. INYECCION	1	0.000000	0.000000	< 1	
AB	1	0.001800	0.001800	4.47	*
AC	1	0.000014	0.000014	< 1	
AD	1	0.000156	0.000156	< 1	
BC	1	0.002300	0.002300	5.71	*
CD	1	0.000079	0.000079	< 1	
ABC	1	0.000000	0.000000	< 1	
el	4	0.001543	0.000386		
F	1	0.026610	0.026610		
G	1	0.000452	0.000452		
e2	46	0.018570	0.000404		
TOTAL	63	0.074200			
e*	50	0.020113	0.000402		

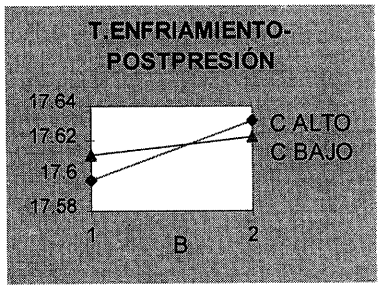
Los factores principales significativos fueron: Tiempo de Postpresión, Tiempo de Enfriamiento, y Temperatura; así como las interacciones: temperatura por tiempo de enfriamiento y tiempo de enfriamiento por postpresión. Las gráficas de estos factores son:











7. RECOMENDACIONES

Con base en los resultados del análisis de la varianza y el apoyo en las gráficas, se tienen las siguientes recomendaciones iniciales:

* Usar 10% de material reciclado.
* Temperatura: 240 C
* Tiempo de postpresión: 5 segundos
* Tiempo de enfriamiento: 35 segundos

8. CONCLUSIONES

Las interacciones, aunque significativas, no contradicen lo que se recomienda para los efectos principales.

Solamente se tiene base estadística para recomendar esta temperatura y estos tiempos. Pero aparentemente, si el proceso permite disminuir temperatura y aumentar tiempos, puede ser recomendable.

Elevar el tiempo de producción resulta una solución contraria a la filosofía de producción; de cualquier manera, se recomienda llevar a cabo mayor experimentación para tener las piezas dentro de las especificaciones deseadas.

REFERENCIAS

Richardson, T.L. and others. (1989) *Industrial plastics: theory and application*, Second Edition Delmar Publishers. Albany. New York.

Das, M.N. and Giri, N.C. (1986). *Design and analisis of experiments*. Halsted Press Book. Wiley. New York.

Ross, P.J. (1989). *Taguchi techniques for quality engineering*. Mc. Graw Hill. New York.

Exceso de Falsas Alarmas en la Aplicación de las Pruebas Estándar a la Carta c, y Alternativas para Reducirlo

OSVALDO CAMACHO CASTILLO y HUMBERTO GUTIÉRREZ PULIDO

Univ. de Guadalajara, México

1. INTRODUCCIÓN

Las ocho pruebas más usuales para detectar cambios especiales en las cartas de Shewhart son las siguientes:

Prueba 1. Un punto fuera de los límites de control.
Prueba 2. Dos de tres puntos consecutivos en la zona A.
Prueba 3. Cuatro de cinco puntos consecutivos en la zona B o más allá, sin salirse de los límites de control.
Prueba 4. Nueve puntos consecutivos de un sólo lado de la línea central, sin salirse de los límites de control.
Prueba 5. Seis puntos consecutivos en aumento (o en disminución).
Prueba 6. Catorce puntos consecutivos alternando entre altos y bajos.
Prueba 7. Ocho puntos consecutivos a ambos lados de la línea central con ninguno en la zona C.
Prueba 8. Quince puntos consecutivos en la zona C, arriba o abajo de la línea central.

Estas pruebas se derivaron a partir del supuesto de que los datos generados por el proceso se distribuyen normal e independientemente, Western Electric (1958). Sin embargo estas pruebas, con algunas restricciones, se han venido aplicando a las cartas de atributos (p, np, c y u); ya que así lo recomienda la literatura clásica de control de calidad, por ejemplo:

- Western Electric (1958) dice "En las más de las cartas donde los límites de control son razonablemente simétricos, es suficientemente seguro aplicar las pruebas estándar".
- Nelson (1987) recomienda "Las pruebas 1, 5 y 6, pueden ser usadas en las cartas p, np, c y u. También la prueba 2, si las distribuciones son suficientemente simétricas. Use las tablas de la distribución Binomial o Poisson para verificar situaciones específicas".
- Montgomery (1991) presenta las pruebas de manera general para las cartas de control de Shewhart, y no asume una posición explícita sobre cuáles se deben usar en las cartas de atributos. Besterfield (1990) señala "Se sugiere al lector que revise la sección -State Control- en el capítulo 3 (cartas para variables continuas), ya que mucha de tal información es aplicable a las cartas de defectos".
- ReVelle and Harrington (1992) dicen "La falta de control estadístico puede ser detectada por cualquiera de las siguientes pruebas... (y a continuación se describe a grandes rasgos las ocho pruebas estándar)".

Como se puede apreciar, en las recomendaciones hay ambigüedades, omisiones y contradicciones, y lo que es peor, algunas están equivocadas, Camacho y Gutiérrez (1995). Lo anterior puede provocar que el usuario de las cartas de control aplique indiscriminadamente las pruebas y eso lleve a declarar con frecuencia que el proceso estuvo fuera de control estadístico, cuando en realidad no ocurrió así.

Siguiendo los trabajos de Camacho y Gutiérrez (1995 ab) para las cartas p y np, en este trabajo se hace algo similar para la carta c, es decir, aquí se presentan los resultados de un estudio de la significancia de las pruebas estándar aplicadas a las cartas c, en los que se demuestra que tal

aplicación genera una mayor cantidad de falsas alarmas que las que se esperarían bajo normalidad, por lo que no es suficientemente seguro aplicar tales pruebas a las cartas c. También se propone y avalúa una modificación en la construcción de la carta c, que permite aplicar con mayor confianza las pruebas estándar y reducir la frecuencia de falsas alarmas, concluyéndose que tal propuesta es viable.

Carta c. Recordemos que a través de la carta c se analiza la variabilidad del número de defectos por unidad (subgrupo), y se supone que tal número se ajusta razonablemente bien a una distribución de Poisson, por lo que los límites de control están dados por

$$\bar{c} \pm 3\sqrt{\bar{c}}$$

donde c es el número promedio de defectos por unidad.

2. ESTUDIO DE SIGNIFICANCIA

Para las pruebas se calculó la probabilidad de que mediante éstas se detecte una señal de falta de control cuando en realidad el proceso no ha cambiado, es decir, se calculó el valor de a para distintos valores del parámetro λ de una distribución Poisson. Para las pruebas 1 a 4, 7 y 8, el cálculo se hizo de manera exacta, mientras que para las pruebas 6 y 7, se estimó la probabilidad mediante el método Monte Carlo. Para evaluar la magnitud de las α encontradas se tomó como punto de referencia la significancia de cada prueba bajo el supuesto de normalidad.

Los resultados se presentan de forma separada para valores de $\lambda<1$, y de $1<\lambda<20$. La síntesis de estos resultados se muestran mediante los estadísticos de la tabla 1. De donde se concluye lo que sigue.

En general las pruebas problemáticas debido al exceso de falsas alarmas son:

Prueba 1, Lado Superior,
Prueba 2, Lado Superior,
Prueba 3, Lado Superior $c<1$ y Lado Inferior $c>1$,
Prueba 4 lado inferior, $c<1$ (c pequeños).
Prueba 8, en toda la carta.

Cabe destacar que los problemas se agudizan cuando c es pequeña, pero aunque disminuyen cuando c se incrementa, éstos siguen persistiendo, en el sentido que para un número importante de valores de c las significancias siguen siendo más altas que bajo normalidad. En particular, se puede ver que la Prueba 1, del Lado Superior que bajo normalidad tiene una significancia de 0.00135, presenta niveles de significancia mayores a este valor. Cuando $c>1$, $\bar{\alpha} = 0.003916$, $S = 0.00183$, y el 75% de las significancias son mayores que 0.0028.

En general donde son pocos los valores de c que generan una mayor cantidad de falsas alarmas respecto a normalidad, y por lo tanto no se tiene el problema que aquí interesa, es en:

La Prueba 1, lado inferior, tiene sentido sólo para valores de $c>9$.

Prueba 2, lado inferior: si $c<1$ no existe la zona A. Si $c>1$, $\bar{\alpha} = 0.007724$, $S = 0.01123$.

La Prueba 3, del lado superior, con $c>1$, tiene un $\bar{\alpha} = 0.002768$, $S=0.001658$ y el 75% de las significancias son menores que 0.00372.

La prueba 4 del lado superior tiene significancias bajas. En este lado de la carta en lugar de que sean 9 puntos consecutivos, bastan 8.

La Prueba 7 es una buena aproximación a la normal.

TABLA 1

Significancias de pruebas estándar aplicadas a la carta c.
(m) mínima, (M) máxima, CI y CS cuartil inferior y superior.
*** La prueba no aplica (no existe la zona).*

Prueba	Lado de la carta	Significancia bajo normalidad	Significancia carta c, c<1	Carta modificada, c<1	Significancia carta c, c>1	Carta modificada, c>1
1	Superior	0.00135	$\bar{\alpha}$ = 0.017589 S = 0.015602 M = 0.08744 CI= 0.0076 CS= 0.021193	$\bar{\alpha}$ = 0.000405 S = 0.000758 M = 0.00410 CS= 0.00032	$\bar{\alpha}$ = 0.003916 S = 0.001803 M = 0.01425 CI= 0.0028 CS= 0.00427	$\bar{\alpha}$ = 0.000899 S = 0.000455 M = 0.00256 CS= 0.00105
	Inferior	0.00135	**		$\bar{\alpha}$ = 0.000114 S = 0.000130 CS= 0.00020	
2	Superior	0.001075	$\bar{\alpha}$ = 0.007724 S = 0.011237 M = 0.05357 CS= 0.00965	$\bar{\alpha}$ = 0.003357 S = 0.004531 M = 0.02 CS= 0.004	$\bar{\alpha}$ = 0.002311 S = 0.001092 M = 0.00661 CS= 0.00302	$\bar{\alpha}$ = 0.001861 S = 0.001078 M = 0.01 CS= 0.002
	Inferior	0.001075	**		$\bar{\alpha}$ = 0.000478 S = 0.000374 CI= 0.0002 CS= 0.00074	
3	Superior	0.0027	$\bar{\alpha}$ = 0.004855 S = 0.007198 M = 0.03193 CI= 0.0001 CS= 0.00666	$\bar{\alpha}$ = 0.003732 S = 0.004980 M = 0.01937 CI= 0.0001 CS= 0.00599	$\bar{\alpha}$ = 0.002768 S = 0.001658 M = 0.00882 CI= 0.0016 CS= 0.00372	$\bar{\alpha}$ = 0.002210 S = 0.001271 M = 0.00590 CI= 0.0013 CS= 0.00305
	Inferior	0.0027	**		$\bar{\alpha}$ = 0.003519 S = 0.003543 M = 0.02875 CI= 0.0016 CS= 0.00443	
4	Superior	0.003822	$\bar{\alpha}$ = 0.001318 S = 0.002734 M = 0.01 CS= 0.001	$\bar{\alpha}$ = 0.001598 S = 0.003372 M = 0.01580 CS= 0.00100	$\bar{\alpha}$ = 0.001641 S = 0.001165 M = 0.00 CS= 0.002	$\bar{\alpha}$ = 0.001730 S = 0.001231 M = 0.00531 CS= 0.00252
4	Inferior	0.003822	$\bar{\alpha}$ = 0.146990 S = 0.228397 M = 1.00 CS= 0.185		$\bar{\alpha}$ = 0.003778 S = 0.004559 M = 0.03 CS= 0.004	
7	Toda la carta	0.000103	$\bar{\alpha}$ = 0.000008 S = 0.000018 CS= 0.00001	$\bar{\alpha}$ = 0.000003 S = 0.000006 CS= 0.00000	$\bar{\alpha}$ = 0.000163 S = 0.000182 CS= 0.00020	$\bar{\alpha}$ = 0.000077 S = 0.000100 CS= 0.00009
8	Toda la carta	0.003254	$\bar{\alpha}$ = 0.166208 S = 0.192526 M = 0.99253 CS= 0.26569	$\bar{\alpha}$ = 0.207958 S = 0.204168 M = 0.99253 CS= 0.33712	$\bar{\alpha}$ = 0.011481 S = 0.074722 M = 0.75009 CS= 0.00500	$\bar{\alpha}$ = 0.015592 S = 0.075740 M = 0.75009 CS= 0.00823

3. ALTERNATIVA

Para corregir el problema anterior, y mejorar la aproximación a la significancia bajo normalidad, se propone que el límite de control superior de la carta c se calcule de la siguiente manera

$$LCS^* = EM[\bar{c} + 3\sqrt{\bar{c}}\,]$$

donde $EM[x]$ es la función entero mayor o igual que x. Así para calcular el LCS^* lo que se hace es calcular el LCS típico, y se le aplica a éste la función mayor entero o igual; por ejemplo si LCS=13.63, entonces $LCS^* = EM\,[13.63] = 14$. Si el límite superior típico para una carta c es un entero, entonces coincide con el LCS^*. La línea central se calcula de la forma tradicional, al igual que el límite de control inferior.

Las zonas de la carta de control modificada, en el lado inferior se obtienen de manera tradicional, y en el lado superior tendrán una amplitud igual a S^*, que se obtiene al dividir entre tres la distancia entre la línea central y el LCS^*, es decir

$$S^* = \frac{LCS^* - \bar{c}}{3}$$

En Camacho y Gutiérrez (1995a) ya se mostró que una modificación similar en los cálculos de los límites de control de las cartas p y np dan buenos resultados, por lo que se espera que ahora en la carta c, ocurra lo mismo.

Para evaluar esta modificación se hizo un estudio de significancia de seis de las ocho pruebas estándar aplicadas a la carta c modificada. El estudio fue el similar que el que se explicó en la sección anterior.

4. CONCLUSIÓN

Es riesgoso aplicar las ocho pruebas a las cartas c, ya que para ciertos valores de c, se tendrá mayor número de falsas alarmas que bajo normalidad que es como han sido diseñadas las pruebas. Los mayores riesgos se presentan cuando se aplica la prueba 1, 2 y 3 del lado superior, y la 8. Agudizándose el riesgos para valores pequeños de c. Por lo anterior se concluye que la aplicación indiscriminada de las pruebas estándar a la carta c, producirá una mayor cantidad de falsas alarmas que en las cartas para procesos con distribución normal. Esto está en contradicción con las recomendaciones de la literatura clásica de control de calidad (ver sección 1). Con la modificación propuesta el problema se resuelve para las pruebas 1, 2 y 3.

REFERENCIAS

Besterfield, D.H. (1990). *Quality control*, 3e. Prentice-Hall, Englewood, New Jersey.

Camacho Castillo, O. y H. Gutiérrez Pulido (1995 a). Estudio de la significancia de las pruebas para detectar causas especiales de variación en las cartas p y np. *Revista de Estadística*, vol. VII, # 9.

Camacho Castillo, O. y H. Gutiérrez Pulido (1995 b). Modificaciones en la construcción de las cartas de control p y np para reducir la frecuencia de falsas alarmas. *Memoria del X Foro Nacional de Estadística.*

Gutiérrez Pulido, H. (1992). *Control Total de Calidad.* Edug, Guad.

Nelson, L. S. (1984). The Shewhart control chart-tests for special causes, *Journal of Quality Technology*, **16**,4, 237-39.

ReVelle, J.B. and H.J. Harrington (1992). Statistical process control en quality engineering handbook. Editado por T. Pyzdek y R.W. Berger. ASQC Quality Press.

Western Electric (1958). *Statistical quality control handbook.* AT&T, Chicago.

Pronósticos Bayesianos con Restricciones en Modelos ARMA. II

ENRIQUE DE ALBA y OMAR AGUILAR CHÁVEZ

ITAM, México *I.S.D.S.*

1. INTRODUCCIÓN

La literatura sobre métodos Bayesianos aplicados al análisis de modelos de series de tiempo tipo ARMA es bastante limitada y generalmente se concentra en modelos AR. En la mayoría de los casos las aplicaciones se limitan a modelos muy simples, o bien sólo pronostican uno o dos periodos. En parte esto se ha debido a la imposibilidad o al costo de resolver los problemas computacionales involucrados. Este artículo presenta un procedimiento Bayesiano para obtener pronósticos de una serie trimestral que sean consistentes con una cifra anual futura dada. La serie de tiempo sigue un modelo ARMA(p,q) y no existen restricciones ni en p ni en q. Los pronósticos restringidos se obtienen de la distribución predictiva condicionando en la restricción dada, de una manera semejante a como se hace en de Alba (1993) para modelos autorregresivos. En de Alba y Aguilar (1995a) se obtienen resultados similares, pero utilizando una aproximación. En estas referencias se puede encontrar una amplia bibliografía. Para facilitar la exposición se supone que en el modelo se utilizan datos trimestrales y que la información futura conocida está disponible en forma anual. Sin embargo los resultados son válidos para situaciones más generales.

Considérese el siguiente modelo ARMA(p,q)

$$\phi(B)W_t = \theta(B)\varepsilon_t \text{ donde } \varepsilon_t \sim N\left(0,\sigma^2\right) \forall t \in Z$$

$\phi(B) = 1-\phi_1 B-\phi_2 B^2-\phi_3 B^3-\ldots-\phi_p B^p$ $\quad \theta(B) = 1-\theta_1 B-\theta_2 B^2-\theta_3 B^3-\ldots-\theta_q B^q$. Se supone $\varepsilon_p = \varepsilon_{p-1} = \varepsilon_{p-2} = \ldots = \varepsilon_{1-q} = 0$ y conocidas las primeras p observaciones, por lo que se tiene

$$\varepsilon_t = W_t - \sum_{i=1}^{p} \phi_i W_{t-1} + \sum_{j=1}^{q} \theta_j \varepsilon_{t-j} \qquad \text{con} \qquad t = p+1, p+2, \ldots, N.$$

La verosimilitud condicional es

$$L\left(\underline{\phi},\underline{\theta},\sigma^2 | S_N\right) \propto \left(\sigma^2\right)^{-\frac{N-p}{2}} \exp\left\{-\frac{1}{2\sigma^2}\sum_{t=p+1}^{N}\varepsilon_t^2\right\}.$$

Si se escribe $W_t = \underline{z}'_t\,\underline{\phi} + \varepsilon_t + \underline{\varepsilon}'_t\,\underline{\theta}$ con $\underline{z}'_t = \left(W_{t-1}, W_{t-2}, \ldots, W_{t-p}\right)$ y $\underline{\varepsilon}'_t = \left(\varepsilon_{t-1}, \varepsilon_{t-2}, \ldots, \varepsilon_{t-q}\right)$; o en notación matricial $\mathbf{W}=\mathbf{X}\underline{\phi}+\mathbf{G}_{\underline{\varepsilon}}$ con $\mathbf{W}=\left(W_{p+1}, W_{p+2}, \ldots, W_N\right)$, $\mathbf{X}' = \left(\underline{z}_{p+1}, \underline{z}_{p+2}, \ldots, \underline{z}_N\right)$, $\underline{\varepsilon} = \left(\varepsilon_{p+1}, \varepsilon_{p+2}, \ldots, \varepsilon_N\right)$, y

$$G=\begin{bmatrix} 1 & 0 & 0 & 0 & \dots & 0 \\ \theta_1 & 1 & 0 & 0 & \dots & 0 \\ \theta_2 & \theta_1 & 1 & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & 0 & \dots & 1 \end{bmatrix}_{(T\times T)}$$

donde $\underline{\varepsilon} \sim N(0, \sigma^2 \boldsymbol{I}_T)$, entonces $\mathbf{W} \sim N_T(\mathbf{X}\underline{\phi}, \sigma^2\mathbf{GG}')$ y la verosimilitud es

$$L(\underline{\phi}, \underline{\theta}, \sigma|\underline{W}) \propto (\sigma)^{-T} \ |\mathbf{V}|^{1/2} \ \exp\left\{-\frac{1}{2\sigma^2}\left(\mathbf{W}-\mathbf{X}\underline{\phi}\right)' V\left(\mathbf{W}-\mathbf{X}\underline{\phi}\right)\right\},$$

con $\mathbf{V}=(\mathbf{GG}')^{-1}$, y $\underline{W}$ representa todos los datos.

2. PRONÓSTICOS SIN RESTRICCIONES

Se utiliza una inicial no informativa para $\underline{\phi}$ **V** y σ (independientes), es decir, $f(\underline{\phi}) \propto$ const, $f(\sigma) \propto \sigma^{-1}$ y $f(\mathbf{V}) \propto$ const. La distribución posterior es, entonces

$$f\left(\underline{\phi}, \underline{\theta}, \sigma|\underline{W}\right) \propto (\sigma)^{-(T+1)}|\mathbf{V}|^{1/2} \exp\left\{-\frac{1}{2\sigma^2}\left(\underline{\phi}-\underline{\hat{\phi}}\right)' \Sigma_G^{-1}\left(\underline{\phi}-\underline{\hat{\phi}}\right)+\nu_G S_G^2\right\},$$

donde $\underline{\hat{\phi}} = (\mathbf{X'VX})^{-1}(\mathbf{X'VW})$, $\Sigma_G^{-1}=(\mathbf{X'VX})$ y $\nu_G S_G^2 = (\mathbf{W} - \mathbf{X}\underline{\hat{\phi}})'\mathbf{V}(\mathbf{W}-\mathbf{X}\underline{\hat{\phi}})$.

Con muestreo de Gibbs se pueden generar observaciones de la posterior de $\underline{\phi}$, σ y **V**. A su vez los valores de $\underline{\theta}$ se pueden obtener de la matriz **V**, ya que $\mathbf{V} = (\mathbf{GG}')^{-1}$ y con la descomposición de Cholesky en $\mathbf{V}^{-1}$ se obtienen G y el vector $\underline{\theta}$.

Si los valores futuros de **W** se escriben de la siguiente forma

$$W_{T+k} - \phi_1 W_{T+k-1} - \phi_2 W_{T+k-2} - \dots - \phi_p W_{T-(p-k)} = \varepsilon_{T+k} - \theta_1 \varepsilon_{T+k-1} - \theta_2 \varepsilon_{T+k-2} - \dots - \theta_q \varepsilon_{T-(q-k)},$$

se puede definir

$\underline{W}_p = \left(W_{T-(p-1)}, \ W_{T-(p-2)}, \dots, W_T\right)'$ las últimas p observaciones

$\underline{\hat{\varepsilon}}_q = \left(\hat{\varepsilon}_{T-(q-1)}, \ \hat{\varepsilon}_{T-(q-2)}, \dots, \ \hat{\varepsilon}_T\right)'$ los últimos q errores, y

$\underline{\varepsilon}_f = \left(\varepsilon_{N+1}, \ \varepsilon_{n+2}, \ \dots, \ \varepsilon_{N+k}\right)'$ los errores de pronóstico,

donde $\underline{\varepsilon}_f \sim N_k\left(\underline{0}, \sigma^2 I_k\right)$ y $\underline{\varepsilon}_q \sim N_q\left(\underline{0}, \sigma^2 I_q\right)$. Si también se definen

$$A_{k\times k}=\begin{bmatrix} 1 & 0 & 0 & \dots & 0 \\ -\phi_1 & 1 & 0 & \dots & 0 \\ -\phi_2 & -\phi_1 & 1 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ -\phi_{k-1} & -\phi_{k-2} & -\phi_{k-3} & \dots & 1 \end{bmatrix}, \qquad C_{k\times k}=\begin{bmatrix} 1 & 0 & 0 & \dots & 0 \\ -\theta_1 & 1 & 0 & \dots & 0 \\ -\theta_2 & -\theta_1 & 1 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ -\theta_{k-1} & -\theta_{k-2} & -\theta_{k-3} & \dots & 1 \end{bmatrix},$$

$$B_{k\times p}=\begin{bmatrix} -\phi_p & -\phi_{p-1} & -\phi_{p-2} & \cdots & -\phi_1 \\ 0 & -\phi_p & -\phi_{p-1} & \cdots & -\phi_2 \\ 0 & 0 & -\phi_p & \cdots & -\phi_3 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & \cdots & -\phi_k \end{bmatrix} \quad y \quad D_{k\times q}=\begin{bmatrix} -\theta_q & -\theta_{q-1} & -\theta_{q-2} & \cdots & -\theta_1 \\ 0 & -\theta_q & -\theta_{q-1} & \cdots & -\theta_2 \\ 0 & 0 & -\theta_q & \cdots & -\theta_3 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & \cdots & -\theta_k \end{bmatrix};$$

y el vector de pronósticos $\underline{W_f} = (W_{T+1}, W_{T+2}, W_{T+3}, \ldots, W_{T+k})'$, entonces la distribución predictiva de $\underline{W_f}$ será

$$f(\underline{W}_f|\ \underline{W}) = \int \ldots \int f(\underline{W}_f|\ \underline{\phi},\underline{\theta},\sigma,\underline{W}) f(\underline{\phi},\underline{\theta},\sigma|\ \underline{W})\, d\underline{\phi}\ d\underline{\theta}\ d\sigma.$$

Se pueden escribir los pronósticos de la siguiente manera

$$\underline{W}_f = -A^{-1}B\underline{W}_p + A^{-1}C\underline{\varepsilon}_f + A^{-1}D\underline{\hat{\varepsilon}}_q,$$

lo cual tiene las siguientes ventajas:

1) Es fácil obtener cualquier número de pronósticos con Monte Carlo y Gibbs,

2) Es fácil obtener la distribución predictiva de los pronósticos, condicional en los parámetros, y

3) Utilizando pérdida cuadrática, la media condicional produce pronósticos con restricciones.

De esta expresión, dadas observaciones hasta T, $\underline{W}$, y todos los parámetros: $\underline{W}_f|\ \underline{\phi},\underline{\theta},\sigma,\underline{W} \sim N_k(\underline{\mu}_f, \Sigma_f)$, con $\underline{\mu}_f = -A^{-1}B\underline{W}_p + A^{-1}D\underline{\hat{\varepsilon}}_q$ y $\Sigma_f = \sigma^2(A^{-1}C)(A^{-1}C)'$. Los pronósticos sin restricciones se obtienen entonces de

$$E(\underline{W}_f|\ \underline{W}) = \int \ldots \int \underline{\mu}_f f(\underline{\phi},\underline{\theta},\sigma|\ \underline{W}) d\underline{\phi}\, d\underline{\theta}\, d\sigma,$$

y la varianza:

$$\mathrm{Var}(\underline{W}_f|\ \underline{W}) = E_{\underline{\phi},\underline{\theta},\sigma}\left(Var(\underline{W}_f|\ \underline{W},\underline{\phi},\underline{\theta},\sigma)\right) + Var_{\underline{\phi},\underline{\theta},\sigma}\left(E(\underline{W}_f|\ \underline{W},\underline{\phi},\underline{\theta},\sigma)\right),$$

donde

$$E_{\underline{\phi},\underline{\theta},\sigma}\left(Var(\underline{W}_f|\ \underline{W},\underline{\phi},\underline{\theta},\sigma)\right) = \int \ldots \int \Sigma_f\, f(\underline{\phi},\underline{\theta},\sigma|\ \underline{W}) d\underline{\phi}\, d\underline{\theta}\, d\sigma$$

Si lo que interesa es $W_c = \underline{i}'\underline{W}_f$, donde $\underline{i}$ es un vector de constantes, se tiene que W_c es Normal con $E(W_c|\ \underline{W},\underline{\phi},\underline{\theta},\sigma) = \underline{i}'\underline{\mu}_f$ y $\mathrm{Var}(W_c|\ \underline{W},\underline{\phi},\sigma) = \underline{i}'\Sigma_f\underline{i}$. Si, por ejemplo, se quiere que z sea la suma entonces $z = \underline{i}'\underline{W}_f$, *con* $\underline{i}' = (1,1,1,1)$, $f(z|\ \underline{W},\underline{\phi},\underline{\theta},\sigma) = N(\underline{i}'\underline{\mu}_f\ \underline{i}'\Sigma_f\underline{i})$ y la densidad predictiva para z es

$$f(z|\ \underline{W}) = \int f(z|\ \underline{W},\underline{\phi},\underline{\theta},\sigma) f(\underline{\phi},\underline{\theta},\sigma|\ \underline{W}) d\underline{\phi}\, d\underline{\theta}\, d\sigma.$$

Estas últimas cuatro ecuaciones se pueden evaluar numéricamente generando muestras a partir de $f(\underline{\phi},\underline{\theta},\sigma|\ \underline{W})$ por muestreo de Gibbs, las cuales se utilizan para llevar a cabo integración por el método de Monte Carlo, ver de Alba y Aguilar (1995a,b).

3. PRONÓSTICOS BAYESIANOS CON RESTRICCIONES

Si se obtienen los pronósticos de $\underline{W}_f$ dados W_c y los datos **W**,

$$f(\underline{W}_f|\ \underline{W}_c,\underline{W}) = \int \cdots \int f(\underline{W}_f|\ W_c,\underline{\phi},\underline{\theta},\sigma,\underline{W}) f(\underline{\phi},\underline{\theta},\sigma|\ W_c,\underline{W}) d\underline{\phi}\, d\underline{\theta}\, d\sigma.$$

Suponiendo independencia entre W_c y $\underline{W}$, $f(\underline{\phi},\underline{\theta},\sigma|\ W_c,\underline{W}) = f(\underline{\phi},\underline{\theta},\sigma|\ \underline{W})$. Para encontrar $f(\underline{W}_f|\ W_c,\underline{\phi},\underline{\theta},\sigma,\underline{W})$, sea

$$\underline{W}^* = \begin{pmatrix} \underline{W}_f \\ W_c \end{pmatrix} = \begin{pmatrix} I_k \\ \underline{i} \end{pmatrix} \underline{W}_f .$$

Entonces $\underline{W}^*|\ \underline{\phi},\underline{\theta},\sigma,\mathbf{W} \sim N_{k+1}(\underline{\mu}^*,\Sigma^*)$, una distribución normal multivariada singular. Si se define

$$\underset{(k+1)\times(k+1)}{\Sigma^*} = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}$$

con $\underset{1\times 1}{\Sigma} = \Sigma_c = \underline{i}'\Sigma_f \underline{i}$, $\underset{1\times k}{\Sigma_{12}} = \Sigma_{21}', \underset{k\times k}{\Sigma_{22}} = \Sigma_f$ y $\underset{k\times 1}{\Sigma_{21}} = \mathrm{Cov}(\underline{W}_f, W_c) = \mathrm{Var}(\underline{W}_f)\underline{i} = \Sigma_f \underline{i}$; entonces $\underline{W}_f|\ W_c,\underline{\phi},\underline{\theta},\sigma,\underline{W} \sim N_k(\underline{\mu}_R,\Sigma_R), \underline{\mu}_R = \underline{\mu}_f + \Sigma_f \underline{i}(\underline{i}'\Sigma_f \underline{i})^{-1}(W_c - \mu_c)$ y $\Sigma_R = \Sigma_f - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}$. Con pérdida cuadrática

$$E(\underline{W}_f|\ W_c,\underline{W}) = \int \cdots \int \underline{\mu}_R f(\underline{\phi},\underline{\theta},\sigma|\ \underline{W}) d\underline{\phi}\, d\underline{\theta}\, d\sigma$$

y $\mathrm{Var}(\underline{W}_f|\ W_c,\underline{W}) = E_{\underline{\phi},\underline{\theta},\sigma}\left(Var(\underline{W}_f|\ W_c,\underline{\phi},\underline{\theta},\sigma,\underline{W})\right) + \mathrm{Var}_{\underline{\phi},\underline{\theta},\sigma}\left(E(\underline{W}_f|\ W_c,\underline{\phi},\underline{\theta},\sigma,\underline{W})\right)$, donde $E_{\underline{\phi},\underline{\theta},\sigma}$ y $\mathrm{Var}_{\underline{\phi},\underline{\theta},\sigma}$ representan esperanza y varianza bajo la posterior de $\underline{\phi},\underline{\theta},\sigma$. De la Sección 2 se tiene que la distribución posterior conjunta es

$$f(\underline{\phi},\underline{\theta},\sigma|\ \underline{W}) \propto (\sigma)^{-(T+1)}|\mathbf{V}|^{1/2} \exp\left\{-\frac{1}{2\sigma^2}(\underline{\phi}-\hat{\underline{\phi}})'\Sigma_G^{-1}(\underline{\phi}-\hat{\underline{\phi}}) + v_G S_G^2\right\}.$$

Las distribuciones condicionales son

$\underline{\phi}|\ \underline{\theta},\sigma,\underline{W} \sim N_p(\hat{\underline{\phi}},\sigma^2\Sigma_G)$;

$\sigma|\ \underline{\phi},\underline{\theta},\underline{W} \sim$ Gamma-Inversa(v_1,S_1^2) con $v_1 = T$ y $S_1^2 = (W - X\underline{\phi})'\mathbf{V}(\mathbf{W} - \mathbf{X}\underline{\phi})$,;

$\mathbf{V}|\ \underline{\phi},\sigma,\underline{W} \sim$ Wishart (Σ,v_2,q), donde $\Sigma = \sigma^2\left[(W - X\underline{\phi})(W - X\underline{\phi}')\right]^{-1}$,

donde $v_2 = q + 2$ y q es el orden de la componente MA del modelo; $\underline{\theta}$ se obtiene de $\mathbf{V}^{-1}$ con la descomposición de Cholesky y $\mathbf{V} = (\mathbf{GG}')^{-1}$. Con estas definiciones es factible obtener muestras de la distribución conjunta de $\underline{\phi}, \underline{\theta}, \sigma$ mediante el muestreo de Gibbs (de Alba y Aguilar, 1995b), y de ahí los pronósticos tanto sin restricciones como con ellas, de acuerdo con las fórmulas anteriores. Para algunos ejemplos del uso de estos resultados véase esta última referencia.

REFERENCIAS

de Alba, E. (1993), Constrained Forecasting in Autoregressive Time Series Models: Bayesian Analysis, *International Journal of Forecasting* **9**, 95-108.

de Alba, E. y O. Aguilar Chávez (1995a), Pronósticos Bayesianos con Restricciones en Modelos ARMA, *Revista de Estadística, INEGI-AME*, en prensa.

de Alba, E. y O. Aguilar Chávez (1995b), Constrained Bayesian Forecasting in ARMA Models, enviado para publicación en: *International Journal of Forecasting*.

El Mercado de Vivienda en México: Un Modelo de Comportamiento

MA. DE LOURDES DE LA FUENTE D.

ITAM, México

1. INTRODUCCIÓN

Los propósitos de este estudio son: proporcionar un marco de referencia sobre los costos directos de las diferentes etapas que intervienen en el proceso de edificación de vivienda, especificar y estimar un modelo econométrico de costos y realizar el análisis del mercado de vivienda con base en los resultados obtenidos en la estimación del modelo. Se incluye la consideración de los factores de demanda y de oferta, factores institucionales y tecnológicos. Finalmente se hacen algunas recomendaciones de política de vivienda en relación a la evolución de los costos directos de edificación de vivienda.

2. ESPECIFICACIÓN DE LA DEMANDA DE VIVIENDA

El análisis de la demanda de vivienda, como el de la demanda de otros bienes durables, considera al consumidor como demandante del flujo de servicios que le proporciona ese activo. De esta manera al decidir el nivel de consumo anual de estos servicios, el consumidor toma en cuenta el costo anual de habitar una vivienda. Al igual que, quien renta, el dueño de una vivienda debe tomar en cuenta la renta potencial de su vivienda o Renta Implícita, que es el costo de oportunidad de habitar una vivienda propia y se mide por los ingresos que deja de obtener si decide habitarla en lugar de rentarla.

La renta de mercado o renta implícita es la renta anual que un dueño de vivienda cobraría a un inquilino por el derecho de ocuparla durante ese periodo. La relación entre el precio de una vivienda y su renta implícita es el concepto más importante en el mercado de vivienda. El valor de mercado de una vivienda es el valor presente neto de las rentas anuales de la vivienda menos sus costos de operación durante la vida útil de la misma. Si la vida útil es de un número grande de años (la aproximación es buena a partir de 30 años de vida útil, como es el caso de las viviendas) esta relación se puede aproximar de la siguiente manera:

$$PVIV = RENT/COSTCAP \qquad (1)$$

donde: *PVIV*: Valor de la vivienda, *RENT*: Renta anual de la vivienda y *COSTCAP*: costo del capital, definido como:

$$COSTCAP = INT + TAX + DEP - GAN + ESF \qquad (2)$$

donde: *INT*: tasa de interés del mercado, *TAX*: tasa de impuesto predial, *DEP*: tasa de depreciación de la vivienda, *GAN*: tasa de ganancias de capital (por el cambio en el valor real de la vivienda)[1], *ESF*: costo unitario de operación del dueño en la actividad de rentar la casa (es la retribución a la actividad empresarial de quien renta una vivienda). Otra determinante fundamental de la cantidad demandada de servicios de la vivienda es el ingreso del consumidor. El "Ingreso" pertinente para estimar esta demanda es difícil de medir. El

[1]La razón del porque está componente del costo entre con signo negativo en la ecuación es que una mayor apreciación de la edificación reduce los costos.

Ingreso anual o "corriente" de un consumidor, no es la medida adecuada pues fluctúa año con año y por estar sujeto a variaciones no esperadas (como por caer en un periodo de desempleo o por recibir una lotería o herencia inesperada). No sería adecuado cambiar el consumo de los servicios de vivienda debido a estos cambios temporales. Los consumidores basan estas decisiones en una medida de ingreso promedio esperado de largo-plazo conocido como Ingreso Permanente. Para los fines de estudio se medirá el ingreso "permanente" a través de la tendencia estimada del ingreso corriente en el periodo considerado. Este procedimiento elimina las fluctuaciones irregulares y deja únicamente las de la tendencia del ingreso y es, por lo tanto, una mejor variable para estimar la respuesta de la demanda de vivienda a la "riqueza" de los individuos.

En base a la información anterior se plantea la siguiente especificación de la demanda de servicios de vivienda:

$$\log VIV = \beta_o + \beta_1 \log R + \beta_2 \log YPER + u \qquad (3)$$

donde: *VIV*: es la construcción de vivienda nueva[2], *R*: es la renta implícita o costo de oportunidad de habitar una vivienda (ya sea propia o rentada), *YPER*: es el ingreso permanente estimado[3], *u*: Es el error aleatorio que satisface los supuestos clásicos del modelo de regresión lineal, β_o, β_1 *y* β_2: son los coeficientes a estimar. β_o representa la demanda de servicios de vivienda autónoma, β_1 mide el efecto que tiene un cambio en el costo de oportunidad de habitar una vivienda sobre la demanda de vivienda, es la elasticidad precio de la demanda y β_2 mide la elasticidad ingreso de la demanda de vivienda.

3. ESPECIFICACIÓN DE LA OFERTA DE VIVIENDA

La principal característica de la oferta de vivienda es que el acervo existente en un momento dado del tiempo es "quasi-fijo" -relativamente constante- en el muy corto plazo. Este acervo se modifica gradualmente a través de nuevas construcciones, demolición de edificaciones de vivienda a otros usos y viceversa.

En un cierto periodo, y dados los costos de edificación de vivienda, se esperarían mayores adiciones al acervo de viviendas -mayor construcción- a mayor precio o "valor" de las mismas. Es importante hacer notar que, a diferencia del consumidor, el oferente de vivienda toma en cuenta este precio de las viviendas, y no su renta anual. De esta manera, los promotores inmobiliarios toman en consideración el precio al que pueden vender las edificaciones y recuperar -con su respectiva utilidad- los costos de construcción.

Dentro de las variables que inciden sobre los costos de construcción de vivienda, están los costos de los insumos de la construcción. Con un estado del conocimiento y la tecnología dados, los principales determinantes del costo de construir una vivienda pueden agruparse en dos grandes apartados: -los costos de los materiales de construcción y -los costos de la mano de obra requerida en la construcción.[4]

De esta manera tenemos que:

$$\log(INCEVIS) = \gamma_0 + \gamma_1 \log(VISMC) + \gamma_2 \log(VISMO) \qquad (4)$$

---

[2]Aproximada por el Producto Interno Bruto de la Cosntrucción.

[3] Estimado como la tendencia del ingreso a partir de un polinomio de grado cuatro

[4]Los costos de los terrenos se consideran constantes para los fines de este trabajo. Los impuestos -como el predial- forman parte del costo del capital. Ver la Ecuación (2).

donde: *INCEVIS*: Índice Nacional del Costo de Edificación de Vivienda, *VISMC:* Componente del Costo de los Materiales de Construcción, *VISMO:* Componente del Costo de la Mano de Obra utilizada en la Construcción, el parámetro γ_0 decrece en el tiempo a una tasa λ equivalente a la tasa de crecimiento de la productividad de la industria de la construcción y γ_1 *y* γ_2: son las participaciones de los costos de materiales y de mano de obra en el costo total. Econométricamente, la relación entre el valor de la vivienda, los costos de edificación y los determinantes de ese costo puede plantearse de la siguiente forma:

$$PVIV = f(\text{Costos de edificación de la vivienda}) \tag{5}$$

donde: f es una función que -para simplificar el análisis- se supone como una constante de proporcionalidad, que mide el "markup" sobre costos, y representa la utilidad de los promotores.[5]

La especificación final de la oferta de vivienda será entonces:

$$log(VIV) = \rho_0 + \rho_1 log(PVIV) + \lambda\, CAMBTECN + w \tag{6}$$

donde: *VIV:* es la construcción de nuevas viviendas, *PVIV:* es el precio relativo de las viviendas, en relación al nivel de precios general en la economía y *CAMBTECN:* es una variable de tendencia que permite la estimación de λ que representa la tasa de crecimiento de la productividad factorial total en la edificación de vivienda.

ρ_0, ρ_1 y λ: son los coeficientes a estimar. ρ_1 mide la elasticidad precio de la oferta de vivienda y w es el error aleatorio que satisface los supuestos clásicos del modelo.

4. EQUILIBRIO EN EL MERCADO DE VIVIENDA

En el caso del mercado de vivienda, el equilibrio no puede determinarse de la manera usual. Nótese que los conceptos de "precio" son diferentes para los oferentes y para los demandantes.

Según la ecuación (1), los demandantes toman en consideración la renta de mercado (en el caso de ser inquilinos) o la renta implícita (en el caso de ser dueños que habitan su propia vivienda). En cambio, según la ecuación (6), los oferentes de vivienda -los promotores inmobiliarios- toman en cuenta el precio o valor total de las viviendas.

Sin embargo, como *RENT=COSTCAP*PVIV* (ver la ecuación (1)), para cada nivel de *RENT*, la curva de la oferta puede ser graficada sobre la curva de demanda, y viceversa.

5. ESTIMACIÓN DE LAS ECUACIONES DE OFERTA Y DEMANDA COMO SISTEMA SIMULTÁNEO

$$log(VIV) = \beta_0 + \beta_1\, logR + \beta_2 log\, y + u \tag{7}$$

$$log(VIV) = \rho_0 + \rho_1\, log(PVIV) + \lambda\, CAMBTECN + w \tag{8}$$

El sistema formado por las ecuaciones (7) y (8) se estimó simultáneamente por medio del método de mínimos cuadrados en 2 etapas con datos anuales para el periodo 1973-1993.

[5]La utilidad de los promotores se supone constante en este análisis.

6. RESULTADOS DE LA ESTIMACIÓN DEL SISTEMA SIMULTÁNEO DE DEMANDA Y OFERTA DE VIVIENDA

$\log(VIV) = 0.197 + 0.781 \log(YPER) - 0.113 \log(R) - 0.175\ D \qquad R^2 = .834$

(1.37) (0.088) (0.029) (0.051)* F = 28.4

(0.44) (8.87) (-3.96) (-3.4) DW = 1.81

$\log(VIV) = 12.09 + 1.43 \log(PVIV)_{t-1} + 0.045\ CAMBTEC - 0.277\ D88^{**} \qquad R^2 = .704$

(.111) (0.539) (0.013) (0.127) F = 8.91

(109.24) (2.65) (3.45) (-2.18) DW = 1.76

La elasticidad ingreso de la demanda de vivienda es 0.78 La elasticidad precio de la demanda de vivienda es -0.11

En el análisis se incluyó una variable "dummy" (D) correspondiente al periodo de la crisis financiera (1982-1987) que resultó altamente significativo y que indica que durante esos años la tendencia de la demanda se redujo 17.5%

La elasticidad precio de la oferta de vivienda es 1.43. La tasa de cambio tecnológico estimada para la industria de la industria de la edificación de vivienda es 4.4%. En la ecuación se incluyó una variable "dummy" correspondiente al año de 1988 (D88), con la intención de captar el impacto del inicio del Pacto (en ese entonces Pacto de Solidaridad Económica) que resultó significativa e indica que la construcción de vivienda cayó durante ese periodo.

Ambos ajustes son razonables -con una R cuadrada del 83% y del 70% respectivamente.

7. SIMULACIÓN DE CAMBIOS EN LOS COSTOS DE EDIFICACIÓN SOBRE EL MERCADO DE VIVIENDA

Cor los parámetros calculados para la oferta y la demanda de vivienda, se procede a la simulación -para el periodo 1994 a 2,000- de los efectos de un cambio en las variables más pertinentes sobre el mercado de vivienda. Las simulaciones que se presentan a continuación son ilustrativas de posibles efectos de las condiciones en el mercado de vivienda en México y pretenden captar tres tipos de efectos:

Efectos Macroeconómicos. Estos se captan a través de variables de escala, como lo es el Producto Interno Bruto del País (y por consecuencia el ingreso de las familias), y de las tasas de interés (que afectan el costo de capital).

Efectos Institucionales. Una mejora administrativa y reducciones en el costo de la vivienda significan mejores condiciones en el mercado de vivienda. Esto se refleja también en el costo de capital considerado -a través de sus componentes de impuestos y de esfuerzo administrativo en el mercado de vivienda.

Efectos de los Precios de Insumos de la Construcción. Las variaciones en los precios de los materiales de construcción y del costo de la mano de obra utilizada se reflejarán en el mercado de vivienda vía los precios de la construcción.

Se formaron los escenarios que se describen a continuación:

7.1 Escenario Pesimista:

---

* Los valores entre paréntesis corresponden a las desviaciones estándar y a las estadísticas T.

** Ecuación corregida por autocorrelación

1) Supone que el ingreso permanente en México crece en términos reales por debajo de su tendencia histórica. 2% anual. 2) El costo del capital -que es alto en términos históricos en la actualidad, permanece constante (en su elevado nivel actual). 3) El precio de los materiales de construcción se incrementa por encima del nivel de inflación de manera que su crecimiento real es del 1%. 4) El precio de la mano de obra utilizada en la construcción aumenta en términos reales a una tasa del 2% (es decir, los salarios en la rama aumentan más que el índice de precios general).

7.2 Escenario Normal:

1) Supone que el ingreso permanente en México crece a aproximadamente su tendencia histórica 3% anual. 2) El costo del capital decrece moderadamente a una tasa del -1% anual. 3) El precio de los materiales de construcción permanece constante. 4) El precio de la mano de obra utilizada en la construcción aumenta en términos reales a una tasa moderada del 1% (es decir, los salarios en la rama aumentan 1% más que el índice de precios general).

7.3 Escenario Optimista:

1) Supone que el ingreso permanente en México crece a tasas altas 3.5% anual. 2) El costo del capital decrece rápidamente.-1.5% anual. 3) El precio de los materiales de construcción decrece rápidamente a una tasa del -1% anual. 4) El precio de la mano de obra utilizada en la construcción aumenta en términos reales a una tasa moderada del -2% (es decir, los salarios en la rama aumentan 1% más que el índice de precios general).
Los Resultados de los Escenarios fueron los siguientes:

7.4 Escenario pesimista

La construcción de vivienda aumenta 38.9% durante el periodo, es decir, 2.03% anual en promedio.
El precio de las viviendas aumenta 51.8% en el periodo, es decir, 2.59% anual en promedio en términos reales.

7.5 Escenario normal

La construcción de vivienda aumenta 81% durante el periodo, es decir, 3.68% anual en promedio.
El precio de las viviendas aumenta 14.9% en el periodo, es decir, 0.87% anual en promedio en términos reales.

7.6 Escenario optimista

La construcción de vivienda aumenta 115.5% durante el periodo, es decir, 4.76% anual en promedio.
El precio de las viviendas decrece 29.76% en el periodo, es decir, baja 2.19% anual en promedio en términos reales.

8. CONCLUSIONES Y RECOMENDACIONES DE POLÍTICA

Los resultados de este estudio sugieren las siguientes recomendaciones:

Un contexto macroeconómico estable y en crecimiento es de importancia significativa para mejorar las perspectivas del mercado de vivienda a través de su efecto sobre el ingreso de las familias.

Reducciones en las tasas de interés, en los costos fiscales de mantener vivienda reducen el costo del capital y con ello estimulan la demanda de vivienda.

Esfuerzos de simplificación administrativa mejoran significativamente las condiciones en el mercado de vivienda -al bajar los costos de operación-.

Mantener bajos los precios de los insumos utilizados en la construcción -y en todo caso al mismo nivel que los precios en general- incrementa la oferta de vivienda y permite que aumente la cantidad de viviendas construidas en equilibrio y que baje su precio Finalmente cabe mencionar que modelos estadísticos del mercado de vivienda resultan útiles para predecir posibles tendencias en el mercado y para planear adecuadamente las políticas y medidas para que funcione adecuadamente en beneficio de las familias mexicanas.

REFERENCIAS

Banco de México. *Indicadores Económicos*. Banco de México: México, varios números.

DeLeeuw, (1971). The Demand for Housing: A Review of the Cross-Section Evidence. *Review of Economics and Statistics*. **53.**

Dougherty, Ann, and Robert Van Order. (1982). Inflation, Housing Costs, and the Consumer *Prince Index. American Economic Review* Vo. **72**: 154-64

Ellwood, David, and A. Mitchell Polinsky. (1979). An Empirical Reconciliation of Micro and Grouped Estimates of the Demand for Housing. *Review of Economics and Statistics* Vo. **61**: 199-205.

Estudio de Frapor *Impacto Macroeconómico de la Vivienda de Interés Social en los principales Insumos de la Construcción*. Sedesol

Gujarati D. (1988). *Econometría Básica* Mc. Graw Hill.

Johnston J. (1984). *Econometric Methods*. Mc. Graw Hill, Third Edition.

Judge. G., C. Hill, W. Griffiths & T. Chao. (1980). *Theory and Practice of Econometrics*. Wiley: New York.

Macro: Asesoría Económica Realidad Económica de México. Iberoamérica: México, varios números.

Mills. E. (1972). *Urban Economics*. Scott Foresman: Glenview.

Polinsky, Mitchell. (1997). The Demand for Housing: A Study in Specification and Grouping. *Econometrica* **45** 447-462.

Pindyck R.S. & Rubinfeld, D.C. (1991). *Econometric Models and Economic Forecasts* Mc. Graw Hill.

Poterba, J. (1980). Inflation, Income Taxes and Owner-Ocupied Housing. NBER, *Working* Paper No. **55** September.

Schwab, Robert. (1983). Real and Nominal Interest Rates and the Demand for Housing. *Journal of Urban Economics* Vol. **13**: 181-95.

Secretaría de Programación y Presupuesto. Sistemas de Cuentas Nacionales: Cuentas de Producción. SPP: México, varios números.

Carta de Control $\overline{X}$ Basada en Muestreo de Grupos Ordenados

ROMAN DE LA VARA S.

CIMAT, México

1. INTRODUCCIÓN

Desde la propuesta original de Shewhart se han desarrollado muchos tipos de cartas y diversos métodos para construirlas, para poder cubrir la gran diversidad de procesos que pueden tenerse en la práctica (ver Montgomery, 1991). Todo este desarrollo e innovaciones de la herramienta se ha dado considerando que los *subgrupos racionales* se obtienen utilizando *muestreo aleatorio simple* (MAS).

Una situación que puede surgir en la práctica son procesos que tienen las dos propiedades siguientes: 1) resulta muy caro obtener cada medición de la característica de interés, en términos de recursos, tiempo y dinero; 2) es posible que una muestra pequeña de unidades pueda ordenarse en relación a la característica de interés sin medirla, ya sea visualmente o utilizando una variable auxiliar fácil de medir, o cualquier otro método mucho más económico. En este tipo de procesos es conveniente utilizar control estadístico basado en *muestreo de grupos ordenados* (MGO), en lugar de muestreo aleatorio simple. En este artículo mostramos como construir la carta de control $\overline{X}$ basada en MGO, y su desempeño se compara con el de carta $\overline{X}$ tradicional basada en MAS.

2. CARTA DE CONTROL $\overline{X}$ BASADA EN MGO

2.1 Muestreo de Grupos Ordenados (MGO)

El muestreo de grupos ordenados consiste en obtener mr muestras aleatorias de tamaño m que pueden ser ordenadas fácilmente para después medir de cada muestra ordenada sólo una estadística de orden: en la primera muestra se mide sólo la unidad con el rango más pequeño, en la segunda muestra se mide la unidad que tiene el segundo rango más pequeño, y así sucesivamente hasta la m-ésima muestra en la cual se mide la unidad que corresponde al rango mayor. Este ciclo se repite r veces lo que da por resultado un total de mr mediciones, que conforman la muestra de grupos ordenados (ver Patil, Sinha y Taillie, 1994).

Para ilustrar el método, consideremos un tamaño de grupo $m = 3$ con $r = 4$ ciclos. Las mediciones que finalmente se harán se marcan con círculo en la Figura 1.

Cada renglón de este diagrama es una muestra ordenada de tamaño $m = 3$ y cada ciclo es un instante de tiempo en el que se obtiene una media basada en las tres unidades marcadas con ⊙. Así pues, se obtienen 36 unidades seleccionadas aleatoriamente en cuatro ciclos, pero sólo $n = mr = 12$ de ellas serán medidas para conformar la muestra de grupos ordenados.

Sean $X_{i1}, X_{i2}, \ldots, X_{im}$, donde $i = 1,2,\ldots, m$, variables aleatorias independientes que tienen la misma distribución $F(x)$. Sean $X_{i(1)}, X_{i(2)},\ldots, X_{i(m)}$ los estadísticos de orden correspondientes a $X_{i1}, X_{i2}, \ldots, X_{im}$ $(i = 1,2,\ldots,m)$. Entonces la muestra ordenada basada en un ciclo está dada

por $X_{1(1)}$, $X_{2(2)}$, ..., $X_{m(m)}$. Denotemos por $X_{(i:m)j}$ al i-ésimo estadístico de orden de la i-ésima muestra en el j-ésimo ciclo. Entonces la media estimada en el ciclo j está dada por

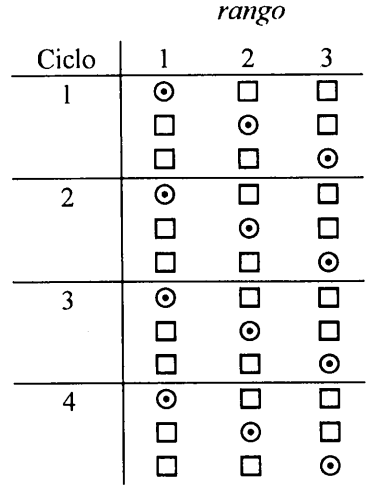


Fig. 1

$$\overline{X}_{(m)j} = \frac{1}{m}\sum_{i=1}^{m} X_{(i:m)j}$$

y es un estimador insesgado de la media del proceso, μ. La varianza de $\overline{X}_{(m)j}$ está dada por

$$\text{var}\left(\overline{X}_{(m)j}\right) = \frac{1}{m^2}\sum_{i=1}^{m} \sigma^2_{(i:m)}$$

donde $\sigma^2_{(i:m)}$ es la varianza de la i-ésima estadística de orden.

2.2 Carta $\overline{X}$ Basada en MGO

En cada instante de tiempo se toma una muestra aleatoria simple de tamaño m^2 de unidades producidas. Sea la unidad X_{ij}, con $i, j = 1,..., m$, tomada de un proceso estable con distribución $N\left(\mu,\sigma^2\right)$. Esta muestra se parte aleatoriamente en m grupos de m unidades, y cada uno se ordena visualmente o utilizando una variable concomitante fácil de medir. Se recomienda utilizar $m = 2$ ó $m = 3$ precisamente para facilitar la ordenación de las unidades en cada grupo. En este artículo suponemos que las unidades se pueden ordenar sin error.

Considerando una corrida de r ciclos, un estimador de la varianza $\sigma^2_{(i:m)}$ es de la forma

$$\hat{\sigma}^2_{(i:m)} = \frac{1}{r-1}\sum_{j=1}^{r}\left(X_{(i:m)j} - \overline{X}_{(i:m)}\right)^2$$

y un estimador de var $\left(\overline{X}_{(m)}\right)$ está dado por

$$\text{var}\left(\overline{X}_{(m)}\right) = \frac{1}{m^2(r-1)} \sum_{i=1}^{m} \sum_{j=1}^{r} \left(X_{(i:m)j} - \overline{X}_{(i:m)}\right)^2$$

donde $\overline{X}_{(i:m)} = \frac{1}{r}\sum_{j=1}^{r} X_{(i:m)j}$. De lo anterior se llega a que los parámetros de la carta $\overline{X}$ basada en la muestra de grupos ordenados son:

$$LSC = \overline{\overline{X}}_{(m)} + 3\sqrt{\frac{1}{m^2(r-1)} \sum_{i=1}^{m} \sum_{j=1}^{r} \left(X_{(i:m)j} - \overline{X}_{(i:m)}\right)^2}$$

$$LC = \overline{\overline{X}}_{(m)}$$

$$LIC = \overline{\overline{X}}_{(m)} - 3\sqrt{\frac{1}{m^2(r-1)} \sum_{i=1}^{m} \sum_{j=1}^{r} \left(X_{(i:m)j} - \overline{X}_{(i:m)}\right)^2}$$

3. COMPARACIÓN CON LA CARTA $\overline{X}$ TRADICIONAL

Sin pérdida de generalidad en el caso normal, consideremos un proceso con distribución normal estándar. Primero se simula una carta $\overline{X}$ tradicional para este proceso considerando un tamaño de muestra n = 3 y una corrida de 50 puntos. Se elige n = 3 para poderla comparar con la carta basada en MGO con $n = m = 3$ y r = 50 ciclos o puntos. El resultado se observa en las Figuras 2 y 3. A simple vista se observa una menor dispersión de la medias estimadas utilizando MGO, lo que habla de un estimador más eficiente de la media μ del proceso.

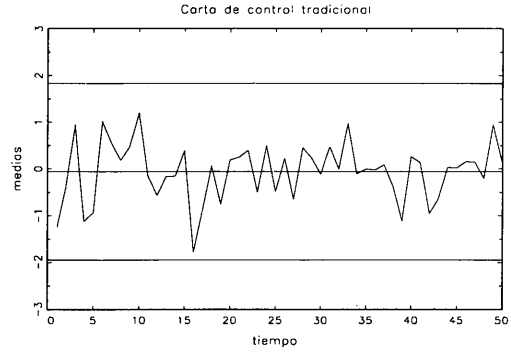


Fig. 2 Carta $\overline{X}$ tradicional, $n = 3$

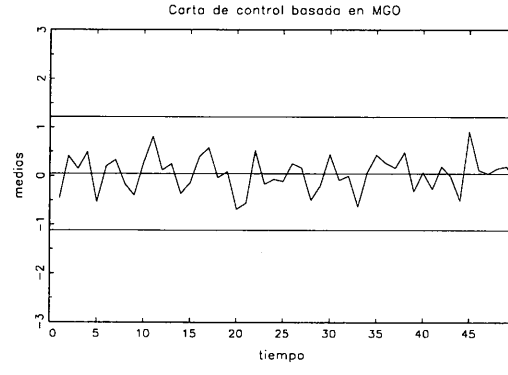


Fig. 3 Carta $\overline{X}$ basada en MGO, $n = m = 3$

3.1 Estudio de ARL's

La comparación visual de las Figuras 2 y 3 no es suficiente, por lo que se hace un estudio de simulación para comparar el desempeño de la carta basada en MGO con respecto a la carta tradicional basada en MAS, ambas con el mismo número de mediciones.

Una estadística muy utilizada para comparar el desempeño de cartas de control es el *ARL* (longitud promedio de corrida), que es el número promedio de puntos que se requieren para que la carta detecte una señal de falta de control. Usualmente el estudio se hace considerando un proceso en control estadístico (desliz $\delta = 0$), e induciendo deslices $\delta = 0.2,0.4,...,3.0$ de la media del proceso, medidos éstos en unidades de desviaciones estándares de las medias muestrales. Esto es, considerando un proceso con distribución conocida $N(\mu,\sigma^2)$, el desliz de la media del proceso está dado por $\delta = \sqrt{n} \mid \mu - \mu_0 \mid / \sigma$. Por ejemplo, un desliz de $\delta = 0.2$ equivale a que la media del proceso se desplazó una distancia de $\mid \mu - \mu_0 \mid = 0.2\sigma / \sqrt{n}$. Cabe recordar que si $\delta = 0$, el proceso se encuentra en control estadístico, y todas las señales de fuera de control registradas en este caso son falsas alarmas.

En este trabajo sólo estudiamos el desempeño de las cartas en relación a la primera regla: un punto fuera de los límites de control. En la Tabla 1 se muestran los ARL's obtenidos por simulación para la carta basada en MAS y para la carta basada en MGO, considerando tamaños de muestra $n = 2$ y 3. Se observa que los ARL's de la carta basada en MGO son bastante más pequeños para todos los deslices de la media del proceso.

TABLA 1. ARL.

δ	n = 2 MGO	n = 3 MGO	n = 2, 3 MAS
0.0	346.00	343.70	370.40
0.2	266.21	252.42	308.43
0.4	155.90	130.04	200.08
0.6	84.63	65.39	119.67
0.8	46.96	33.99	71.55
1.0	26.92	18.93	43.89
1.2	16.41	11.14	27.82
1.4	10.45	6.98	18.25
1.6	6.96	4.65	12.38
1.8	4.87	3.28	8.69
2.0	3.56	2.45	6.30
2.5	1.95	1.47	3.24
3.0	1.35	1.14	2.00

Aunque la carta de control $\overline{X}$ basada en MGO detecta un poco más de falsas alarmas, ésta parece tener el potencial para aplicarse en procesos donde se busca hacer un mínimo de mediciones, sin perder la potencia de detección.

REFERENCIAS

Montgomery D. C. (1991). *Introduction to Statistical Quality Control.* Second Edition. Wiley, New York.

Patil, G. P., Sinha, A. K. and Taillie, C. (1994). Ranked set sampling. In Handbook of Statistics, Vol. **12**: *Environmental Statistics*, G. P. Patil and C. R. Rao, eds., *North-Holland Elsevier*, New York, 167-200.

Aproximaciones a la Verosimilitud Perfil en el Caso de Muestras Finitas

ELOISA DIAZ FRANCÉS y ENRIQUE VILLA D.

CIMAT, México

1. INTRODUCCIÓN

Consideramos el caso donde nos interesa hacer afirmaciones de estimación sobre uno de los parámetros de localización (φ_1) de una mezcla finita de distribuciones Gumbel con parámetros de localización desconocidos y parámetros de escala conocidos, iguales a 1. Este incluye el caso de una mezcla de distribuciones exponenciales con parámetros de escala desconocidos, ya que la transformación logaritmo de una variable aleatoria exponencial con parámetro de escala θ dá una variable aleatoria con distribución Gumbel con parámetros (log (θ), 1). La mezcla de c densidades Gumbel puede expresarse como

$$f(x) = \sum_{i=1}^{c} p_i \exp\{(x-\varphi_i) - \exp(x-\varphi_i)\},$$

donde $\varphi_1,\ldots, \varphi_c$ son los parámetros de localización de cada componente Gumbel en la mezcla y $p_1,\ldots, p_c$ son las proporciones de mezcla; así, $\sum_{i=1}^{c} p_i = 1$. El espacio paramétrico es multidimensional, $\Phi = (\varphi_1,\ldots, \varphi_c, p_1,\ldots, p_{c-1})$. Sin embargo, en el caso aqui considerado solo hay un parámetro de interés, φ_1, y los restantes $(2c-2)$ parámetros se toman como parámetros de estorbo.

Si se requieren afirmaciones de estimación por intervalo sobre φ_1, el procedimiento usual es considerar la distribución asintótica del estimador de máxima verosimilitud de φ_1 para construir la siguiente cantidad pivotal que converge en distribución a una variable aleatoria ε que tiene una distribución normal estándar conforme crece el tamaño muestral. Esto es,

$$u_{\varphi 1} = u(\varphi_1, X) = (\varphi_1 - \hat{\varphi}_1)\sqrt{I_{\varphi_1}} \xrightarrow{d} \varepsilon \sim N(0, 1).$$

donde u_{φ_1} es una cantidad pivotal (i.e. una variable aleatoria que es función de parámetros pero que tiene una distribución conocida e independiente de parámetros desconocidos), X es la muestra observada, $\hat{\varphi}_1$ es el estimador de máxima verosimilitud del parámetro de interés φ_1 y I_{φ_1} es la información observada. Entonces las afirmaciones de estimación para φ_1 en su forma más simple serían

$$\varphi_1 = \hat{\varphi}_1 \pm \sqrt{I_{\varphi_1}}\,\varepsilon, \quad \text{donde } \varepsilon \sim N(0,1). \tag{1}$$

Sin embargo, puede ocurrir que para muestras finitas de mezclas de distribuciones Gumbel el pivotal u_{φ_1} no tiene distribución normal; aún más, la distribución asociada podría ser asimétrica. En tales casos, que ocurren frecuentemente, aún para muestras de tamaño 1000, usualmente consideradas "grandes", Las afirmaciones de estimación (1) podrían ser completamente equivocadas.

El procedimiento aquí sugerido es calcular la verosimilitud perfil o maximizada del parámetro de interés, considerando los parámetros restantes en la mezcla como parámetros de estorbo. El objetivo es obtener una cantidad pivotal lineal aproximada u_{φ_1}, con una densidad log F adecuada, con aproximación suficiente a la verosimilitud perfil observada del parámetro de interés. El procedimiento es similar al descrito en Viveros y Sprott (1987). Esta aproximación pivotal puede llevar a inferencias sobre φ_1 simples y eficientes de la forma

$$\varphi_1 = \hat{\varphi}_1 - \sqrt{I_{\varphi_1}}\varepsilon, \quad \text{donde} \quad \varepsilon \sim \log F(m,n), \tag{2}$$

donde m y n son los grados de libertad que se deben determinar en base a la muestra observada. Eficiencia significa aquí que los intervalos deducidos de (2) corresponden a los de la muestra o verosimilitud perfil observada de φ_1. Adicionalmente, estos intervalos son intervalos de verosimilitud, en el sentido que valores de φ_1 dentro del intervalo son más plausibles, dada la muestra observada, que otros valores fuera del intervalo. Además los intervalos de confianza obtenidos en (2) serán exactos, i.e. ellos presentan frecuencias de cobertura correctas. La propiedad de exactitud *per se* no es suficiente, ya que un procedimiento de estimación puede ser exacto y sin embargo incluir valores del parámetro que son no plausibles o irrelevantes de acuerdo a la muestra observada. (Ver Chamberlin y Sprott, 1991, para una descripción general y discusión de conjuntos de verosimilitud-confianza, además de fidelidad, eficiencia y exactitud de afirmaciones de estimación).

2. PIVOTALES LINEALES APROXIMADOS Y VEROSIMILITUD PERFIL

En el esquema de una mezcla de densidades Gumbel aqui tratado, en donde solo es de interés un parámetro de localización φ_1 de una de las componentes de la mezcla, la función de verosimilitud perfil de φ_1, es

$$L_p(\varphi_1, X) = L(\varphi_1, \varphi_2^*..., \varphi_c^*, p_1^*, ..., p_{c-1}^*, X),$$

donde $L(.)$ es la función de verosimilitud de la mezcla y $\varphi_2^*..., \varphi_c^*, p_1^*, ..., p_{c-1}^*$ son las estimaciones de maxima verosimilitud dado φ_1. Aunque una expresión explicita para esta verosimilitud maximizada como función de φ_1 no puede encontrarse facilmente en el caso de una mezcla de Gumbels, la verosimilitud perfil para φ_1 se puede calcular via el algoritmo iterativo EM propuesto por Dempster, Laird y Rubin (1977) para una colección de puntos φ_1

El objetivo aquí es encontrar cantidades pivotales lineales en φ_1, que reproducen bien la función de verosimilitud perfil de φ_1 observada. El uso de este pivotal aproximado puede llevar a inferencias simples, eficientes y exactas para φ_1 que toman en cuenta la información sobre el parámetro de interés, contenida en la muestra y en los parámetros de estorbo. Por evidencia empírica, la función de verosimilitud perfil observada de φ_1 puede ser asimétrica aún para muestras consideradas como grandes, razón por la que es importante buscar una cantidad pivotal lineal aproximada cuya distribución se encuentre en la familia log F y así poder hacer inferencias sobre φ_1 de la forma 2. Se comparan tres métodos para obtener pivotales lineales log F para el parámetro de localización de interés.

El método analítico esta propuesto en Viveros y Sprott (1987), y consiste en considerar la aproximación de Taylor de la función log-verosimilitud perfil relativa,

$$r(\varphi_1, X) = \log\left(R(\varphi_1 X)\right) = \log\left(\frac{L_P(\varphi_1, X)}{L_P(\hat{\varphi}_1, X)}\right) = \log\left(\frac{L_P(\varphi_1, X)}{L(\hat{\varphi}_1, \hat{\varphi}_2 \ldots, \hat{\varphi}_c, \hat{p}_1, \ldots, \hat{p}_{c-1}, X)}\right),$$

en el estimador de máxima verosimilitud $\hat{\varphi}_1$, conservando los cuatro primeros términos de la aproximación.

El método de aproximación por mínimos cuadrados de la log-verosimilitud relativa observada, por la log-verosimilitud de una distribución log F, consiste en estimar los parámetros de la distribución por el método de mínimos cuadrados ponderados.

El método de aproximación minimizando la distancia dirigida de Kullback-Leibler entre la verosimilitud perfil relativa de φ_1, reescalada para ser una función de densidad, y la correspondiente densidad log F, vista como una función de sus parámetros m and n. Estos dos últimos métodos se explican ampliamente en Diaz-Francés y Villa (1995).

3. COMPARACIÓN DE MÉTODOS

En Figura 1, se tiene la verosimilitud perfil relativa observada de φ_1 y las correspondientes aproximaciones log F obtenidas con los tres métodos mencionados, para una muestra de tamaño 50, de una mezcla de Gumbels, con parámetros $\varphi_1 = \log(10)$; y $\varphi_2 = \log(200)$, y proporciones de mezcla, $p_1 = p_2 = 0.5$, con estos parámetros la muestra tiene un comportamiento de muestra grande, ya que coinciden la verosimilitud perfil relativa de φ_1 y la aproximación normal. Esta Figura muestra que todas las aproximaciones estan cercanas a la verosimilitud perfil relativa observada, esto es, las aproximaciones log F asi como tambien la normal asintótica son eficientes, ya que reproducen bien la información contenida en la función de verosimilitud perfil relativa. En este caso todas las aproximaciones dan intervalos de verosimilitud-confianza, y además las afirmaciones de estimación serán eficientes y exactas.

En la Figura 2 se presentan las gráficas ya mencionadas para una muestra también de tamaño 50, pero con parámetros $\varphi_1 = \log(10)$; y $\varphi_2 = \log(75)$, y proporciones de mezcla, $p_1 = p_2 = 0.5$. Con estos parámetros la muestra tiene un comportamiento de muestra pequeña, ya que difieren la verosimilitud perfil relativa observada y la aproximación normal. Sin embargo, las aproximaciones log F en este caso muestran una buena aproximación.

4. CONCLUSIONES

De los tres métodos propuestos para obtener aproximaciones log F de la verosimilitud perfil, el de la minima distancia de Kullback-Leibler y el de mínimos cuadrados ponderados resultan igualmente recomendables por ser relativamente fáciles de implementar y ofrecer buenos resultados inferenciales. El método analítico resulta sumamente complicado cuando la dimensión del espacio paramétrico aumenta.

Fig. 1

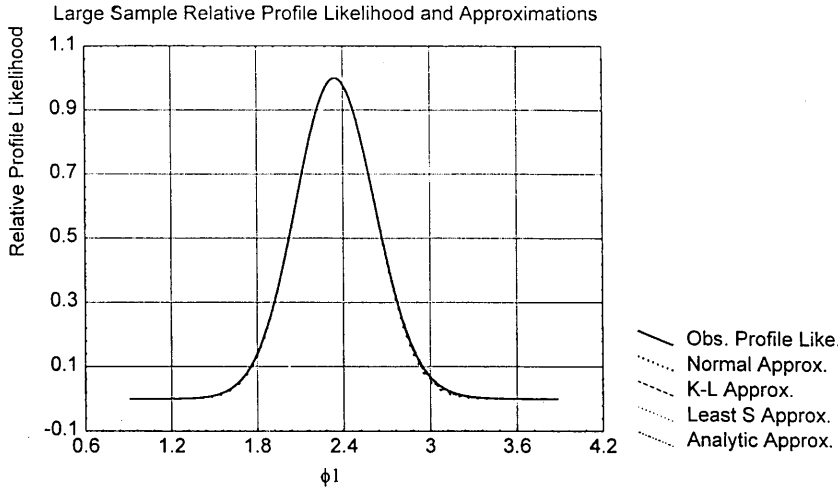
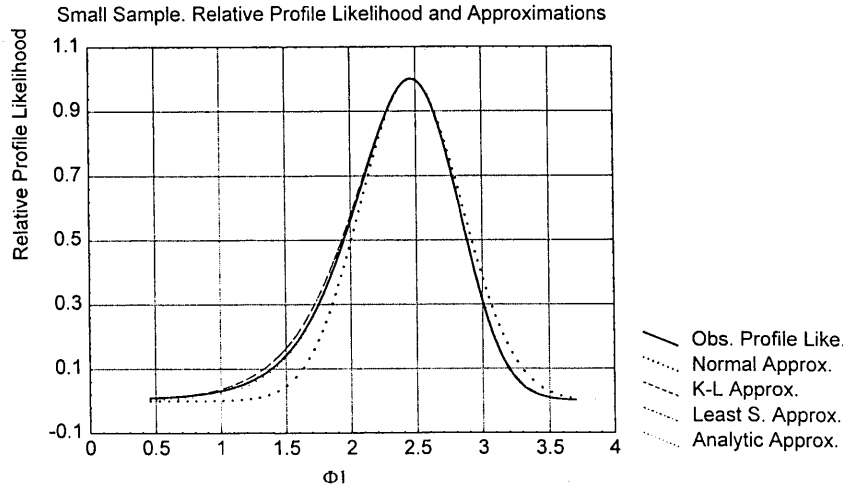


Fig. 2



REFERENCIAS

Chamberlin, S. R. and Sprott, D. A. (1991). Inferential Estimation, Likelihood, and Maximum Likelihood Linear Estimating Functions'. In Godambe, V. P. (Editor); Estimating Functions; Oxford: *Oxford University Press*, pp. 255-266.

Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977). Maximum Likelihood from Incomplete Data Via the EM Algorithm. *J. Royal Statist. Soc.*, Series B, V. **39**, 1-38.

Diaz-Francés, E. y Villa, E. (1985). Approximations to the profile Likelihood of a Location Parameter of a Mixture of Gumbel Distributions, *Comunicaciones del CIMAT*.

Viveros, R. and Sprott, D. A. (1987). Allowance for Skewness in Maximum Likelihood Estimation with Application to the Location-Scale Model, *Canad. J. Statist.*, V. **15**, 349-361.

Apuntes Sobre la Modelización de Series Diarias de Actividad Económica

ANTONI ESPASA y J. MANUEL REVUELTA

Universidad Carlos III de Madrid, España

1. INTRODUCCIÓN

Estamos acostumbrados a ver como el análisis de series temporales se centra en series de nivel de agregación mensual o superior. Son varias las razones por las que series de mayor frecuencia no han sido todavía fruto de estudios más profundos. Fundamentalmente, esto ha sido debido a las múltiples complicaciones añadidas que trae el reducir el nivel de agregación y que pueden hacer dudar en un principio de la bondad de los ajustes conseguidos mediante técnicas de series temporales. Por ejemplo, el comportamiento de series diarias de actividad económica relacionadas con el consumo de energía eléctrica, ventas en grandes superficies, ocupación de medios de transporte, niveles de contaminación, etc., viene caracterizado, entre otros, por aspectos tales como el solapamiento de estacionalidades de distintos periodos (fundamentalmente, semanal, mensual y anual), por su sensibilidad al efecto calendario o por una compleja dependencia de variables exógenas como la temperatura. Todo ello da una idea de la dificultad del problema.

Otro factor que ha actuado como freno para el estudio de estas series ha sido la falta de datos. Por un lado, ha sido práctica habitual considerar sólo el agregado por su mayor facilidad de almacenamiento. Además, muchas de las series de interés están relacionadas con la actividad privada de empresas, las cuales son, en general, remisas a hacer pública esta información y prefieren financiar sus propios estudios de forma que se preserve la confidencialidad. Todo esto ha hecho que se genere un cierto ocultismo que en nada favorece el desarrollo de técnicas adecuadas.

Por otro lado, es fácil imaginarse las grandes necesidades de cálculo que requiere el análisis de un volumen tan grande de datos. Esto exige prestaciones computacionales que sólo han estado disponibles de forma generalizada en estos últimos años.

Particularizando a España, dos de los modelos con mayor trascendencia a la vista de sus muy buenos comportamientos, han sido los construidos para la predicción de la circulación fiduciaria, patrocinado por el Banco de España, y para la predicción del consumo peninsular de energía eléctrica, encargado por REE. Ambos fueron realizados por A. Espasa y J.R. Cancelo y se emplean con éxito desde 1986 y 1988, respectivamente. Para hacernos una idea de la complejidad de estos modelos baste decir que dependen de más de 80 y 170 parámetros, respectivamente.

Son evidentes las ventajas que aportaría a una empresa o institución el dotarse de este tipo de modelos. Por un lado, la tarea de predicción puede convertirse en un trabajo rutinario, lo que evita una dependencia continua respecto a expertos. Además, éstos no pueden cuantificar, en general, conjuntamente todos los factores que es capaz de considerar el modelo. Por otro lado, múltiples parámetros que conforman la parametrización son, en si mismos, herramientas de control y gestión de enorme utilidad.

Esto justifica nuestro interés por este tipo de análisis, cuyas líneas generales describimos a continuación.

2. CARACTERÍSTICAS PRINCIPALES

Las series diarias de actividad económica muestran, con carácter general, varias de las siguientes características:

(1) Tendencias:
Tendencias u oscilaciones locales de nivel.
(2) Oscilaciones Estacionales:
(a) Semanal, (b) anual, (c) efectos de principio y fin de mes.
(3) Efectos Calendario:
Oscilaciones, cambios de tendencia y de periodo estacional debidas a la presencia de fiestas y periodos vacacionales.
(4) Dependencia de Variables Exógenas:

Relación compleja respecto a variables explicativas exógenas y en especial de las variables meteorológicas.

A éstas hay que añadir otras específicas a cada caso en estudio, lo que hace necesario siempre un estudio particular. El proceso se complica cuando la presencia de alguna de ellas cancela o enmascara el efecto de otras.

En las figuras 1 y 2 se ven ejemplos de este tipo de series, en los que se pueden apreciar con una simple inspección alguna de las características enumeradas anteriormente. En los siguientes gráficos se analiza la importancia del *efecto calendario* en una serie de demanda eléctrica para la comunidad autónoma de Andalucía. Posteriormente analizamos esta cuestión con mayor profundidad.

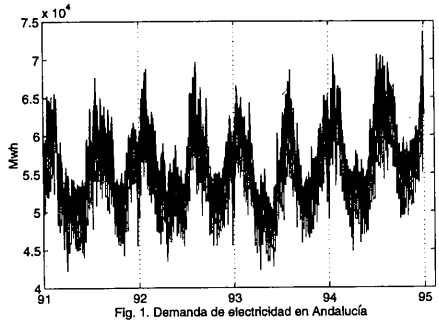


Fig. 1. Demanda de electricidad en Andalucía

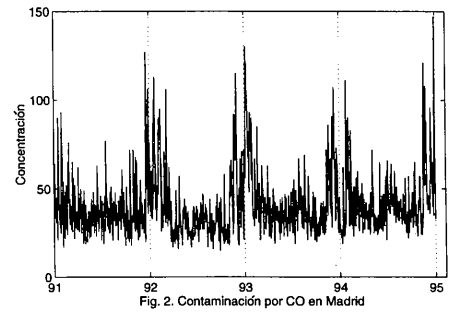


Fig. 2. Contaminación por CO en Madrid

Lo anterior sugiere enfocar el problema de la modelización tomando como base una lista de posibles características que puedan estar presentes en los datos, y a partir de ahí realizar una diagnosis que muestre cuáles son las que realmente aparecen en nuestra aplicación particular. Esta caracterización es conveniente realizarla por sectores para conseguir una homogeneización del estudio.

Una vez establecida esa lista, el proceso de modelización puede ser el siguiente:

(a) Determinación de las ortogonalidades o cuasi-ortogonalidades de entre el listado de características.
(b) Clasificación de estas características según su nivel de influencia en los datos. Así establecemos una jerarquía en su tratamiento.

(c) Modelización mediante indicadores de aquellas de las características anteriores para las que esto sea posible.
(d) Selección de esquemas determinísticos o estocásticos adecuados para la modelización del resto de las características.

El desarrollo de un esquema para la modelización sistemática se vería potenciado si se pudiera disponer de características lo más ortogonales posible y así poder proceder a su estudio de forma independiente. Lo anterior no siempre es factible y en tal caso habrá que recurrir a agrupamientos que representen efectos conjuntos. Este es el aspecto al que se refiere el punto (a), para cuya implementación será importante la opinión de expertos.

También queda por definir un criterio sobre el cual basar nuestras decisiones a la hora de cuantificar los efectos de las distintas características (punto (b)) o de elegir entre distintas alternativas de modelizacion para éstas (puntos (c) y (d)). Un criterio que se puede adoptar es el de *reducción de la varianza residual* a la hora de primar unos esquemas frente a otros. Aquí cabe decir que no hay una completa base teórica para defender este método, aunque una justificación intuitiva es clara y la experiencia en el tratamiento de este tipo de series abala sus buenos resultados.

Para tener en cuenta todo lo anterior, el mecanismo de estimación que se propone incorpora un proceso de estimación en cascada. Se comienza con la modelización de la primera característica de la lista del punto (b), eliminando su efecto de los datos. Posteriormente se pasa a la modelización de la segunda. En un siguiente paso se estiman ambas características conjuntamente, eliminando su efecto y pasando entonces a la tercera característica. De nuevo a esto seguiría una estimación conjunta y la eliminación de su efecto. El proceso continuaría hasta que se llega a la modelización del último efecto y a la consecuente estimación conjunta. El proceso debe culminarse con un análisis adecuado de validación.

Normalmente, la jerarquización de la lista de características presentes en el proceso es similar a la ordenación expuesta cuando enumeramos las características relevantes. La característica dominante suele ser la tendencia, seguida de las distintas estacionalidades. Dentro de éstas, su importancia dependerá del tipo de serie con que tratemos.

Respecto a la elección del esquema de modelización de cada característica, lo primero que habrá que analizar es la posible existencia de una variable indicador explicativa, y muy especialmente alguna variable meteorológica. Es fácil darse cuenta de la importancia de la temperatura en el caso de series diarias o de la plubiosidad para series de ocupación de medios de transporte. Cuando no exista un indicador habrá que recurrir bien a esquemas determinísticos o a esquemas estocásticos según sea más adecuado.

El proceso anterior puede resultar complejo. No obstante, es frecuente la posibilidad de agrupar series según aspectos tales como el sector a que pertenezcan, lo que permite un ahorro en recursos importante al poder homogeneizar partes del estudio.

3. ALTERNATIVAS DE MODELIZACIÓN SIN INDICADORES

3.1. Tendencia y Oscilaciones Locales

3.1.1. Esquemas Deterministicos

El esquema básico determinístico para la modelización de la tendencia es:

$$x_t = a + b_t + u_t$$

Este esquema presenta muy poca flexibilidad por lo que prácticamente no se usa, prefiriendo en general esquemas estocásticos.

3.1.2. Esquemas Estocásticos

Para la modelización de la tendencia o de oscilaciones locales mediante esquemas estocásticos se recurre al operador en diferencias

$$\Delta = (1 - L)$$

representando L al operador retardo ($Lx_t=x_{t-1}$). En Espasa y Peña (1995) se caracteriza cada proceso con una expresión *I(d,m)*, representando d el número de diferencias tomadas en el modelo (Δ^d) y m el indicador de la presencia (m=1) o no (m=0) de media determinística no nula en la tranformación estacionaria. En función de esto, el comportamiento de la función de predicción de la serie vendrá marcado por un polinomio tendencial de orden ($d+m-1$). En la mayoría de las series económicas el comportamiento encontrado es cuasilineal, es decir, responde a esquemas *I(1,1)* ó *I(2,0)*.

3.2. Estacionalidad

3.2.1. Esquemas Deterministicos

Para la modelización de estacionalidades mediante esquemas deterministicos se recurre a variables artificiales. Suponiendo que pretendemos analizar una estacionalidad S_j de periodo P_j, en el caso más general se recurre a las variables artificiales $VA_j(h)$, con $h=1,...,P_j$, donde $VA_j(h)$ toma el valor uno en aquellas observaciones correspondientes al momento estacional h de la estacionalidad S_j, y cero en el resto.

Un problema que se plantea es la dificultad para definir el periodo de algunas estacionalidades como la mensual (meses de 28, 29, 30 ó 31 días) o anual (años de 365 ó 366 días). Soluciones a esto pueden encontrarse en Espasa (1993).

Otro problema es la imposibilidad de considerar una variable artificial para cada momento estacional cuando el periodo es largo como en el caso mensual o anual. La solución a esto pasa por el agrupamiento en la misma variable artificial de momentos estacionales de comportamientos homogéneos respecto a la estacionalidad tratada. Por ejemplo, para la estacionalidad mensual suele ser suficiente considerar tres variables artificiales que consideren el efecto principio y fin de mes.

Otro aspecto a considerar, sobre todo en estacionalidades de periodo corto como la semanal, es la posibilidad de que la estructura estacional cambie a lo largo del año. Esto requiere recurrir a un juego de variables artificiales distintas para cada régimen de comportamiento. Además, puede ser que variables exógenas, como la temperatura, también afecten a los coeficientes de las distintas variables artificiales.

Esto nos da idea de la complejidad que tendrá este tipo de tratamiento. A diferencia de lo que ocurría para la tendencia, en la que los esquemas deterministicos tienen escaso interés, su uso sí es frecuente en la modelización de estacionalidades.

3.2.2. Esquemas Estocásticos

La modelización estocástica de una estacionalidad de periodo P_j se implementa mediante la aplicación a la serie original del operador suma:

$$U_{Pj-1}(L) = \left(1 + L + L^2 + ... + L^{Pj-1}\right)$$

Cuando este operador se aplica a una serie que también requiere un operador diferencia, ambos pueden combinarse dando lugar al operador diferencia estacional:

$$(1 - L)U_{Pj-1}(L) = \left(1 - L^{Pj}\right).$$

El operador estacional se puede descomponer en función de armónicos correspondientes a sus raíces, cuya interpretación es la de distintos subciclos que en su conjunto componen el ciclo estacional S_j. Por eso, a veces es suficiente, sobre todo en estacionalidades largas como la anual, recurrir directamente a los armónicos que realmente estén presentes y no al agregado. Por otro lado, esto nos da idea también de por qué cuando una serie presenta varias estacionalidades, el uso de distintos esquemas estocásticos para su modelización puede llevarnos a sobrediferenciaciones, al darse coincidencia de armónicos. Por ejemplo, para cualquier armónico del operador semanal o mensual, existe un armónico del operador anual que prácticamente lo solapa.

4. MODELIZACIÓN DE VARIABLES EXÓGENAS

La importancia de las variables exógenas en las series diarias depende en gran medida del sector con el que estemos tratando. No obstante, por las características de las series que hemos descrito hasta ahora, es fácil darse cuenta de que las variables meteorológicas tienen, en general, un papel importante. Características como el ciclo anual de series eléctricas o de contaminación, pueden considerarse suficientemente explicadas recurriendo a la temperatura.

Su incorporación al modelo no es simple puesto que, en general, su efecto es no lineal. Una formulación que ha dado muy buenos resultados en la práctica es la planteada en Cancelo y Espasa (1995), donde se recurre a una *modelización por umbrales*. El analista define, a partir de una búsqueda empírica, una serie de umbrales en el rango de la variable exógena dentro de los cuales el comportamiento de la variable a modelizar se pueda linealizar. De esta forma conseguimos una función lineal por tramos de fácil manejo.

Un problema importante en este tipo de aplicaciones es la heterogeneidad de los patrones de dependencia respecto a las variables exógenas a lo largo del año. Esto puede hacer necesaria la definición de distintos regímenes en la dependencia según la época del año.

5. MODELIZACIÓN DEL EFECTO CALENDARIO

A medida que bajamos el nivel de agregación temporal de nuestra serie, ésta se hace más sensible a aspectos tales como la existencia de fiestas generales o locales, periodos vacacionales, huelgas, elecciones, etc. En el caso de series diarias, una modelización correcta de este efecto es fundamental como paso previo para cualquier tipo de análisis posterior.

La complejidad de su modelización es grande. En general, se recurre a variables artificiales de distinto tipo según la naturaleza del efecto, acompañadas de un filtro que marca su dependencia dinámica. De nuevo, dependiendo de aspectos tales como el día de la semana en que cae la fiesta, la época del año, la posición en el mes o la temperatura, puede resultar necesario recurrir a distintos juegos de variables artificiales y filtros dinámicos.

Para evitar que se dispare el número de parámetros de nuestro modelo, es conveniente agrupar los distintos tipos de efecto calendario en bloques de comportamiento homogéneo.

En los gráficos 3 a 7, aparecen ilustrados distintos aspectos de este epígrafe para el caso de la demanda de electricidad en Andalucía.

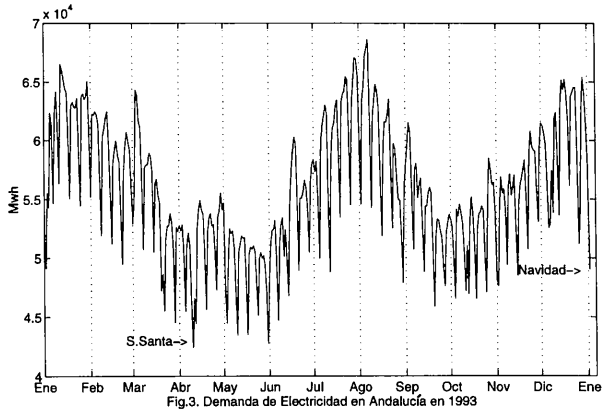


Fig.3. Demanda de Electricidad en Andalucía en 1993

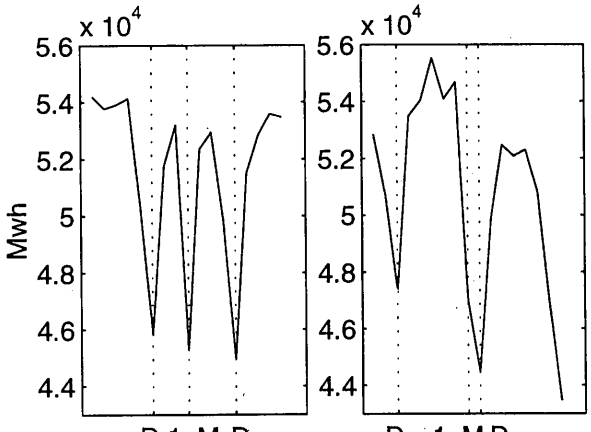


Fig.4. 1 Mayo de 1991 1 Mayo de 1993

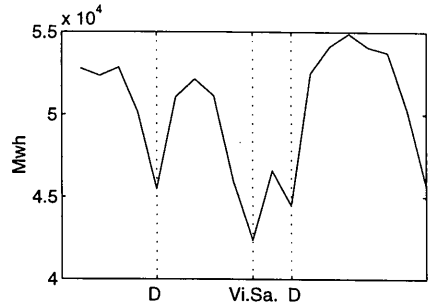


Fig.5. Semana Santa 1993

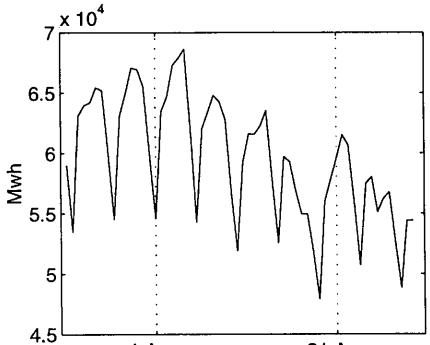


Fig. 6. Agosto 1993

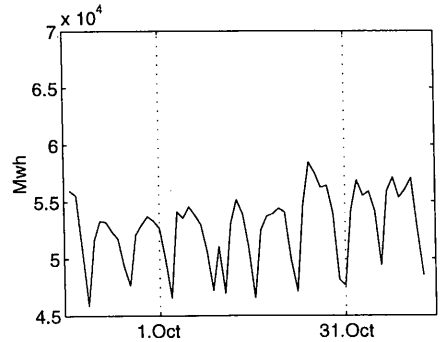


Fig. 7. Octubre 1993

6. CONCLUSIONES

En esta breve exposición hemos intentado, fundamentalmente, ilustrar la problemática asociada al tratamiento de las series diarias de actividad económica, describiendo los aspectos principales que las hace distintas de otras series de mayor nivel de agregación temporal, destacando la existencia simultánea de varias estacionalidades de distinto periodo. También es relevante su mayor sensibilidad al efecto calendario o su posible dependencia de variables exógenas.

Hemos analizado como una modelización adecuada puede requerir esquemas de una gran complejidad, incorporando un gran número de parámetros. Esto en general exige una estrategia de modelización disciplinadora.

REFERENCIAS

Box, G.E. y Jenkins, G.M. (1970). *Time Series Analysis: Forecasting and Control.* San Francisco, CA: Holden-Day.

Cancelo, J.R. and Espasa, A. (1991). Forecasting Daily Demand for Electricity with Multiple- -Input Nonlinear Transfer Function Models: A case Study. Working Paper **91-**05. *Universidad Carlos III de Madrid.*

Cancelo, J.R. and Espasa, A. (1995). Modelización del efecto temperatura en el consumo de electricidad: un ejercicio de búsqueda de especificación en relaciones dinámicas no lineales. *Estadística Española*, **37**: 183-200.

Cancelo, J.R. and Espasa, A. (1996). Using high-frecuency data and time series models to improve yield management. Working Paper. *Universidad Carlos III de Madrid.*

Espasa, A. (1993). Modelling Daily Series of Economic Activity. *Proceedings of the BES section of the Amer. Stat. Assoc.*

Espasa, A. and Peña, D. (1995). The Decomposition of Forecast in Seasonal ARIMA Models. *Journal of Forecasting*, **14**:565-583.

Un Estimador para el Análisis de Tablas de Vida con Muestras Complejas

MARTIN H. FELIX MEDINA y FELIPE DE J. PERAZA GARAY

Univ. Autónoma de Sinaloa. Culiacán, México

1. INTRODUCCIÓN

En un estudio sobre parasitosis intestinal realizado por la Escuela de Ciencias Químico Biológicas de la Universidad Autónoma de Sinaloa, se realizó un muestreo por conglomerados (familias). Los individuos infectados con cierto parásito se trataron con un medicamento antiparasitario. Se observaron durante un año los tiempos de reinfección y se construyó una tabla de vida. El objetivo del estudio fue estimar el tiempo en el cual la proporción de pacientes reinfectados alcanzaba un determinado valor.

Bajo el supuesto de observaciones independientes e idénticamente distribuidas (o.i.i.d.), la tabla de vida de este estudio se analiza mediante el método actuarial. Sin embargo, con muestras complejas, este estimador ignora la información del diseño muestral y, como se demuestra en este trabajo, puede producir resultados erróneos. La incorporación del diseño muestral en el análisis de tablas de vida no se encontró reportado en la literatura. Sin embargo, dentro del contexto del análisis de supervivencia, Chambles y Boyle (1985) y Binder (1993), entre otros, han estudiado el ajuste de modelos de regresión de riesgos proporcionales con datos de muestras complejas. La conclusión de estos autores es que el diseño muestral sí se debe de tomar en cuenta en la etapa de análisis de la información.

En este estudio se considera el problema de la estimación de la función de supervivencia a partir de una tabla de vida formada con datos obtenidos mediante un muestreo por conglomerados y estratificado. Se propone un estimador que se obtiene del estimador actuarial ordinario, al ponderar las observaciones por los inversos de sus probabilidades de inclusión. Se demuestra que si las distribuciones de los tiempos de vida de cada estrato son diferentes, y el diseño es no autoponderado, el estimador actuarial es asintóticamente sesgado, mientras que el estimador propuesto no presenta este problema. Mediante un estudio de simulación se confirma el resultado anterior, así como la falta de robustez del estimador de la varianza del estimador actuarial a la existencia de correlación intraconglomerado.

2. DEFINICIONES Y NOTACIÓN

Supóngase una población finita dividida en L estratos, el estrato h en N_h conglomerados, y el conglomerado i del estrato h formado por M_{hi} elementos. Mediante un diseño muestral por conglomerados y estratificado se toma una muestra de $n = \sum_{h=1}^{L} n_h$ conglomerados, donde n_h es el número de conglomerados muestreados del estrato h.

Los elementos muestreados son observados durante un cierto periodo y para cada uno de ellos se registra el intervalo $I_k = [\tau_{k-1}, \tau_k)$, $k = 1,\ldots,m+1$, $0 = \tau_0 < \tau_1 < \ldots < \tau_{m+1}$, en el cual ocurre un evento de interés o hay abandono del estudio. Con los datos obtenidos se construye una tabla de vida, y el objetivo es estimar la función de supervivencia $S(t)$ del

tiempo T de ocurrencia del evento de interés. Sea $p_u = 1 - q_u = Pr\{T > \tau_u | T \geq \tau_{u-1}\}$, u = 1, 2, ... ,m. Y para el elemento j del conglomerado i del estrato h, definanse las siguientes variables:

d_{hiju} = si el evento ocurre en I_u, y $d_{hiju} = 0$ en otro caso;

$v_{hiju} = 1$ si el evento no está presente al inicio de I_u, y $v_{hiju} = 0$ en otro caso, y

$w_{hiju} = 1$ si el elemento deja el estudio en I_u, y $w_{hiju} = 0$ en otro caso.

3. MÉTODO ACTUARIAL ORDINARIO

Bajo el supuesto de o.i.i.d., $S(\tau_k)$ se estima mediante el estimador actuarial ordinario (Lawless, 1982; pp. 53-59): $\hat{S}_{\text{ord}}(\tau_0) = 1$; $\hat{S}_{\text{ord}}(\tau_k) = \prod_{u=1}^{k} \hat{p}_u$, k = 1,2,...,m, donde

$$\hat{p}_u = 1 - \hat{q}_u,\ \hat{q}_u = \frac{\sum_{h=1}^{L}\sum_{i=1}^{n_h}\sum_{j=1}^{M_h} d_{hiju}}{\sum_{h=1}^{L}\sum_{i=1}^{n_h}\sum_{j=1}^{M_h} r_{hiju}} = \frac{\sum_{h=1}^{L}\sum_{i=1}^{n_h} d_{hi\bullet u}}{\sum_{h=1}^{L}\sum_{i=1}^{n_h} r_{hi\bullet u}},\ r_{hiju} = v_{hiju} - \frac{1}{2} w_{hiju},\ d_{hi\bullet u} = \sum_{j=1}^{M_h} d_{hiju}$$

y $r_{hi\bullet u} = \sum_{j=1}^{M_h} r_{hiju}$.

Un estimador de la varianza de $\hat{S}_{\text{ord}}(\tau_k)$, obtenido mediante el Método Delta es

$$\hat{V}\left[\hat{S}_{ord}(\tau_k)\right] = \left[\hat{S}_{ord}(\tau_k)\right]^2 \sum_{u=1}^{k} \frac{\hat{q}_u}{\hat{p}_u \sum_{h=1}^{L}\sum_{i=1}^{n_h} \tau_{hi\bullet u}}. \qquad (1)$$

Un intervalo de aproximadamente 100(1 - α)% de confianza para $S(\tau_k)$ es $\hat{S}_{\text{ord}}(\tau_k)$, $\pm$ $Z_{1-\alpha/2}\ \sqrt{\hat{V}\left[\hat{S}_{ord}(\tau_k)\right]}$, donde $Z_{1-\alpha/2}$ es el cuantil $1-\alpha/2$ de la distribución normal estándar.

4. EL ESTIMADOR PROPUESTO

Como lo señala Pfeffermann (1993), una de las maneras de introducir la información del diseño muestral en los estimadores es substituir las sumas que aparecen en las expresiones de los estimadores ordinarios por sumas ponderadas por los inversos de las probabilidades de inclusión. Con esta estrategia, a partir de la expresión de $\hat{S}_{\text{ord}}(\tau_k)$, se obtiene el siguiente estimador de $S(\tau_k)$: $\hat{S}^*(\tau_0) = 1$; $\hat{S}^*(\tau_k) = \prod_{u=1}^{k} \hat{p}_u^*$,

k = 1,2, ..., m, donde $\hat{p}_u^* = 1 - \hat{q}_u^*$ y $\hat{q}_u^* = \dfrac{\sum_{h=1}^{L} \frac{N_h}{n_h} \sum_{i=1}^{n_h} d_{hi\bullet u}}{\sum_{h=1}^{L} \frac{N_h}{n_h} \sum_{i=1}^{n_h} r_{hi\bullet u}}$

Un estimador de la varianza de $\hat{S}^*(\tau_k)$, obtenido mediante el método Delta es:

$$\hat{V}\left[\hat{S}^*(\tau_k)\right] = \left[\hat{S}^*(\tau_k)\right]^2 \hat{V}\left[\log \hat{S}^*(\tau_k)\right], \qquad (2)$$

donde $\hat{V}\left[\log \hat{S}^*(\tau_k)\right] = \sum_{u=1}^{k} \hat{V}\left[\log \hat{p}_u^*\right] + 2 \sum_{1 \le u < u' \le k} \mathrm{Cov}\left[\log \hat{p}_u^*, \log \hat{p}_{u'}^*\right]$, y $\hat{V}\left[\log \hat{p}_u^*\right]$ y $\mathrm{Cov}\left[\log \hat{p}_u^*, \log \hat{p}_{u'}^*\right]$ son los correspondientes elementos del estimador de la matriz de varianzas y covarianzas de $\log \hat{p}^* = \left(\log \hat{p}_1^*, \ldots, \log \hat{p}_m^*\right)^t$, dada por

$$\hat{V}\left[\log \underline{\hat{p}}^*\right] = A \left\{ \sum_{h=1}^{L} \frac{N_h}{n_h(n_h-1)} \sum_{i=1}^{n_h} \left(\underline{e}_{hi\bullet} - \underline{\bar{e}}_{h\bullet\bullet}\right)\left(\underline{e}_{hi\bullet} - \underline{\bar{e}}_{h\bullet\bullet}\right)^t \right\} A,$$

donde $A = \mathrm{diag}\left\{ \frac{1}{\hat{p}_1^* r_{\bullet\bullet\bullet}1}, \ldots, \frac{1}{\hat{p}_m^* r_{\bullet\bullet\bullet}m} \right\}$, $r_{\bullet\bullet\bullet k} = \sum_{h=1}^{L} \frac{N_h}{n_h} \sum_{i=1}^{n_h} r_{hi\bullet k}$; $k = 1,2,\ldots, m$,

$\underline{e}_{hi\bullet} = \left(d_{hi\bullet 1} - \hat{q}_1^* r_{hi\bullet 1}, \ldots, d_{hi\bullet m} - q_m^* r_{hi\bullet m}\right)^t$, y $\underline{\bar{e}}_{hi\bullet} = \frac{1}{n_h} \sum_{i=1}^{n_h} \underline{e}_{hi\bullet}$.

Un intervalo de aproximadamente 100(1 - α)% de confianza para $S(\tau_k)$ es: $\hat{S}^*(\tau_k) \pm t_{n-L(1-\alpha/2)}, \sqrt{\hat{V}\left[\hat{S}^*(\tau_k)\right]}$, donde $t_{n-L(1-\alpha/2)}$, es el cuantil $(1 - \alpha/2)$ de la distribución t-Student con $(n - L)$ grados de libertad.

5. EL EFECTO DE LA SELECCIÓN Y DEL DISEÑO MUESTRAL SOBRE $\hat{S}_{\mathrm{ord}}(\tau_k)$ Y $\hat{S}^*(\tau_k)$

Es una conjetura de los autores que, si el tiempo de vida está asociado a la variable diseño y el muestreo es no autoponderado, entonces el estimador $\hat{S}_{\mathrm{ord}}(\tau_k)$ presenta problemas de sesgo, lo cual no ocurre con el estimador $\hat{S}^*(\tau_k)$. En este trabajo se demuestra esta conjetura en el caso particular de una población estratificada, en la cual los tiempos de vida de los diferentes estratos son generados por diferentes distribuciones. Concretamente, se prueba que asintóticamente $E_{\xi p}\left(\hat{q}_u^*\right) \cong q_u$ y $\mathrm{E}_{\xi p}\left(\hat{q}_u\right) \neq q_u$, $u = 1,\ldots,m$; donde los subíndices ξ y p del operador esperanza, denotan respectivamente la distribución que genera los tiempos de vida y la distribución generada por el diseño. La demostración de este resultado se puede solicitar a los autores.

En cuanto al efecto del diseño, es un resultado conocido que si el coeficiente de correlación intraconglomerado (ρ) es positivo, la varianza de $\hat{q}_u$ es mayor que la que se obtiene a partir del supuesto de o.i.i.d. Y puesto que $\hat{V}\left[\hat{S}_{ord}(\tau_k)\right]$, dada por (1), se deriva bajo este supuesto, este estimador subestimará a $V\left[\hat{S}_{ord}(\tau_k)\right]$. Este problema no se presenta con el estimador (2), ya que toma en cuenta la posible existencia de la correlación intraconglomerado.

6. ESTUDIO DE SIMULACIÓN

Con el fin de observar si los resultados asintóticos de la sección anterior son válidos con muestras de tamaño fijo, y de comparar el desempeño de ambos estimadores, se realizó un estudio de simulación. Se consideraron poblaciones con dos estratos y con 500 conglomerados cada uno. Los tamaños de los conglomerados de los estratos 1 y 2 fueron respectivamente 6 y 3. Las muestras se obtuvieron mediante los siguientes dos diseños muestrales: diseño autoponderado (A), con 50 conglomerados de cada estrato, y diseño con probabilidad de selección proporcional al tamaño del conglomerado (ppt), con 66 conglomerados del primer estrato y 33 del segundo. En ambos diseños, los conglomerados de cada estrato se obtuvieron mediante muestreo aleatorio simple con reemplazo.

Los tiempos de vida de cada estrato se generaron mediante la distribución lognormal $\Lambda\left(\mu_h, \sigma_h^2\right)$, $h = 1, 2$; y de tal manera que los tiempos de vida de los elementos del mismo conglomerado tuvieran una correlación intraconglomerado ρ. Aproximadamente el 10% de los datos se censuraron aleatoriamente, con tiempos de censura con distribución lognormal.

TABLA 1.

Resultados de la estimación de $S(\tau_k)$, basados en 1000 simulaciones.
Tiempos de vida de ambos estratos con distribución $\Lambda(1,1)$.

ρ_{int}	diseño	Est.	s(1)	s(2)	s(3)	s(4)	s(5)	s(6)
	A	$\hat{S}_{ord}$	0.020	0.036	0.052	0.066	0.081	0.095
			0.945	0.956	0.947	0.943	0.933	0.938
		$\hat{S}*$	0.020	0.036	0.052	0.066	0.081	0.095
			0.948	0.960	0.952	0.946	0.939	0.939
	ppt	$\hat{S}_{ord}$	0.020	0.035	0.047	0.062	0.076	0.091
			0.939	0.951	0.955	0.959	0.943	0.954
		$\hat{S}*$	0.021	0.037	0.050	0.064	0.079	0.096
			0.929	0.938	0.946	0.941	0.932	0.929
	A	$\hat{S}_{ord}$	0.032	0.059	0.082	0.102	0.121	0.144
			0.777	0.775	0.768	0.792	0.802	0.787
		$\hat{S}*$	0.032	0.059	0.082	0.102	0.121	0.144
			0.953	0.957	0.958	0.962	0.949	0.946
	ppt	$\hat{S}_{ord}$	0.033	0.061	0.086	0.110	0.132	0.150
			0.747	0.743	0.725	0.713	0.740	0.746
		$\hat{S}*$	0.033	0.061	0.085	0.107	0.129	0.149
			0.929	0.939	0.941	0.946	0.932	0.941
	A	$\hat{S}_{ord}$	0.040	0.072	0.099	0.127	0.152	0.177
			0.677	0.701	0.695	0.678	0.688	0.699
			0.040	0.072	0.099	0.127	0.152	0.177
			0931	0.939	0.945	0.938	0.943	0.944
	ppt	$\hat{S}_{ord}$	0.039	0.072	0.096	0.121	0.144	0.168
			0.692	0.669	0.679	0.694	0.702	0.706
		$\hat{S}*$	0.038	0.070	0.093	0.117	0.139	0.163
			0.922	0.945	0.953	0.952	0.959	0.960

(La primera entrada en cada celda es la raíz cuadrada del error cuadrático medio relativo, la segunda entrada el porcentaje de cobertura).

Los resultados se presentan en las tablas 1 y 2. Estos resultados muestran que:

- En poblaciones con correlación intraconglomerados, las propiedades de cobertura del estimador $\hat{S}_{ord}(\tau_k)$ se distorsionan. Este problema se debe a que el estimador $\hat{V}\left(\hat{S}_{ord}(\tau_k)\right)$ subestima la verdadera varianza. En el caso del estimador $\hat{S}_{ord}(\tau_k)$ no se presenta este tipo de problema.
- En poblaciones con tiempos de vida de ambos estratos con igual distribución, la eficiencia de $\hat{S}_{ord}(\tau_k)$ y $\hat{S}^*(\tau_k)$, medida en términos de $\sqrt{ECM_{rel}}$ es muy similar.
- En poblaciones con diferentes distribuciones de los tiempos de vida de cada estrato, la eficiencia de los estimadores depende del diseño muestral. Con el diseño autoponderado, ambos estimadores coinciden y por tanto, su eficiencia también. Sin embargo, con el diseño ppt, la eficiencia de $\hat{S}_{ord}(\tau_k)$ se reduce notablemente. La reducción de la eficiencia se debe al sesgo del estimador.
- Aún en el caso óptimo para $\hat{S}_{ord}(\tau_k)$ (o.i.i.d.), la eficiencia y las propiedades de cobertura de ambos estimadores son muy similares.

TABLA 2

Resultados de la estimación de $S(\tau_k)$, basados en 1000 simulaciones. Tiempos de vida del estrato 1 con distribución $\Lambda(1,1)$, y del estrato 2 con distribución $\Lambda(-0.05,1)$.

ρ_{int}	diseño	Est.	s(1)	s(2)	s(3)	s(4)	s(5)	s(6)
	A	$\hat{S}_{ord}$	0.036	0.059	0.087	0.115	0.146	0.173
			0.961	0.967	0.949	0.953	0.950	0.954
		$\hat{S}^*$	0.036	0.059	0.087	0.115	0.146	0.173
			0.954	0.948	0.928	0.935	0.931	0.941
	ppt	$\hat{S}_{ord}$	0.089	0.146	0.182	0.205	0.225	0.244
			0.385	0.373	0.429	0.537	0.618	0.697
		$\hat{S}^*$	0.034	0.062	0.091	0.128	0.168	0.207
			0.966	0.929	0.939	0.910	0.908	0.903
	A	$\hat{S}_{ord}$	0.062	0.093	0.125	0.153	0.185	0.225
			0.762	0.816	0.829	0.874	0.880	0.866
		$\hat{S}^*$	0.062	0.093	0.125	0.153	0.185	0.225
			0.944	0.945	0.956	0.956	0.955	0.937
	ppt	$\hat{S}_{ord}$	0.100	0.163	0.199	0.225	0.247	0.270
			0.446	0.422	0.468	0.539	0.588	0.636
		$\hat{S}^*$	0.056	0.095	0.132	0.174	0.214	0.261
			0.962	0.952	0.941	0.921	0.910	0.917

(La primera entrada en cada celda es la raíz cuadrada del error cuadrático medio relativo, la segunda entrada el porcentaje de cobertura)

7. CONCLUSIONES GENERALES

La conclusión principal que se obtiene de este estudio, es la necesidad de considerar el diseño muestral en la estimación de la función de supervivencia a partir de una tabla de vida formada con datos de muestras complejas. Los resultados obtenidos con el estimador

propuesto son bastante alentadores, sin embargo, debido a lo limitado del estudio de simulación, se requiere observar su comportamiento con otro tipo de distribuciones y de diseños muestrales. Finalmente, cabe aclarar que la extensión del estimador propuesto a diseños multietápicos y/o con probabilidades de inclusión dentro de cada estrato es directa.

REFERENCIAS

Binder, D.A. (1992). Fitting Cox's proportional hazards models from survey data. *Biometrika* **79**, 139-47.

Chambles, L.E. and Boyle, K.E. (1985). Maximum likelihood methods for complex sample data: logistic regression and discrete proportional hazards models. *Communications in Statistics - Theory and methods* **14**, 1377-92.

Lawless, J.F. (1982). *Statistical Models and Methods for Lifetime Data*. Wiley: New York.

Pfeffermann D. (1993). The role of the sampling weights when modeling survey data. *International Statistical Review* **61**, 317-37.

Modelos de Superviviencia en Doble Censura con Parámetros Variables en el Tiempo y Covariables

ARTURO J. FERNANDEZ JOSE I. BRAVO IÑIGO DE FUENTES

Univ. de La Laguna. Tenerife, España

1. INTRODUCCIÓN

Una cuestión de interés esencial en los análisis de Fiabilidad y Supervivencia es la estimación de la distribución de la variable X, *tiempo de vida*, en estudio. Debido a los métodos de muestreo y a ciertos factores sin control experimental, en algunas ocasiones aparecen observaciones doblemente censuradas. Los artículos de Bravo y Esteban (1993a,1993b), Gu y Zhang (1993), Jewell et al. (1994) o Bravo et al. (1995) son algunos de los muchos trabajos que consideran este tipo de observaciones.

Suponemos que, además del tiempo de vida, hemos observado sobre cada individuo cierto número de covariables que afectan a la distribución de la supervivencia. El modelo general de riesgos proporcionales, introducido por Cox (1972), considera que el efecto de las covariables sobre la tasa de fallo es multiplicativo. De este modo, para un individuo con vector columna p-dimensional de covariables $z=(z_1,\ldots,z_p)'$, la tasa de fallo es $h(x;z)=\psi(z)\lambda(x)$, donde $\psi(z)$ es alguna función no negativa y $\lambda(x)=f_0(x)/S_0(x)$ es la tasa de fallo base, es decir, para un individuo en condiciones normales $(z=z_0)$, mientras que $S_0(x)=Pr(X>x|\ z=z_0)$ y $f_0(x)$ son respectivamente las funciones de supervivencia y de densidad de X cuando $z=z_0$ respectivamente. Luego las funciones de supervivencia y de densidad de X, dado el vector de covariables z, serían respectivamente

$$S(x;z)=\{S_0(x)\}^{\psi(z)} \quad y \quad f(x;z)=f_0(x)\cdot\psi(z)\cdot\{S_0(x)\}^{\psi(z)-1}$$

La función $\psi(z)$ puede ser parametrizada como $\psi(z;\beta)$, donde $\beta=(\beta_1,\ldots,\beta_p)$ es un vector fila p-dimensional de parámetros desconocidos. La formulación log-lineal, $\psi(z;\beta)=\exp(\beta z)$, es la más popular.

Desde el punto de vista teórico existen principalmente dos diferentes líneas de investigación. En la primera, no se conoce la tasa de fallo base; en la segunda, se supone un modelo completamente paramétrico para ésta. Nuestro artículo se incluye dentro de la segunda línea de investigación y estudia el modelo de riesgos proporcionales con datos doblemente censurados, considerando un modelo paramétrico general para la distribución base de la supervivencia (apartado 2) o un modelo definido a trozos (apartado 3).

2. MODELO PARAMÉTRICO PARA LA DISTRIBUCIÓN BASE

Supongamos que, mediante estudios preliminares, sabemos que la distribución base de X está dada por un modelo paramétrico que depende de un vector $\theta=(\theta_1,\ldots,\theta_q)$ de parámetros desconocidos. Considerando el modelo de riesgos proporcionales con multiplicador $\psi(z;\beta)$, nos proponemos estimar el vector de parámetros desconocidos $\phi=(\theta,\beta)$ a partir de N observaciones independientes (t_i,δ_i,z_i) $i=1,..,N$, donde t_i es el tiempo

de vida observada, δ_i, indica el tipo de censura (0, 1 ó 2 si la observación t_i está no censurada, censurada por la derecha o por la izquierda respectivamente) y z_i es el vector de p covariables, correspondientes al sujeto i-ésimo.

Supuesto que el mecanismo de censura es no informativo e independiente de las covariables, la función de verosimilitud es proporcional a

$$L(\phi) = \prod_{i \in I_0} h(i)\exp\{-H(i)\} \prod_{j \in I_1} \exp\{-H(j)\} \prod_{k \in I_2} \left[1 - \exp\{-H(k)\}\right]$$

y el sistema de $(q + p)$ ecuaciones de verosimilitud es:

$$\begin{cases} \dfrac{\partial}{\partial \theta_r} \ln L(\phi) = \displaystyle\sum_{i \in I_0} \frac{\lambda_r(i)}{\lambda(i)} + \sum_{j \in I} \{w(j) - 1\} \frac{\Lambda_r(j)}{\Lambda(j)} H(j) = 0, \quad r = 1, \ldots, q \\ \dfrac{\partial}{\partial \beta_s} \ln L(\phi) = \displaystyle\sum_{i \in I_0} \frac{\psi_s(i)}{\psi(i)} + \sum_{j \in I} \{w(j) - 1\} \frac{\psi_s(j)}{\psi(j)} H(j) = 0, \quad s = 1, \ldots, p, \end{cases} \tag{1}$$

donde $\lambda(i)$, $\Lambda(i)$, $\psi(i)$ son respectivamente los valores de la tasa de fallo base, tasa de fallo acumulativa base y multiplicador ψ, para el sujeto i-ésimo, $H(i) = \psi(i)\,\Lambda(i)$, h(i) = $\psi(i)\lambda(i)$ y $w(i) = I(\delta_i = 2)/[1 - \exp\{-H(i)\}]$ para $i = 1,.., N$, $I = \{1,..., N\}$ e $I_j = \{i \in I: \delta_i = j\}$, $j = 0,1,2$, y los subíndices r y s indican respectivamente $\partial/\partial\beta_s$ de la correspondiente función.

Hallaríamos los estimadores máximo-verosímiles $\hat{\theta}$ y $\hat{\beta}$ maximizando la función log-verosimilitud ln $L(\phi)$ o resolviendo el sistema de $(q + p)$ ecuaciones de verosimilitud mediante el método de Newton-Raphson o algún algoritmo de relajación.

Podemos utilizar la inversa de la matriz de información observada de Fisher para estimar la matriz de covarianzas asintótica del estimador $\hat{\phi} = (\hat{\theta}, \hat{\beta})$, contrastar hipótesis sobre los valores de los parámetros desconocidos y estimar, mediante el método delta, la varianza asintótica de cualquier función de ϕ con derivadas parciales continuas.

Nótese que, como la función de supervivencia de X dado z es $S(x; z, \phi)$, entonces $S(X; z, \phi)$ está uniformemente distribuida, con lo cual $H(X; z, \phi)$= -ln $S(X; z, \phi)$ tiene distribución exponencial de media uno. Una vez estimado el vector ϕ, es posible contrastar la validez del modelo ajustado analizando la variable residual $R = H(X; z, \hat{\phi})$, cuya distribución debería ser aproximadamente exponencial unitaria.

3. DISTRIBUCIÓN BASE CON PARÁMETROS VARIABLES

Suponemos ahora que los parámetros que caracterizan a la distribución base de la supervivencia pueden variar en el tiempo. Por esta razón, introducimos una partición del intervalo de tiempo $(0, \infty)$ en $(k + 1)$ subintervalos mediante los puntos de cambio $a_1, \ldots, a_k$, de forma que la parametrización de la distribución base se mantiene en cada subintervalo, aunque con diferentes parámetros en cada uno de ellos. Por conveniencia, definimos $a_0 = 0$ y $a_{k+1} = \infty$. Puede esperarse que estos modelos definidos a trozos describirán la mayoría de los procesos de supervivencia con riesgos proporcionales que incluyan una serie de puntos en los cuales se alteren las condiciones imperantes.

Considerando el modelo log-lineal de riesgos proporcionales y que la distribución base de la supervivencia es:

- Exponencial a trozos: $g(t) = \ln\Lambda(t) = \ln t + \varepsilon_j$
- Weibull a trozos: $g(t) = \ln\Lambda(t) = \alpha_j \ln t + \varepsilon_j$
- De Valor Extremo Generalizada a trozos: $g(t) = \ln\Lambda(t) = \alpha_j t^{\gamma_j} + \varepsilon_j$

para $t \in (a_{j-1}, a_j]$, $j = 1, .., k+1$, obtenemos a partir de (1) los correspondientes sistemas de ecuaciones de verosimilitud que generalizan los resultados obtenidos, entre otros, por Aitkin y Clayton (1980), Noura y Read (1990) y García et al. (1993) al caso de doble censura. En los dos últimos casos suponemos la continuidad de la función $g(t)$.

Si definimos m_i como el índice del subintervalo al que pertenece el i-ésimo dato t_i, para $i = 1, .., N$, esto es, $m_i = j$ si $t_i \in (a_{j-1}, a_j]$, para $j = 1, ..., k+1$, obtenemos, supuesto que la distribución base es Weibull a trozos, que $\alpha_j \ln a_j + \varepsilon_j = \alpha_{j+1} \ln a_j + \varepsilon_{j+1}$ para j = 1, ..., k, con lo cual $\varepsilon_j = \varepsilon_1 + \sum_{r=2}^{j} (\alpha_{r-1} - \alpha_r) \ln a_{r-1}$, $j = 2, ..., k+1$. Luego $g(t_i) = \varepsilon_1 + \alpha_{mi} \ln t_i + \sum_{r=2}^{mi} (\alpha_{r-1} + \alpha_r) \ln a_{r-1}, i = 1, ..., N$, donde se omite el sumatorio en r, si $m_i = 1$.

Si la distribución base es de valor extremo generalizada a trozos se verifica que $\alpha_j a_j^{\gamma_j} + \varepsilon_j = \alpha_{j+1} a_j^{\gamma_{j+1}} + \varepsilon_{j+1}$ $j = 1, ..., k$, de lo cual obtenemos que $\varepsilon_j = \varepsilon_1 + \alpha_1 a_1^{\gamma_1} + \sum_{r=2}^{j-1} \alpha_r \left(a_r^{\gamma_r} - a_{r-1}^{\gamma_r}\right) - \alpha_j a_{j-1}^{\gamma_j}$, j = 2, ..., k + 1, donde la suma en r se omite si $j = 2$. De este modo, resulta que

$$g(t_i) = \varepsilon_1 + I(m_i = 1)\alpha_1 t_i^{\gamma_1}$$

$$+ I(m_i > 1)\left\{\alpha_{mi}\left(t_i^{\gamma_{mi}} - a_{mi-1}^{\gamma_{mi}}\right) + \alpha_1 a_1^{\gamma_1} + \sum_{r=2}^{m_i-1} \alpha_r \left(a_r^{\gamma_r} - a_{r-1}^{\gamma_r}\right)\right\}$$

para $i = 1, ..., N$, donde se omite la suma en r si $m_i \leq 2$.

Nótese que si $\gamma_j = 1$ para $j = 1, .., k+1$, obtenemos la distribución de valor extremo y si hacemos $\gamma_j \to 0$ aparece la distribución de Weibull. Por tanto, podemos utilizar el estadístico de razón de verosimilitudes para contrastar la hipótesis nula de distribución de Weibull (o de valor extremo) frente a la alternativa de una distribución de valor extremo generalizada.

En nuestro análisis es preferible utilizar la fórmula

$$\frac{\partial}{\partial \theta_r} \ln L(\phi) = \sum_{i \in I} \left\{\frac{g'_r(i)}{g'(i)} + g_r(i)\right\} + \sum_{j \in I} \{w(j) - 1\} g_r(j) H(j), \quad r = 1, ..., q,$$

donde $g'(i) = \partial g(t_i)/\partial t$ y $H(i) = \exp\{g(t_i) + \beta z_i\}$ para $i = 1, ..., N$.

Como análisis gráfico preliminar es aconsejable agrupar los individuos en subconjuntos con valores similares de las principales covariables y examinar gráficamente la estimación autoconsistente (Turnbull, 1974; Gu y Zhang, 1993), $S^{\alpha}(t)$, de la función de supervivencia para cada grupo. Podemos considerar apropiadas las distribuciones de Weibull

(exponencial), de valor extremo o de valor extremo generalizada si la gráfica de $\ln\{-\ln S^e(t)\}$ es aproximadamente lineal a trozos en $\ln t$ (y con pendiente uno), en t o en algunas potencias de t, respectivamente. Además este procedimiento puede sugerirnos la localización aproximada de los puntos de cambio.

Una vez ajustado el modelo es posible comprobar su validez analizando los residuos $r_i = \hat{H}(i)$, $i = 1, .., N$. Si el modelo propuesto es apropiado $r_1, .., r_N$ deben comportarse aproximadamente como un conjunto de datos (doblemente censurados) de la distribución exponencial unitaria, aunque no serán independientes.

De forma similar al método semiparamétrico analizado en Bravo et al. (1995), proponemos la estimación de la supervivencia residual $\tilde{S}(r)$ para el caso de doble censura como el valor esperado de la supervivencia residual empírica, $S^e(r) = N^{-1}\sum_{i=1}^{N} I(R_i > r)$, esto es,

$$\tilde{S}_r = \frac{1}{N}\left\{\sum_{i\in I} I(r_i > r) + \sum_{j\in I_{10}} e^{-(r-r_j)} I(r_j \le r) - \sum_{k\in I_2} \frac{1-e^{-r}}{1-e^{-rk}} I(r_k > r)\right\}.$$

Si el modelo es válido, la representación gráfica de $-\ln\tilde{S}(r)$ frente a r debe aproximarse a la bisectriz del primer cuadrante. Para evitar la excesiva variabilidad de $\tilde{S}(r)$ cuando r es grande, y considerando que $\tilde{S}(r)$ es esencialmente una probabilidad binomial para cada r, también es aconsejable representar la transformación estabilizadora de la varianza arcsen $\sqrt{\tilde{S}(r)}$ frente a arcsen$\left(e^{-r/2}\right)$.

REFERENCIAS

Aitkin, M. and Clayton, D. (1980), The fitting of exponential, Weibull and extreme value distributions to complex censored survival data using GLIM, *Appl. Statist.*, **29**, 156-163.

Bravo, J. I. and Esteban, M. D. (1993a), Nonparametric estimation of survival functions for doubly censored observations when the lifetime hazard rate is proportionally related to the censoring hazard rates, *Comm. Statist.-Theory Meth.*, **22**, 3237-3253.

Bravo, J. I. and Esteban, M. D. (1993b), Partially parametric estimation of survival functions for doubly censored observations, ASMDA 93: Proc. Int. Symp. Appl. Stochastic Models and Data Analysis (eds. J. Janssen and C. H. Skiadas), *World Scientific Publ.* Co., 106-113.

Bravo, J. I., De Fuentes, I. and Fernández, A. J. (1995)}, A semi-parametric estimation of a survival function from incomplete and doubly censored data, *Comm. Statist.-Theory Meth.*, **24**(11), (en prensa).

Cox, D. R. (1972), Regression models and life-tables (with discussion), J. R. *Statistics Soc. B*, **34**, 187-220.

García, J., Lara, A. M., Ollero, J. and Pérez, R. (1993)}, A study of a survival model with a piecewise generalized extreme value distribution. ASMDA 93: Proc. Int. Symp. Appl. Stochastic Models and Data Analysis (eds. J. Janssen and C. H. Skiadas), *World Scientific Publ. Co.*, 265-274.

Gu M. G. and Zhang C. H. (1993), Asymptotic properties of self-consistent estimators based on doubly censored data, *Ann. Statist.*, **21**, 611-624.

Jewell, N. P., Malahi, H. M. and Vittinghoff, E. (1994)}, Nonparametric estimation for a form of doubly censored data, with applications to two problems in AIDS, J. *Amer. Statist. Assoc.*, **89**, 7-18.

Noura, A. A. and Read, K. L. Q. (1990), Proportional hazards changepoint models in survival analysis, *Appl. Statist.*, **39**, 241-253.

Turnbull, B. W. (1974)}, Nonparametric estimation of a survivorship function with doubly censored data, J. *Amer. Statist. Assoc.*, **69**, 169-173.

Análisis Estadístico sobre Infertilidad en México

TATIANA FERNANDEZ N. y MA. DE LOURDES DE LA FUENTE D.

ITAM, México

1. INTRODUCCIÓN

Durante los últimos 15 años se han logrado grandes avances en el estudio de la infertilidad. Estos avances se refieren al conocimiento de los mecanismos que intervienen en el proceso de reproducción, tanto en personas normales, como en aquellas en estado patológico. Sin embargo, aún no se cuenta con una visión lo suficientemente completa que permita el desarrollo de nuevas y mejores técnicas para el diagnóstico y tratamiento de este padecimiento.

El objeto de este trabajo es analizar el comportamiento de una base de datos[1] de una población de clase media y media alta de casos de infertilidad en México. En un estudio anterior : (Fernández y de la Fuente, 1994), se llevó a cabo el análisis exploratorio de esta base de datos con objeto de describir los perfiles de la mujer, del hombre y de la pareja infértil. Asimismo, a partir del cruce de algunas variables se observaron algunas relaciones entre las características del perfil de la mujer infértil y las causas fisiológicas de la infertilidad.

En esta parte del estudio, se obtuvieron frecuencias condicionales asociadas a la propensión de embarazo a partir de la construcción de árboles de clasificación y posteriormente, con base en la metodología de Modelos de Elección Cualitativa, se estimaron modelos Probit y Logit con objeto de poder determinar qué variables tienen un efecto significativo sobre la probabilidad de embarazo, así como poder cuantificar la magnitud de estos efectos. A partir de los resultados obtenidos en la estimación de estos modelos se hicieron algunas simulaciones para determinar como afectan los atributos considerados a dicha propensión de embarazo. La 'unidad de información'[2], para fines del estudio, es el ciclo menstrual de la mujer. Este es caracterizado por la edad de la mujer, la causa de infertilidad, el tipo de medicamento empleado, el número de folículos observados y finalmente, si hubo embarazo o no lo hubo. Los medicamentos que fueron tomados en cuenta fueron Omifin, Pergonal y HCG. Su función es la siguiente: el Omifin estimula, a través del hipotálamo, la secreción de hormonas hipofisiarias, conocidas como gonadotropinas, que a su vez estimulan a los folículos; el Pergonal consiste de gonadotropinas, que generan un estímulo directo sobre los folículos. En ocasiones, para 'asegurar' el estímulo sobre el folículo, se suministran conjuntamente el Omifin y el Pergonal. Finalmente, el HCG actúa estimulando la ruptura del folículo, provocando con esto que se libere el óvulo.

Las características a analizar por ciclo menstrual fueron la edad de la mujer, la causa de infertilidad, el tratamiento (medicamento), el número de folículos observados, el uso del HCG y la presencia o ausencia de embarazo. A partir de la metodología de árboles de

---

[1] La base de datos fue proporcionada por un especialista en endocrinología de la reproducción y consiste en 436 ciclos mentruales.

[2] La selección de las características de interés se sujetó a juicio y sugerencia del especialista.

clasificación, se obtuvieron frecuencias condicionales asociadas al embarazo de acuerdo a los atributos edad, causa y tratamiento. Los resultados fueron los siguientes:

Atributo	Frecuencias Condicionales (Embarazo/Atributo)
Edad	
Menos de 35 años	0.129
35 años ó más	0.052
Causa	
Anovulación	0.178
Anovulación y Otras Causas	0.077
Tratamiento	
Omifin	0.104
Pergonal	0.067
Omifin y Pergonal	0.127

Como puede observarse las mujeres menores de 35 años presentan embarazo con una frecuencia 2.5 veces mayor que las de 35 años y más. Análogamente, aquellas con anovulación se embarazan con una frecuencia 2.3 veces mayor que aquellas con anovulación y otras causas. Con respecto al tratamiento, el más efectivo en términos de la frecuencia de embarazo resultó ser el de Omifin y Pergonal, representando una frecuencia 1.22 veces mayor con respecto a Omifin y 1.9 veces mayor con respecto a Pergonal. De acuerdo a lo anterior es posible concluir que, aún cuando en términos generales la frecuencia condicional de embarazo dado cualquier atributo no resulta ser alta, las características que favorecen el embarazo son: edad menor de 35 años, infertilidad por anovulación como causa única y el tratamiento conjunto de Omifin y Pergonal

2. MODELOS DE ELECCIÓN CUALITATIVA

Los Modelos de Elección Cualitativa (Pindyck y Rubinfeld, 1992), son aquellos en que la variable dependiente involucra dos o más opciones de naturaleza cualitativa, el supuesto que los sustenta es que la inclinación sobre una de las alternativas depende de características específicas del individuo u objeto en cuestión. El objetivo de éstos es determinar la propensión hacia una alternativa sobre otra(s), dado un conjunto de atributos. Permiten encontrar la relación entre el conjunto de atributos y la probabilidad asociada a una alternativa, dada la naturaleza del problema que nos ocupa, esto es, estimar la propensión de embarazo a partir de ciertas características del perfil de la mujer. El interés se centra sobre este tipo de modelos cuando la elección es binaria. La especificación de los modelos fue la siguiente:

Modelo Probit

$$Z_i = \alpha + \beta_1 EDAD + \beta_2 D_1 + \beta_3 D_2 + \beta_4 D_3 + \beta_5 No.FOLI + \beta_6 HCG$$

Modelo Logit

$$L_N \frac{P_i}{1 \quad D} = \beta_0 + \beta_1 EDAD + \beta_2 D_1 + \beta_3 D_2 + \beta_4 D_3 + \beta_5 No.FOLI + \beta_6 HCG$$

Donde:

$$D_1 = \begin{cases} 1 & Si\ se\ presento\ anovulacion \\ 0 & en\ otro\ caso \end{cases} \quad D_2 = \begin{cases} 1 & Si\ se\ administro\ Omifin \\ 0 & Otro\ caso \end{cases} \quad D_3 = \begin{cases} 1 & Si\ se\ administro\ Pergonal \\ 0 & En\ otro\ caso \end{cases}$$

Resultados de la Estimación del Modelo Probit

$$Z_i = -0.6541974 - 0.0549318EDAD + 0.5192374D_1 + 0.03345744No.FOLI^3$$

(0.706) (0.019) (0.173)
(0.147)[4].....
(0.354) (0.005) (0.003)
(0.023).....

Como se aprecia en las ecuaciones hay coincidencia entre los signos esperados y los obtenidos de los coeficientes. Aún cuando las variables correspondientes al tratamiento, D2 y D3, al igual que HCG, resultaron ser no significativas para el modelo, es clara la relación a nivel clínico entre el tratamiento y el número de folículos observados. Esta última variable resultó ser significativa, guardando relación directa con la probabilidad de embarazo, ya que al incrementarse el número de folículos se incrementa dicha probabilidad; la edad presenta una relación inversa con respecto a la probabilidad de embarazo. Finalmente, tanto la causa como el número de folículos, presentaron una relación directa con respecto a la probabilidad de embarazo, siendo el efecto del número de folículos mayor en términos relativos al efecto de la causa. El índice del cociente de verosimilitudes ρ (Maddala, 1983), definido como $\rho = 1 - \{L(\beta^*)/L(\sigma)\}$ donde $L(\beta^*)$ es el valor de la función log-verosimilitud evaluada en el máximo y $L(\sigma)$ es el valor de la función log-verosimilitud cuando todos los parámetros son iguales a cero, resultó ser igual a 0.08. Esto significa que gran parte de la variabilidad de la propensión al embarazo queda aún sin explicar. No obstante, el objetivo de este modelo era cuantificar la relación entre los atributos y la propensión a embarazo, y no el de predecir probabilidades, por lo que las conclusiones a las que se llegó con el modelo siguen siendo válidas.

Resultados de la Estimación del Modelo Logit

$$Zi = -1.1004533 - 0.1015421EDAD + 0.9824115D1 + 0.6277259No.FOLI \qquad 4.15^5$$

(1.380) (0.037) (0.322) (0.294).....
(0.425) (0.006) (0.002) (0.033).....

Como es posible apreciar, los signos obtenidos para las βi's son consistentes tanto con los signos esperados como los obtenidos en el modelo probit. En este caso $\rho = 0.0811$, y su interpretación resulta ser la misma que para la ρ obtenida con el Modelo Probit.

A partir de los resultados de la estimación de estos modelos se simularon las probabilidades de embarazo de una mujer menor a 35 (28 años) y para el caso de una mujer

[3] Sólo se incluyen las variables que resultaron estadísticamente significativas.
[4] Los valores entre paréntesis corresponden a las desviaciones estandar y a las estadisticas T
[5] Una vez más, HCG, y D2 y D3 resultaron ser atributos no significativos dentro del modelo (Ver valor ❷ correspondiente a cada variable). Por lo que se obtuvo una segunda estimación de éste.

mayor a 35 (36 años)[6] de acuerdo a las posibles causas de infertilidad. Los resultados fueron los siguientes

Modelo Probit
Edad = 28 < 35

Número de Folículos

Causa.	0	1	2	3 ó más
Anovulación	0.047	0.090	0.158	0.252
Anovulación y Otras Causas	0.014	0.032	0.064	0.117

Edad = 36 > 35

Número de Folículos

Causa.	0	1	2	3 ó más
Anovulación	0.017	0.038	0.075	0.134
Anovulación y Otras Causas	0.004	0.011	0.025	0.052

La mayor probabilidad de embarazo, la presentan mujeres menores de 35 años, con anovulación y 3 folículos o más. Independientemente de la edad, dada anovulación con 0 y 1 folículos, la probabilidad de embarazo es el triple de la probabilidad dado el mismo número de folículos y anovulación y otras causas. Para el caso en que se presentaron 2 y 3 folículos la probabilidad de embarazo dada anovulación es el doble de la probabilidad con respecto a anovulación y otras causas.

Modelo Logit
Edad = 28 < 35

Número de Folículos

Causa.	0	1	2	3 ó más
Anovulación	0.049	0.088	0.154	0.254
Anovulación y Otras Causas	0.019	0.035	0.064	0.113

Edad = 36 > 35

Número de Folículos

Causa.	0	1	2	3 ó más
Anovulación	0.023	0.041	0.075	0.131
Anovulación y Otras Causas	0.009	0.016	0.029	0.054

[6] La edad de la mujer para las simulaciones se escogió arbitrariamente

La mayor probabilidad de embarazo la presenta la mujer menor de 35 años, con anovulación y 3 ó más folículos.

3. RESULTADOS SOBRESALIENTES DE LAS ESTIMACIONES

En términos generales, la probabilidad de embarazo resulta ser baja, siendo que en el mejor de los casos (Edad menor a 35 años, Anovulación y 3 ó más folículos), uno de cada cuatro casos, aproximadamente, presentan embarazo. El hecho de ser mayor de 35 años tiene un efecto negativo sobre la probabilidad de embarazo de manera significativa, al igual que el hecho de tener infertilidad por causa múltiple. En lo que se refiere al efecto del número de folículos en la probabilidad de embarazo, si se considera el caso de mujeres menores de 35 años, con anovulación, cuando se observa tan sólo un folículo: 1 de cada 10 presenta embarazo; cuando se tienen 2 folículos: 3 de cada 20 se embarazan; y por último, en el caso de 3 folículos: 1 de cada 4 presentan embarazo.

Se hizo la comparación entre las probabilidades promedio obtenidas de los modelos Probit y Logit y, así como entre las frecuencias condicionales, de acuerdo a los atributos de edad, causa y tratamiento, obteniéndose resultados consistentes en todos los casos.

1. Para la edad:

Probabilidades Promedio

	Frecuencia Condicional	Modelo Probit	Modelo Logit
Menos de 35 años	0.129	0.131	0.129
35 años o más	0.052	0.051	0.053

Se puede apreciar la relación de 2.5:1 entre la probabilidad de embarazo para las menores de 35 años y las de 35 años y más.

2. Para la causa:

Probabilidades Promedio

	Frecuencia Condicional	Modelo Probit	Modelo Logit
Anovulación	0.178	0.186	0.188
Anovulación y Otras Causas	0.077	0.076	0.075

Se observa una relación aproximada de 3:1, entre las mujeres que presentan infertilidad por anovulación y las que la presentan por anovulación acompañada por otras causas.

3. Para el tratamiento

Probabilidades Promedio

	Frecuencia Condicional	Modelo Probit	Modelo Logit
Omifin	0.104	0.089	0.089
Pergonal	0.067	0.098	0.099
Omifin y Pergonal	0.127	0.116	0.116

El efecto de Omifin y Pergonal, suministrados conjuntamente, resultan ser el tratamiento más eficiente, incrementando la probabilidad de embarazo en poco menos de 0.02 con respecto a los valores correspondientes a Omifin o Pergonal.

4. CONCLUSIONES

Los factores significativos de incidencia en la propensión de embarazo resultaron ser la edad, la causa de infertilidad y el número de folículos observados durante el ciclo. Relacionándose inversamente el primer atributo y directamente los restantes, con la probabilidad de embarazo. Aún cuando para fines de los modelos, las variables de tratamiento no resultaron ser significativas, es un hecho que éstas variables inciden de manera indirecta en la propensión de embarazo, vía estimulación de folículos. Por último, la conjunción de atributos con los que la probabilidad de embarazo es máxima resultaron ser: edad menor a 35 años, causa única anovulación y 3 o más folículos. En general, el hecho de que los resultados obtenidos con cada uno de los modelos sean muy similares, da mayor validez y consistencia a éstos.

REFERENCIAS

Ehrenberg, A.S.C. (1975). *Data Reduction*. Great Britain. John Wiley & Sons

Maddala, G.S. (1983). *Limited Dependent and Qualitative Variables in Econometrics*. Econometric Society Monographs, Cambridge University Press

Fernández, T. y de la Fuente, M.L. (1994). Análisis Exploratorio de la Infertilidad en México. *Memorias del IX Foro Nacional de Estadística*. INEGI

Pindyck, R. & Rubinfeld, D. (1992) *Econometric Models and Economic Forecasts*. 3th. edition. Mc Graw Hill.

Ehrenberg, A.S.C. (1975). *Data Reduction*. Great Britain: John Wiley and Sons.

Modelos de Curvas de Crecimiento con Errores No Estacionarios

EVA FERREIRA GARCIA, VICENTE NÚÑEZ ANTÓN

Univ. del País Vasco, España

y

JUAN M. RODRÍGUEZ PÓO

Univ. de Cantabria, España

1. INTRODUCCIÓN

El análisis de curvas de crecimiento, cuando la muestra aleatoria de unidades experimentales (normalmente asociadas a diferentes tratamientos) se observa a través de un periodo de tiempo, no se ha tratado quizá suficientemente en el análisis de regresión.

Si consideramos m unidades experimentales, cada una de ellas con n medidas repetidas de la variable respuesta $y_i(.)$, es común considerar el proceso de generación de datos $y_i(t_j) = f_i(t_j) + \varepsilon_i(t_j)$, con $i = 1, ..., m$, y $j = 1, ..., n$. En este contexto, $f(\cdot)$ se restringe a alguna familia de funciones paramétricas, y el término aleatorio $\varepsilon_i(t_j)$ se supone normalmente distribuido con media cero y errores independientes o, en todo caso, con correlación estacionaria a lo largo del tiempo. La independencia entre las distintas unidades se considera como procedente del proceso de selección de la muestra. Además de estas situaciones, también se han estudiado casos en los que aparecen observaciones incompletas e irregularmente espaciadas.

Existen multitud de trabajos dedicados a relajar las hipótesis usuales. Rosner y Muñoz (1988) analizaron modelos autorregresivos con datos perdidos para observaciones igualmente espaciadas, utilizando un método de estimación basado en regresión no lineal. Jennrich y Schluchter (1986) investigaron el análisis de observaciones completas e igualmente espaciadas con varios tipos diferentes de estructuras de covarianzas, incluyendo modelos autorregresivos de primer orden. Poca atención se ha prestado, sin embargo, a la hipótesis de estacionariedad en la estructura de correlación de los errores. Nuñez-Antón y Woodworth (1994) consideraron una transformación de la escala del tiempo que produce estructuras de covarianzas no estacionarias, donde la estacionariedad es un caso particular. Por otra parte, necesitaron suponer normalidad en los errores así como una forma paramétrica en la curva de regresión, con objeto de derivar los estimadores de los parámetros en la estructura de covarianzas, y su distribución asintótica.

En este trabajo, proponemos un método para derivar las estimaciones de los parámetros de la estructura de covarianzas, sin necesidad de especificar una forma funcional para la curva de regresión, y sin la hipótesis de normalidad en los errores. El método de estimación sigue dos etapas. En primer lugar, se computan los estimadores, por mínimos cuadrados no lineales, de los parámetros en la estructura de covarianzas, obteniéndose resultados de consistencia y normalidad asintótica. En segundo lugar, estimamos no parametricamente la curva de regresión mediante un estimador tipo kernel. En este segundo paso utilizamos los valores estimados en el primero para elegir el parámetro de suavizado en el kernel. La curva así estimada es consistente y la selección del parámetro de suavizado. En la Sección 2

presentamos los métodos de obtención tanto de los estimadores de los parametros de la estructura de covarianzas, como de la curva de regresión. Las demostraciones de los resultados enunciados, así como varios ejemplos ilustrativos realizados mediante simulación, se pueden encontrar en Ferreira, Nuñez y Rodríguez (1994).

2. ESTIMACIÓN DE LA ESTRUCTURA DE COVARIANZAS Y DE LA CURVA DE REGRESIÓN

El modelo que consideramos aquí es similar al utilizado por Azzalini (1984), o por Hart y Wehrly (1986), pero en este seremos capaces de tomar en cuenta posibles casos de no estacionariedad en la estructura de correlación del termino de error. En concreto, consideraremos el modelo

$$y_i(t_j) = f(t_j) + \varepsilon_i(t_j) \tag{1}$$

para $i = 1, ..., m, j = 1, ..., n$, y t_j fijos y uniformemente espaciados en $[0,1]$. La función $f(.)$ es desconocida y debe ser estimada, para lo cual exigiremos que sea al menos dos veces diferenciable. $y_i(.)$ es la variable respuesta y los errores $\varepsilon_i(\cdot)$'s tienen media cero y la siguiente estructura de covarianzas

$$cov\left(\varepsilon_i(t_j), \varepsilon_k(t_l)\right) = \begin{cases} \sigma^2 \rho^{\frac{|t_j^\lambda - t_l^\lambda|}{\lambda}} & \text{si } i = k \\ 0 & \text{si } i \neq k, \end{cases}$$

donde $\sigma^2 > 0$, $0 < \rho < 1$ y $\lambda > 0$ son parámetros que han de ser estimados.

Esta estructura permite correlación a lo largo del tiempo dentro de cada unidad, pero supone que las observaciones para las diferentes unidades son independientes. Nótese que cuando λ es 1 la estructura de correlación es estacionaria. Sin embargo, si λ es diferente de 1, tenemos errores no estacionarios. Por tanto, en este último caso, la correlación depende del orden de las medidas (de forma positiva si $\lambda < 1$ y negativa si $\lambda > 1$).

Procedemos a continuación a estimar los parámetros de la estructura de covarianzas. Para ello definimos

$$\hat{cov}_{ij} = \frac{1}{m-1} \sum_{k=1}^{m} \left(y_k(t_i) - \bar{y}(t_i)\right)\left(y_k(t_j) - \bar{y}(t_j)\right),$$

para $i, j = 1, ..., n$. Definimos también $\hat{C} = \left\{\hat{cov}_{ij}\right\}_{i=1,...,n}^{j=1,...,n}$.

Dado que $E\left(\hat{cov}_{ij}\right) = \sigma^2 \rho^{|t_i^\lambda - t_j^\lambda|/\lambda}$ podemos obtener un estimador del verdadero vector de parámetros $\theta_0 = \left(\sigma_0^2, \lambda_0, \rho_0\right)$ mediante la minimización de la siguiente función criterio respecto de $\theta = (\theta_1, \theta_2, \theta_3) = \left(\sigma^2, \lambda, \rho\right)$

$$Q_n\left(\sigma^2, \lambda, \rho\right) = \frac{1}{n^2} \sum_{i,j} \left(\hat{cov}_{ij} - \sigma^2 \rho^{\frac{|t_i^\lambda - t_j^\lambda|}{\lambda}} \right)^2 \tag{2}$$

El estimador resultante $\hat{\theta}_N = \left(\sigma_N^2, \hat{\lambda}_N, \hat{\rho}_N\right)$ es el estimador por mínimos cuadrados no lineales de $\theta_0 = \left(\sigma_0^2, \lambda_0, \rho_0\right)$, basado en los n^2 valores de $c\hat{o}v_{ij}$. Se puede probar que este estimador cumple las condiciones suficientes para garantizar su consistencia fuerte y su distribución asintóticamente normal (Ferreira, Nuñez y Rodríguez, 1994).

Para estimar la curva de regresión, utilizamos el estimador de regresión no paramétrica de Gasser- Muller (Gasser y Muller, 1979)

$$\hat{f}_h(t) = \frac{1}{h}\sum_{j=1}^{n} \bar{y}\left(t_j\right)\int_{sj-1}^{sj} K\left(\frac{t-s}{h}\right)ds, \tag{3}$$

con $s_0 = 0$, $s_n = 1$ y $s_j = (t_j + t_{j+1})/2, j = 1, ..., n-1$. La función $K(.)$ es un kernel de orden 2, h es la anchura de banda o parámetro de suavizado (Hardle, 1990), y

$$\bar{y}\left(t_j\right) = \frac{1}{m}\sum_{i=1}^{m} y_i\left(t_j\right), \qquad j = 1,...,n.$$

Los siguientes dos teoremas proporcionan algunos resultados acerca del comportamiento de la curva estimada. Del primero se obtienen cotas de convergencia asintóticas de la curva estimada en términos del error cuadrático medio; esto es,

$$MSE\left(\hat{f}_h(t)\right) = E\left(\hat{f}_h(t) - f(t)\right)^2 = BIAS^2\left(\hat{f}_h(t)\right) + VAR\left(\hat{f}_h(t)\right),$$

donde

$$BIAS\left(\hat{f}_h(t)\right) = E\hat{f}_h(t) - f(t),$$

y

$$VAR\left(\hat{f}_h(t)\right) = E\left(\hat{f}_h(t) - E\hat{f}_h(t)\right)^2.$$

Teorema 1. Supongamos que $f(\cdot)$ es una función dos veces continuamente diferenciable, con segunda derivada no nula, y que $0 < t < 1$. Entonces, cuando $n, m \to \infty$ y $h \to 0$,

$$\begin{aligned} MSE\left(\hat{f}_h(t)\right) &= \mathrm{E}\left(\hat{\mathrm{f}}_{\mathrm{h}}(t) - f(t)\right)^2 \\ &= \frac{\sigma^2}{m}\left(1 + h(\log\rho)t^{\lambda-1}C_k\right) + \frac{h^4}{4}f''(t)^2\sigma_k^4 + \\ &\quad \mathrm{o}\left(\mathrm{h}^4\right) + O\left(n^{-2} + m^{-1}n^{-\delta} + h^2n^{-1}\right), \end{aligned} \tag{4}$$

donde

$$C_k = \int_0^1\int_0^1 |r-s|K(r)K(s)drds$$

y

$$\sigma_k^2 = \int_{-1}^{1} u^2K(u)du$$

son dos constantes que dependen únicamente del kernel, y $\delta = \min\{\lambda, 1\}$

Por tanto, para la consistencia puntual de $\hat{f}_h(t)$ solamente necesitamos las condiciones $m, n \to \infty$, $h \to 0$, usuales en estimación no paramétrica (Eubank, 1988; Hardle, 1990).

El siguiente teorema compara la anchura de banda resultante de la minimización de $\mathrm{MSE}(\hat{f}_h(t))$ con las obtenidas anteriormente bajo estacionariedad y bajo errores incorrelados.

Teorema 2. Bajo los supuestos del Teorema 1, y si además $n^{\delta}/m = O(1)$, entonces,

$$MSE\left(\hat{f}_{h_m^*}(t)\right) \leq MSE\left(\hat{f}_{h_{n,m}}(t)\right), \tag{5}$$

donde

$$h_m^*(t) = \left[-\frac{(\log\rho)t^{\lambda-1}C_k\sigma^2}{\left[f''(t)\right]^2\sigma_k^4} \right]^{1/3} m^{-1/3}, \tag{6}$$

y $h_{n,m}$ es la anchura de banda que minimiza $\mathrm{MSE}\left(\hat{f}_h(t)\right)$ bajo errores incorrelados; esto es,

$$h_{n,m}(t) = \left[\frac{\sigma^2 \int K^2(u)du}{\left[f''(t)\right]^2\sigma_k^4} \right]^{1/5} (mn)^{-1/5} \tag{7}$$

En el caso estacionario ($\lambda = 1$), $h_m^*(t)$ coincide con la anchura de banda óptima derivada por Hart y Wehrly (1986),

$$h_m^{**}(t) = \left[-\frac{(\log\rho)C_k\sigma^2}{\left[f''(t)\right]^2\sigma_k^4} \right]^{1/3} m^{-1/3}, \tag{8}$$

habiendo ellos probado que

$$MSE\left(\hat{f}_{h_m^{**}}(t)\right) < MSE\left(\hat{f}_{h_{n,m}}(t)\right). \tag{9}$$

Sin embargo, en el caso no estacionario, $h_m^{**}(t)$ es subóptima, en el sentido de que

$$MSE\left(\hat{f}_{h_m^*}\right) < MSE\left(\hat{f}_{h_m^{**}}(t)\right), \tag{10}$$

siendo $h_m^*(t) > h_m^{**}(t)$ si la correlación crece con el orden temporal de las medidas ($\lambda < 1$), y lo contrario si la correlación decrece ($\lambda > 1$). En definitiva, la selección del suavizado en datos temporales correlados, si se ignora la estructura no estacionaria de las covarianzas, puede producir bien estimaciones subsuavizadas o sobresuavizadas de la curva de regresión. Considerando medidas globales del error tomamos

$$MISE\left(\hat{f}_h\right) = \int_0^1 E\left(\hat{f}_h(t) - f(t)\right)^2 dt,$$

con lo que la anchura de banda óptima global, definida como el valor de h que minimiza *MISE*, será

$$h_m^* = \left[-\frac{(\log\rho)C_k\sigma^2}{\lambda\int_0^1 \left[f''(t)\right]^2 dt\sigma_k^4} \right]^{1/3} m^{-1/3} \tag{11}$$

De nuevo, ignorar la no estacionariedad si ese es el caso, puede llevar a estimaciones erróneas de las curvas.

Teniendo en cuenta que la anchura de banda óptima depende de derivadas de la función desconocida, esta anchura no puede ser utilizada en la práctica. Aquí proponemos un criterio basado en Rice (1984) y elegiremos el valor de h, denotado por $\hat{h}_m$, como el valor que minimiza la función criterio

$$M(h) = RSS(h) - \frac{\hat{\sigma}^2}{m} + \frac{2\hat{\sigma}^2}{nmh}\sum_{i,j}^{n} \hat{r}_{ij}\int_{s_{j-1}}^{s_j} K\left(\frac{t_i - s}{h}\right)ds, \tag{12}$$

donde

$$RSS(h) = \frac{1}{n}\sum_{i=1}^{n} \left(\bar{y}(t_i) - \hat{f}_h(t_i)\right)^2,$$

y

$$\hat{r}_{ij} = \sigma^2\hat{\rho}^{\frac{\left|t_i^{\hat{\lambda}} - t_j^{\hat{\lambda}}\right|}{\hat{\lambda}}},$$

siendo $\hat{\sigma}^2, \hat{\rho}$ y $\hat{\lambda}$, los parámetros estimados anteriormente. La elección de esta función criterio esta basada en el hecho de que

$$E\left[RSS(h)\right] = \frac{\sigma^2}{m} + MISE(h) - \frac{2}{mnh}\sum_{i,j}^{n} \sigma^2\rho^{\frac{\left|t_i^{\lambda} - t_j^{\lambda}\right|}{\lambda}} \int_{s_{j-1}}^{s_j} K\left(\frac{t_i - s}{h}\right)ds. \tag{13}$$

Además, es posible probar que el valor así seleccionado $\hat{h}_m$, = $arg\ \min_h M(h)$ proporciona las mismas propiedades asintóticas hasta el segundo orden que la anchura de banda optima h_m^* arg $\min_h MSE(h)$ (Ferreira, Nuñez y Rodríguez, 1994).

REFERENCIAS

Azzalini, A. (1984). Estimation and Hypothesis Testing for Collections of Autoregressive Time Series. *Biometrika*, **71**, 85-90.

Eubank, R. (1988). *Spline Smoothing and Nonparametric Regression*. New York, Marcel Dekker.

Ferreira, E., Nunez, V. y Rodr¡guez, J. (1994). Regression Models with Nonstationary Errors. *Documento de Trabajo* DT94.12, Universidad del Pais Vasco, Bilbao, Espana.

Gasser, T. y Muller, H.G. (1979). Kernel Estimation for Regression Functions. En T. Gasser and M. Rosenblatt, editors, Smoothing Techniques for Curve Estimation, 23-68. *Heidelberg: Springer-Verlag.*

Hardle, W. (1990). *Applied Nonparametric Regression. Cambridge*: Cambridge University Press.

Hart, J.D. y Wehrly, T.E. (1986). Kernel Regression Estimation Using Repeated Measurements Data. *Journal of the American Statistical Association*, **81**, 1080-1088.

Jennrich, R.L. y Schluchter, M.D. (1986). Unbalanced Repeated-Measures Models with Structured Covariance Matrices. *Biometrics*, **42**, 805-820.

Nunez Anton, V. y Woodworth, G.G. (1994). Analysis of Longitudinal Data with Unequally Spaced Observations and Time Dependent Correlated Errors. *Biometrics*, **50**, 445-456.

Rice, J. (1984). Bandwidth Choice for Nonparametric Regression. *The Annals of Statistics*, **12**, 1215-1230.

Rosner, B. y Munoz, A. (1988). Autoregressive Modelling for the Analysis of Longitudinal Data with Unequally Spaced Examinations. *Statistics in Medicine*, **7**, 59-71.

Muestrea: Herramienta Computacional en el Salón de Clase

LILIANA FIGUEROA y RUBEN HERNÁNDEZ C.

ITAM, México

1. INTRODUCCIÓN

En las últimas décadas mucha atención se ha dedicado al desarrollo de software que facilite los cálculos necesarios en todas las ramas de la Estadística. No obstante, una sección que ha sido continuamente descuidada es la pedagógica. Sin duda alguna, esta situación resulta grave si consideramos que cada vez más personas de todas las disciplinas están tomando cursos de Estadística durante sus estudios profesionales y muchas veces la teoría les parece demasiado complicada, más aún cuando no han tenido una sólida formación matemática.

Preocupados por la ausencia de herramientas computacionales que apoyen la enseñanza estadística, hemos decidido dar un paso en esta dirección al crear una rutina para asistir a los profesores en la impartición de un curso clásico de una de sus ramas, el Muestreo.

Al considerar que la impresión visual que un fenómeno deja en el ser humano es generalmente el mejor medio para asimilarlo, nuestra propuesta tendrá como base el uso de imágenes. Esta idea se concretará en la rutina *Muestrea*, cuyo propósito será tomar una muestra de puntos de una imagen mediante distintos esquemas de muestreo y con distintos tamaños de muestra. A partir del conjunto de puntos elegido, intentaremos recuperar la imagen original por medio de un criterio de inferencia sobre las distancias entre los puntos de la imagen.

A través de esta rutina, un profesor será capaz de presentar la forma en que se aplica el muestreo a una población finita, de modo que el alumno pueda visualizar el proceso; podrá, además, mostrar cómo se ve esta muestra, qué información puede extraerse de ella y hará, finalmente, comparaciones entre distintos esquemas al haber sido aplicado el criterio de inferencia.

Los esquemas de muestreo que se consideran en la rutina son:

1. Muestreo Aleatorio Simple (con y sin restitución).
2. Muestreo Estratificado.
3. Muestreo por Conglomerados.
4. Muestreo Sistemático.

Debe ser claro que con nuestra propuesta no pretendemos reemplazar los textos o las clases de muestreo, sino que *Muestrea* debe ser tomada sólo como un complemento.

2. ESPECIFICACIONES Y REQUERIMIENTOS DE MUESTREA

La rutina fue programada en el lenguaje Turbo Pascal para Windows v. 1.5, seleccionado por su velocidad y flexibilidad, así como por la numerosa cantidad de funciones que posee para manipular imágenes y por presentarse en el ambiente Windows, que nos permitirá hacer una presentación más agradable.

Muestrea fue escrito usando una PC HP Vectra 486 operando con el sistema MS-DOS v. 60.0 bajo la versión 3.1 de Windows. Tiene 1800 líneas de código divididas en un archivo ejecutable (muestrea.exe) y cinco unidades (Ayudabus, Info, Auxiliar, Nuevo2 y Métodos);

utiliza, además, un archivo de recursos (muestrea.res) que contiene las imágenes, los cuadros de diálogo y los menús, con lo que reúne, en total, 394,016 bytes. Las imágenes usadas tienen el formato bitmap y fueron tomadas del acervo de Paintbrush.

Para ser ejecutado, el programa requiere que cada uno de los archivos mencionados anteriormente estén presentes. Los requerimientos mínimos de hardware y software son una computadora 386 con 4MB de RAM, adaptador de gráficos (Hercules, EGA o VGA), un ratón y la versión 3.1 de Windows.

3. ESTRUCTURA DE *MUESTREA*

El programa cuenta con el formato típico de las aplicaciones de Windows, es decir, una ventana con un menú de control, botones para maximizar, minimizar o restablecer la ventana, una barra de título y una barra de menú (Imagen 1). El menú principal tiene 2 opciones: Muestreo y Ayuda.

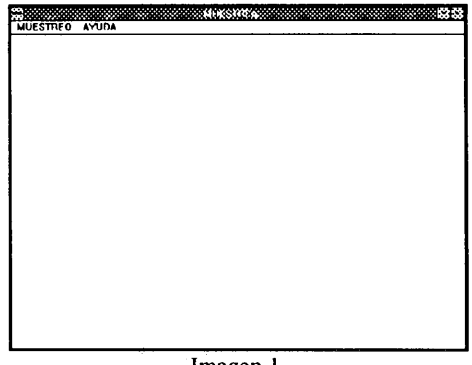


Imagen 1.

Ayuda tiene a su vez 3 opciones:

Contenido despliega en pantalla una ventana con una breve explicación del objetivo del programa y la forma en que pretendemos alcanzarlo.

Buscar es un diccionario. Cuando se selecciona esta opción aparece una ventana con una lista de términos estadísticos utilizados a lo largo del trabajo y que servirán para orientar al usuario cuando tenga que proporcionar algún dato o hacer una elección dentro de un conjunto de opciones. Para elegir el concepto cuya explicación se desea, debe hacerse click (presionar el botón izquierdo del ratón) sobre el término y presionar el botón Despliega. Una breve explicación aparecerá en el recuadro inferior. Cuando se termine de usar deberá presionarse el botón Cerrar. La opción buscar está disponible en todas las etapas de la rutina.

Acerca de... proporciona alguna información relevante del programa.

Muestreo cuenta únicamente con la opción **Nuevo**, que será la necesaria para comenzar el muestreo y recuperación de una imagen. Al seleccionarla aparece en pantalla un menú de imágenes preestablecidas y en la esquina inferior derecha se encontrarán dos botones: *Cerrar* y *Siguiente* (Imagen 2). Si el usuario presiona el botón Cerrar en cualquier momento, se borrará completamente la pantalla y podrá comenzarse un nuevo muestreo. Una observación importante que debemos hacer es que si se desea hacer un nuevo muestreo cuando ya se ha comenzado uno, deberá seleccionarse primero *Cerrar* y después la opción **Nuevo**. En caso contrario, el programa no correrá adecuadamente. El botón *Siguiente* es el

encargado de llevar paso a paso la rutina, con lo que el usuario puede aprovechar el programa con todo su potencial sin complicaciones, al evitar que se extravíe en la elección de múltiples menús. Cuando se desee pasar a la siguiente etapa sin haber completado lo requerido para el buen funcionamiento de *Muestrea*, aparecerá un cuadro de diálogo que indicará qué debe hacer (Imagen 3).

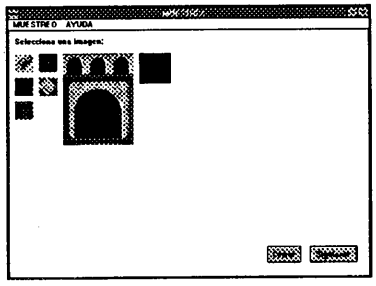


Imagen 2.

Imagen 3.

El usuario escogerá una de las imágenes haciendo click sobre cualquier punto en ella. Una vez hecho esto, la imagen aparecerá en la esquina superior izquierda de la pantalla, borrándose todo lo demás, y se podrá ir al siguiente paso. A continuación el programa

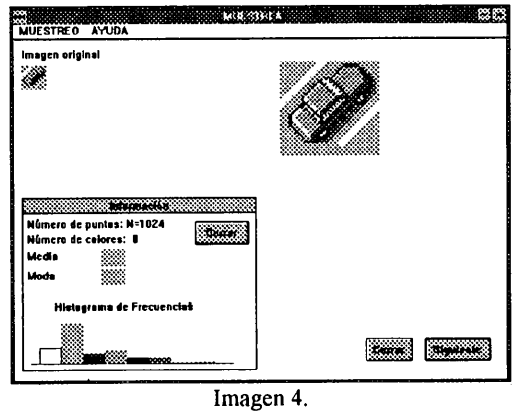


Imagen 4.

despliega la imagen ampliada y una ventana con información sobre ella (Imagen 4). La información que hemos incluído es el número de pixeles (puntos) que conforman a la imagen, el número de colores que contiene, la media del color, la moda y un histograma de frecuencias de los colores.

El siguiente paso es escoger uno de los métodos por medio de los cuales se extraerá una muestra de puntos y el tamaño de muestra (si el esquema elegido es el Sistemático, deberá indicarse el intervalo de muestreo; si el esquema es el de muestreo por conglomerados, deberá indicarse el número de conglomerados a seleccionar).

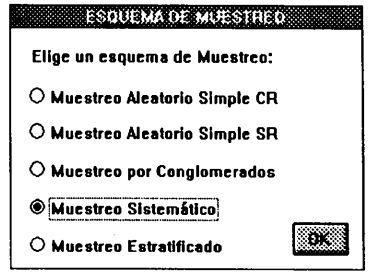


Imagen 5.

La información obtenida de la muestra será la posición de cada punto y su color. Lo siguiente es observar la pantalla la muestra y, después de algunos segundos, comenzará la recuperación de la imagen (Imagen 6). Por restricciones de espacio, el tamaño de muestra no podrá ser mayor a 3600 puntos ni exceder el número total de pixeles en la imagen.

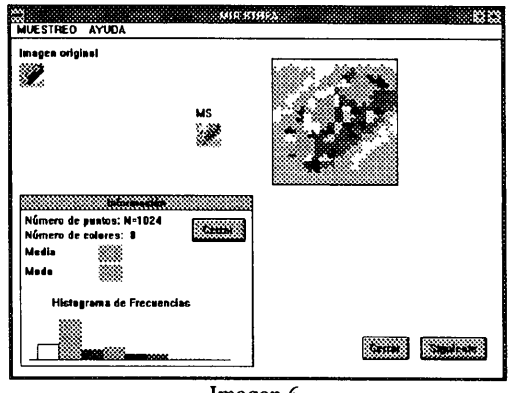


Imagen 6.

Como hemos mencionado antes, el criterio de inferencia que hemos utilizado está basado en las distancias. Esto es, para cada punto que no aparece en la muestra calculamos las distancias entre él mismo y los puntos de la muestra. El punto de la muestra más cercano es el que determina el color del punto, asignándosele exactamente el mismo color.

El usuario captará en pantalla la reconstrucción de la imagen punto a punto. Para esto hemos hecho también una ampliación tanto de la muestra como de la nueva imagen. Una vez completada la reconstrucción, aparece una nueva ventana que proporciona la información dada antes, pero ahora sobre la imagen que hemos obtenido. Con esto podemos comparar, al menos en parte, qué tan buena resultó nuestra elección. Esta nueva ventana se encuentra debajo de la ventana de información que apareció al principio.

Así, cuando se han seleccionado varios esquemas de muestreo para hacer una recuperación de la imagen, la pantalla de la ventana principal se verá como en la Imagen 8. En este ejemplo hemos seleccionado el bitmap del auto que presentamos en la Imagen 4. Hemos tomado un tamaño de muestra n=100 (Imagen 7) para cada uno de los esquemas. En el caso del esquema por conglomerado seleccionamos 25 de ellos ya que la población fue dividida en grupos de 2×2 pixels. Para el esquema de muestreo sistemático tomamos un intervalo de muestreo k = 3, el tipo de muestreo sistemático elegido fue el no alineado, siguiendo el enfoque de Quenouille. La estratificación de la imagen se hizo en cuadrantes ya

que el criterio de inferencia trabaja sobre ellos, con lo que tenemos control local; la asignación fue proporcional.

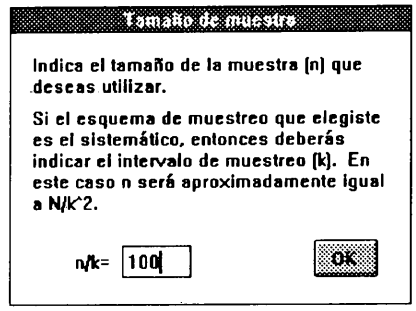


Imagen 7.

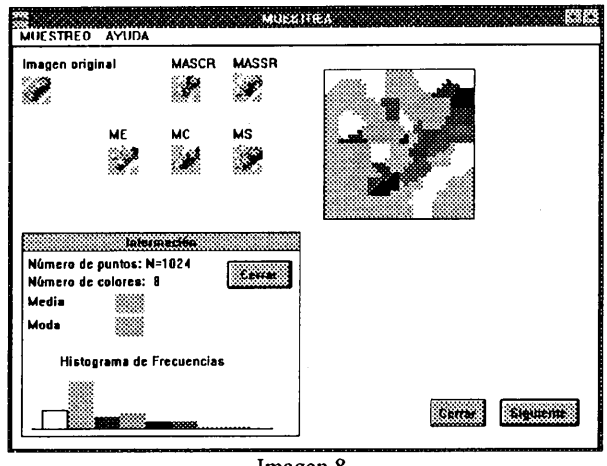


Imagen 8.

Pero *Muestrea* no termina aquí y al oprimir nuevamente el botón *Siguiente* tenemos la posibilidad de escoger un nuevo método de muestreo. La reconstrucción de la imagen no se empalmará sobre la primera a menos de que el esquema sea el mismo. Este proceso puede seguir varias veces.

4. SUGERENCIAS PARA EL BUEN FUNCIONAMIENTO DE *MUESTREA*

Para un mejor aprovechamiento del programa sugerimos al usuario que tenga, en todo momento, maximizada la ventana de aplicación. Al tener varias imágenes, cuadros de diálogo y ventanas sobre la ventana principal, se tendrá una mejor visión y cobertura del proceso en esta forma.

Una recomendación adicional es no tratar de modificar el tamaño de la ventana o minimizarla cuando se hayan efectuado ya algunas recuperaciones de la imagen y se desee conservarlas en pantalla para hacer comparaciones. Si, por ejemplo, en cierto momento se minimiza la ventana, las imágenes de recuperación desaparecerán de la pantalla, ya que no han sido grabadas en forma alguna en la memoria de la máquina.

Aunque el propósito principal de la rutina es el muestreo en sí, *Muestrea* puede servir también para aprender algo sobre Estadística Descriptiva. En las primeras etapas del

programa se despliega una ventana con información que describe a la imagen elegida, así que si el propósito es solamente observar esto, pueden seguirse los pasos iniciales y observar, para cada imagen, el color medio, el color nada y el histograma de frecuencias de los colores. Más tarde se puede elegir el botón *Cerrar* para repetir el proceso con otra imagen.

Como última observación, tenemos que el programa ha sido escrito de forma tal que cualquier imagen en el formato bitmap puede integrarse para ser muestreada si se hacen los cambios pertinentes tanto en el archivo de recursos como en el código del programa.

5. SUGERENCIAS PARA EL DESARROLLO FUTURO DE *MUESTREA*

Como toda primera versión de un trabajo, éste tiene huecos y limitaciones que deberán ser superadas. El usuario notará que el ambiente de la rutina resulta todavía restringido, por lo que en el futuro quizás se deba trabajar en la reducción de esta condición. Sin embargo, la versión actual de *Muestrea* puede ser ya usada en los cursos.

A continuación presentamos una serie de elementos que consideramos importante integrar al programa en el futuro con el objeto de dar mayor libertad al usuario, además de brindarle una presentación más completa del tema, a saber:

1. Otorgar al usuario la libertad de elegir el tamaño y la forma de los conglomerados.
2. Permitir de alguna forma la definición de estratos.
3. Implementar la realización de muestreos multietápicos con selección de distintos esquemas en cada etapa.
4. Crear intervalos de confianza para los estimadores.
5. Incluir funciones de costo que podrían tomar en cuenta el tiempo máquina.

REFERENCIAS

Cochran, William G. (1963). *Técnicas de Muestreo*, CECSA.

Hansen, M. H., Hurwitz, W.N., Madow, W.G. (1953*). Sample Survey Methods and Theory I.* Rnd. edn. New York: John Wiley & Sons, Nueva York.

Hansen, M. H., Hurwitz, W. N., Madow, W.G. (1953*). Sample Survey Methods and Theory II.* Rnd. edn. New York: John Wiley & Sons, Nueva York (2da. ed.)

Kish, Leslie (1965). *Survey Sampling.* John Wiley & Sons, Nueva York

Krishnaiah, P.R., Rao, C.R., eds,. (1988). *Handbook of Statistics, Vol. 6.* Amsterdam: Elsevier Science Publishers B. V.

Raj Des (1972). *The Design of Sample Surveys.* McGraw-Hill, Nueva York

Raj, Des (1968). *Sampling Theory.* Nueva York: McGraw-Hill.

Sukhatme, P.V., Sukhatme, B.V. (1954). *Sampling Theory of Surveys with Applications.* Iowa State University Press.

Evolución de la Mortalidad Infantil en México hasta 1990

GEORGINA Y. GALLARDO HURTADO

Colegio de México, México

1. INTRODUCCIÓN

Las deficiencias de la información de las estadísticas vitales en cuanto al registro de los nacimientos y defunciones en el primer año de vida, han sido una permanente preocupación por los sesgos que arrojan en la estimación de la mortalidad infantil en México Mina (1984 y 1992).

Tradicionalmente se utilizan métodos directos para medir la fecundidad y la mortalidad, pero al haber subregistro de las defunciones, así como de los nacimientos, se tiene una subestimación de los niveles de la mortalidad infantil en México.

El uso de técnicas indirectas nace de la necesidad de solucionar este problema para poder dar apreciaciones más reales de lo que sucede con la mortalidad y la fecundidad en países con errores en la captación de información acerca de nacimientos y defunciones Hill (1984).

La base para la estimación no convencional de la mortalidad infantil son reportes de las mujeres respecto de los nacimientos de hijos vivos y la sobrevivencia de esos hijos. Naciones Unidas (1986). La forma más simple de estimación requiere solamente del número total de hijos nacidos vivos por mujer y el número de ellos que aún están vivos en el momento del censo o la encuesta. Hill (1984). En el censo de 1980 se captan por primera vez el número de hijos nacidos vivos y el número de hijos sobrevivientes por mujer por grupo quinquenal de edad lo que permite la estimación de tasas de mortalidad y probabilidades de muerte infantil aplicando diferentes métodos indirectos como son los de Brass, Trussell, Sullivan y Feeney.

Para efectos de este trabajo la República Mexicana se dividió en cinco regiones:
Región Caribe comprendiendo los estados de *Campeche, Quintana Roo y Yucatán.* **Región Centro** formada por los estados de *Aguascalientes, Distrito Federal, Durango, Guanajuato, Hidalgo, México, Morelos, Querétaro, Puebla, San Luis Potosí, Tlaxcala y Zacatecas.* **Región Frontera** conformada por *Baja California, Coahuila, Chihuahua y Nuevo León.* **Región Golfo** formada por los estados de *Veracruz, Tamaulipas y Tabasco.* **Región Pacífico** formada por los estados de *Baja California Sur, Chiapas, Colima, Guerrero, Jalisco, Michoacán, Nayarit, Oaxaca, Sinaloa y Sonora.*

El objetivo de este trabajo es estimar niveles, tendencias y diferenciales de la mortalidad en México utilizando información censal, particularmente la captada en 1980 y 1990 a través de las preguntas sobre hijos nacidos vivos e hijos sobrevivientes de las mujeres de 15 a 49 años cumplidos.

2. FUENTES DE DATOS

Se utilizaron el X Censo de Población y Vivienda 1980 y el XI Censo de Población y Vivienda 1990, ambos en su sección referida al número de hijos nacidos vivos y número de hijos sobrevivientes por grupo de edad de las mujeres censadas así como la información del número de mujeres por grupo de edad y por edad individual.

3. METODOLOGÍA

Los posibles errores en la declaración de la edad son un factor crucial del análisis en cualquier estudio de la población. Nuñez (1984). Esta información se corrige aplicando la fórmula de graduación de 1/16.

Para evaluar la calidad de la información en cuanto a la distribución de las edades, se aplicó el método de Whipple. Este método estima el grado de preferencia hacia los dígitos 0 y 5 por la población censada que declaró su edad entre los 23 y 62 años.

Los supuestos en que se basa Brass (1974), son el de que se trabaja con una población cerrada, la fecundidad y la mortalidad son estables, se espera que la declaración de las edades sea correcta y que no haya omisión en la declaración del número de hijos vivos y sobrevivientes.

Brass asocia a la proporción de hijos fallecidos una probabilidad de morir de los 0 a los n años, dependiendo del grupo de edad de las mujeres es la n que asocia.

Los métodos de Sullivan y Trussell se basan en el de Brass con ligeras variaciones en la estimación de nq0.Feeney modifica los supuestos de Brass ya que permite cambios en la mortalidad (no necesariamente estable) y obtiene *Tasas de Mortalidad Infantil* ubicadas en el tiempo, en los años previos al censo.

A partir de las probabilidades de muerte y de las tasas de mortalidad infantil pueden obtenerse las demás series de una tabla abreviada de vida.

4. RESULTADOS

Los tres métodos se aplicaron para ambos censos utilizando en ambos casos, y cuando resultó pertinente, las regiones sur y oeste de las tablas modelo de Coale-Demeny.

En 1980, al tomar el patrón oeste, las tasas de mortalidad infantil más altas se presentan en la región Centro en donde 155 niños de entre 0 y 1 años exactos fallecen por cada mil nacidos vivos mientras que en la región Frontera únicamente 73 de cada mil. La media nacional se encuentra entre 72 fallecidos por mil y 100 fallecidos por mil combinando todos los grupos y técnicas. De acuerdo a la técnica de Feeney se obtienen fechas de entre 2.3 y 15.5 años anteriores al censo.

En las estimaciones para 1990, tomando el patrón oeste, la tasa más alta la presenta la región Pacífico para el grupo 45-49 años, esta es de 81 por mil y para el primer grupo de edad de 54 por mil. La región Frontera muestra las tasas más bajas. Los años anteriores al censo que se obtienen en 1990 se encuentran entre 2.4 y 15.6 años para el primer y último grupo de edad respectivamente.

5. ESTIMACIÓN DE LA ESPERANZA DE VIDA AL NACER

En 1980, en la República Mexicana la esperanza de vida al nacer en el grupo de edad de las mujeres de 20-24 años, la cual se utilizará para obtener la tabla de vida, se encuentra entre 57.4 años y 58.7 años. En 1990 las estimaciones para la República Mexicana van de 72.1 años a 74 años

6. CONCLUSIONES

En 1980 había una diferencia de 1.22 hijos entre hijos nacidos vivos e hijos sobrevivientes en la República Mexicana, para el último grupo de edad, en 1990 esta diferencia se redujo a

0.58 hijos solamente. Las tasas de mortalidad infantil se reducen en una cuarta parte para la República Mexicana que de presentar una tasa de 105 por mil en el grupo 45-49 años, pasa a una tasa de 78 por mil.

En cuanto a las esperanzas de vida al nacer, hay una ganancia considerable en todas las regiones consideradas. La República Mexicana pasó de una esperanza de vida de 58.7 años a una de 74 años, gana en sólo diez años 15.3 años. La región Centro pasa de 48.1 años a 64.2 años, la región Frontera pasa de 61.2 años a 76.5 años y la región Pacífico de 55 años a 62.7 años.

REFERENCIAS

Brass, W. (1973). A critique of methods for estimating population growth iin countries with limited data, *Laboratories for Population Statistics*, Reprint Series no. 4.

Brass, W. (1977) *Cuatro Lecciones de...*, Centro Latinoamericano de Demografía, Santiago de Chile.

Brass, W. (1974). *Métodos para estimar la fecundidad y la mortalidad en poblaciones con datos limitados*. Selección de trabajos de William Brass, Centro Latinoamericano de Demografía.

Corona, V. R. (1991). Confiabilidad de los resultados preliminares del XI Censo General de Población y Vivienda de 1990, *Estudios Demográficos y Urbanos*, Vol. **6**, núm. 1, pp. 33-68.

Feeney, G. (1983). *Estimación de la mortalidad infantil y de la niñez en condiciones de mortalidad variable*, Centro Latinoamericano de Demografía, San José, Costa Rica.

Feeney, G. (1977*). Estimación de parámetros demográficos a partir de información censal y de registros*, Centro Latinoamericano de Demografía, Santiago de Chile.

Hill, K. (1984). *An evaluation of indirect methods for estimating mortality, Methodologies for the collection and analysis of mortality data*, edited by J. Vallin, J. Pollard and L. Heligman, Liège, Belgium: Ordina Editions.

Hill, K. (1991). Approaches to the measurement of childhood mortality: a comparative *review Population Index*, vol. **57,** no. 3, 1991. pp. 368-382

INEGI (1985). X Censo General de Población y Vivienda 1980, Resumen General.

INEGI (1990). XI Censo General de Población y Vivienda 1990, Resumen General

Mina, V.A. (1992). Curso básico de demografía, Serie de notas de clase, Vínculos Matemáticos no. **118**-1992. *Publicaciones del Departamento de Matemáticas* de la Facultad de Ciencias UNAM. Tercera edición.

Mina, V.A. (1992). Elaboración y utilidad de la tabla abreviada de mortalidad, Serie: Notas de clase, Vínculos matemáticos # 138-1992, *Publicaciones del Departamento de Matemáticas de la Facultad de Ciencias, UNAM*, tercera edición.

Mina, V.A. (1984). La medición indirecta de la mortalidad infantil y en los primeros años de vida en México, *Seminario La mortalidad en México: Niveles, tendencias y*

determinantes. Centro de Estudios Demográficos y de Desarrollo Urbano, El Colegio de México.

Mina, V.A. (1994). Notas de análisis demográfico Sserie: notas de clase, Vínculos Matemáticos no. 203, 1994 *Publicaciones del Departamento de Matemáticas de la Facultad de Ciencias, UNAM.*

Mina, V.A. (1992) Notas de demografía matemática, El Colegio de México enero de 1992.

Naciones Unidas, Manual X (1986) Técnicas indirectas de estimación demográfica, Departamento de Asuntos Económicos y Sociales Internacionales, Estudios de Población, no. 81, Naciones Unidas Nueva York.

Nuñez, L. (1984). Una aproximación al efecto de la mala declaración de la edad en la información demográfica recabada en México, Secretaría de Gobernación, Dirección General del Registro Nacional de Población e Identificación Personal.

Ordorica, M. y Medina, V. (1986). Evaluación de la información censal sobre fecundidad. Instituto Nacional de Estadística, Geografía e Informática, México.

Pressat, R. (1983). El análisis Demográfico, fondo de Cultura Económica, México.

Proyecto Sedesol. Poblamiento de las zonas costeras de México, (en prensa).

Rabell, C. A. y Mier y Terán, R.M. (1986). El descenso de la mortalidad en México de 1940 a 1980 Estudios Demográficos y Urbanos 1 (1): pp 39-72

Tapinos, G. (1988). *Elementos de Demografía* Espasa-Calpe S.A., Madrid.

Confiabilidad de Items Cuya Degradación se Modela a Partir de una Longitud de Iniciación

CHRISTIAN GARRIGOUX

ITESM, Campus Monterrey, México

1. INTRODUCCIÓN

El desarrollo de un defecto en un componente (o en un sistema) sometido a cierto estrés puede llevar a una falla. Existen dos estrategias principales para manejar el control de la confiabilidad en tales casos.

(a) Se puede retirar el componente de servicio en un tiempo predeterminado tal que la probabilidad de falla antes de este tiempo sea aceptablemente baja. Esta estrategia es estándar para componentes para los cuales las inspecciones son destructivas. Sin embargo, es ineficiente usarla cuando se pueden efectuar inspecciones no destructivas pues se retiran muchos componentes que podrían quedar en servicio por más tiempo con un riesgo calculado.

(b) La otra estrategia consiste precisamente en efectuar inspecciones no destructivas sobre los componentes mientras degradan de manera que se determina si pueden seguir sin mayor riesgo hasta la inspección siguiente o si se deben retirar de inmediato.

Varias tecnologías nuevas permiten efectuar pruebas no destructivas cada vez más precisas. Por lo tanto, resulta de interés estudiar las propiedades y ventajas que presenta la segunda estrategia, es decir cómo el uso de pruebas no destructivas puede ayudar a mantener un nivel de confiabilidad adecuado. Este problema ha sido estudiado bajo diferentes enfoques y está presentado en varios libros, por ejemplo, Jorgenson, McCall y Radner (1967), Barlow y Proschan (1981), Anders (1990) etc. El trabajo presentado en este artículo está desarrollado en base a un modelo matemático descrito con detalles en Garrigoux y Meeker (1994).

A continuación se describen, a grandes rasgos, las características de dicho modelo.

(a) Se asume una distribución de probabilidad conocida para el nivel inicial de degradación, nivel medido por un solo parámetro en el tiempo cero.

(b) El nivel de degradación aumenta a través del tiempo según una ley de degradación conocida. En la ausencia de inspecciones, el item se retira de servicio, sea porque se alcanzó un nivel donde ocurre una falla, sea porque el item alcanzó su vida útil de funcionamiento u horizonte.

(c) La ley de degradación está descrita por un vector de parámetros aleatorios (de un item a otro), pero fijos a través del tiempo. Su distribución conjunta de probabilidad es conocida.

(d) Mientras está en servicio, el item se inspecciona para determinar su nivel de degradación. Debido a los posibles errores de medición, el nivel de degradación observado es una variable aleatoria cuya distribución de probabilidad condicionada sobre el nivel real desconocido de degradación es conocida.

(e) Las inspecciones se pueden efectuar en unos tiempos predeterminados llamados oportunidades para inspección. Si se observa un nivel de degradación inaceptable, el item es retirado de servicio.

En ese modelo, se obtienen fórmulas para medidas de confiabilidad como la distribución de probabilidad para los tiempos de falla, las probabilidades de rechazo en inspecciones y la tasa instantánea de fallas. Un programa computacional escrito en el lenguaje S (Becker, Chambers y Wilks 1985) con interfase con Fortran evalúa estas fórmulas y genera gráficas que facilitan el análisis de confiabilidad. Ciertos aspectos matemáticos relacionados con esas fórmulas bajo diferentes enfoques se pueden encontrar en artículos como Kitagawa e Hisada (1997), Harris y Lim (1983), Yang y Chen (1985).

La generalidad de la formulación matemática y del programa computacional anteriores permiten su uso en aplicaciones físicas de diferentes naturalezas así como extensiones a otras variantes del problema de degradación. En la sección siguiente, se comentarán las modificaciones necesarias para adecuar la matemática y su implementación computacional al caso donde la degradación se modela a partir de un mismo nivel, pero en diferentes tiempos en vez de empezar con diferentes niveles de degradación en un mismo tiempo (ver fig. 1 y 2). La motivación para esta modificación proviene de un caso de degradación por fatiga mecánica donde la distribución de probabilidad para el nivel de degradación inicial (en este caso el tamaño inicial de una grieta en cierto componente metálico) no puede ser conocida poruqe la longitud inicial de los defectos está por debajo del nivel de detección. Sin embargo, existe una longitud típica llmada de iniciación a partir de la cual no solamente el defecto es observable, sino que su ley de degradación es conocida (ley de París). Se observa el componente y se mide cuánto tiempo transcurre hasta observar una grieta con longitud igual a la de iniciación. Con tales observaciones, se construye experimentalmente la distribución de probabilidad del tiempo durante el cual un componente debe estar en servicio

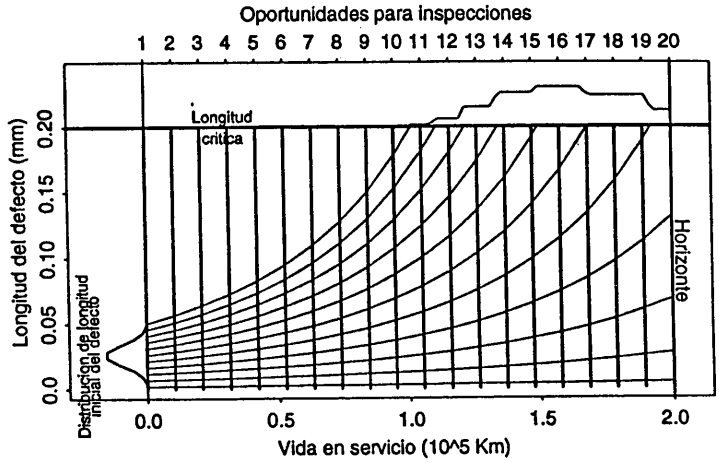


Figura 1. Curvas de degradación a partir de una distribución de longitud inicial del defecto.

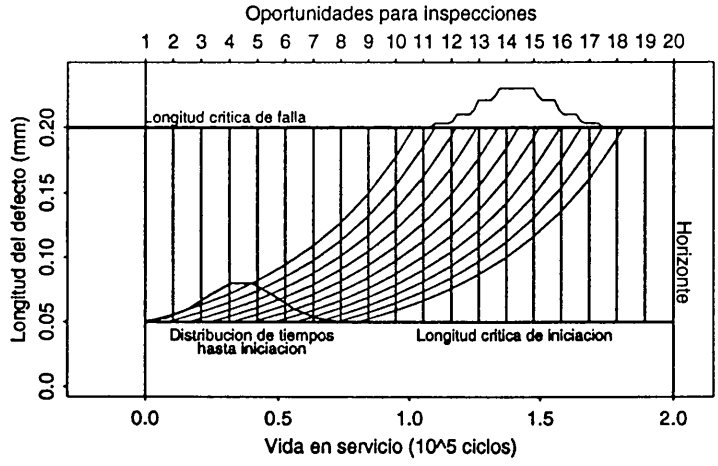


Figura 2. Curvas de degradación a partir de una distribución de tiempos hasta iniciación.

para que grietas o defectos de manufactura indetectables en la puesta en servicio se transformen en grietas con longitud de iniciación. A partir de ese momento, la grieta es observable y su degradación se puede modelar matemáticamente. Las características de confiabilidad que se requieren evaluar siguen siendo claramente la densidad de probabilidad de tiempos de falla; las probabilidades de rechazo o aceptación en inspecciones; y la tasa instantánea de fallas, cuya obtención es directa a partir de la densidad de probabilidad de los tiempos de falla.

2. EL NUEVO MODELO DE DEGRADACIÓN

2.1 Modificaciones necesarias en la formulación matemática

Presentamos aquí las nuevas fórmulas que permiten los cálculos de probabilidad para fallas y rechazos en inspecciones. Acompañamos estas fórmulas con unas explicaciones/interpretaciones intuitivas y referimos a quienes estén interesados en los detalles matemáticos a Garrigoux y Meeker, (1994)

2.2 Densidad de probabilidad de falla.

La fórmula para la densidad de probabilidad de los tiempos de falla es:

$$g_T(t) = \begin{cases} \int_{D(\underline{w})} g\left[\tau_{\underline{w}}(t)\right] \dfrac{d\tau_{\underline{w}}(t)}{dt} \prod_{j=I(\tau_{\underline{w}}(t))}^{j=i} P_A\left[\underline{w}, \tau_{\underline{w}}(t), t_j\right] dG(\underline{w}), & t_i < t < t_{i+1}, \quad i = 1,\ldots,N \\ \int_{D(\underline{w})} g\left[\tau_{\underline{w}}(t)\right] \dfrac{d\tau_{\underline{w}}(t)}{dt} dG(\underline{w}), & t_0 < t < t_1 \end{cases}$$

donde $g_t(t)$ representa de probbilidad de la variable aleatoria tiempo de servicio T evalúada en t.

τ se llama tiempo de iniciación, es el tiempo en que el defecto alcanza el nivel de iniciación y $\tau_w(t)$ es el valor de τ tal que el nivel de falla se alcanzará en el tiempo t, dado que el vector de parámetors aleatorios $\underline{W}$ que describe la velocidad de degradación toma valores $\underline{w}$. $g(\tau)$ denota la densidad de probabilidad de τ, D($\underline{w}$) el dominio de valores $\underline{w}$ y G($\underline{w}$) la distribución conjunta acumulada de $\underline{w}$.

Finalmente, P_A simboliza la probabilidad de aceptación del componente en una inspección. Esta probabilidad depende del tamaño real del defecto que depende a su vez del tiempo de iniciación τ, de la velocidad de degradación descrita a través de $\underline{w}$ y del tiempo de inspección (t_j para la j ésima inspección). $l\left[\tau_{\underline{w}}(t)\right]$ representa el índice de la primera inspección efectuada sobre un componente después del tiempo de iniciación $\tau_{\underline{w}}(t)$ de su grieta.

2.3 Probabilidad de rechazo en la i-ésima inspección

La probabilidad de rechazo en la i-ésima inspección del esquema preestablecido de inspecciones, denotada $Pr(T = t_i)$ es dada por

$$Pr(T=t_i)=\begin{cases}\displaystyle\int_{D(\underline{w})}\int_{\tau_{\underline{w}}(t_i)}^{t_i}\prod_{j=l(\tau)}^{i-1}P_A[\underline{w},\tau,t_i]g(\tau)d\tau dG(\underline{w}), & i=2,\ldots,N\\ \displaystyle\int_{D(\underline{w})}\int_{\tau_{\underline{w}}(t_1)}^{t_1}P_R[\underline{w},\tau,t_1]g(\tau)d\tau dG(\underline{w}), & i=1\end{cases}$$

donde $P_R(\underline{w}, \tau, t_j)$ es la probabilidad complementaria de $P_A[w, \tau, t_j]$ definida anteriormente. Esta fórmula se usa también para la probabilidad de alcanzar el horizonte mediante la colocación de una inspección en t_{N+1}= tiempo horizonte, con $P_R = 1$.

2.4 Modificaciones necesarias en la implementación computacional

Así como se puede observar en las figuras 1 y 2, el mayor cambio entre los casos distribución de nivel de degradación inicial y distribución de tiempos hasta iniciación estriba en que en el primero todos los items empiezan su degradación (observable) en un mismo tiempo y por lo tanto podrán ser rechazados desde la primera oportunidad común de inspección. Por otra parte, en el segundo caso un item podría, con ciertos inputs, empezar a degradar en forma detectable hasta después de la última inspección. El modelo computacional desarrollado para el caso 1 se ve afectado por esas características nuevas del caso 2. En efecto, las curvas de degradación están construidas en base a una matriz que rastrea el crecimiento de los niveles iniciales de degradación de una oportunidad para inspección a otra, hasta mapearse sobre el nivel de falla. Con el propósito de minimizar los cambios necesarios en el algoritmo computacional, se conservó esta matriz de curvas de degradación agregando indicaciones para que los elementos de la matriz que corresponden a puntos (de curvas de degradación) anteriores al nivel de iniciación no se tomen en cuenta en los cálculos o gráficas.

3. EJEMPLOS

En las figuras 3 y 4, representan casos típicos de outputs gráficos para el análisis de confiabilidad con una inspección respectivamente en las oportunidades 9 y 17. La curva

vertical graficada en los tiempos de inspección está relacionada con la probabilidad de detectar un defecto en función de su longitud. Algunos de los datos usados aquí provienen del Departamento de Mecánica de la Northwestern University y describen el crecimiento de una grieta en una pieza metálica cuando ésta última se somete a un estrés cíclico. Sin embargo, la distribución de tiempos de iniciación está modelada arbitrariamente como una normal con percentiles .01 en $\tau = 0$ y .99 en $\tau = 1.2$. Las oportunidades para inspección están igualmente espaciadas y se puede observar cómo una inspección temprana (en la oportunidad 9) traslada la distribución de los tiempos de falla hacia la derecha (figura 3) mientras que una inspección tardía trunca esta distribución a la derecha (figura 4). En este último caso, algunos defectos con menor longitud, pasan la inspección siin ser detectados y alcanzan a fallar como lo indica el repunte en la distribución de tiempos de falla cerca del horizonte. Otros outputs disponibles y no presentes en las gráficos son, por ejemplo, los porcentajes de items rechazados en inspecciones, el porcentaje que llega a horizonte, y percentiles de tiempos de falla.

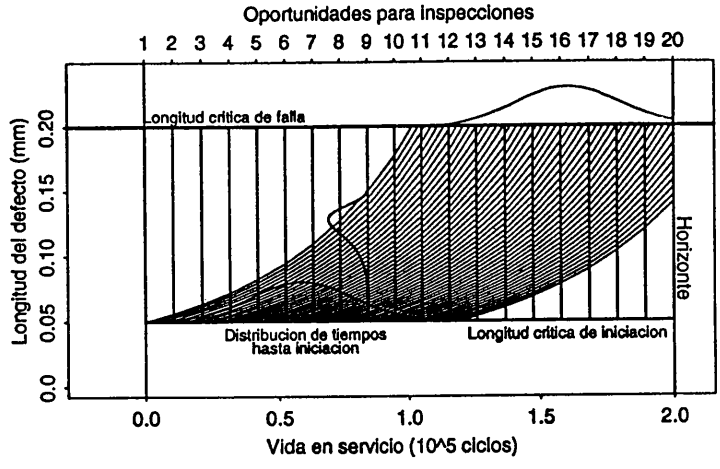


Figura 3. La inspección en 9 rechaza los items que fallarían primero, trasladando la distribución de tiempos de falta hacia la derecha.

4. OBSERVACIONES Y CONCLUSIÓN

El hecho de poder conserva una matriz (modificada) para curvas de degradación en la sección anterior permitió generalizar el programa en vez de escribir uno nuevo específico del caso de degradación a partir de un nivel de iniciación. De esta forma, el usuario dispone de un código computacional que le permite analizar indiferentemente una degradación que se modela a partir de una distribución de niveles iniciales de degradación o de una distribución de tiempos hasta iniciación. Varios aspectos de este problema requieren más investigación, por ejemplo:

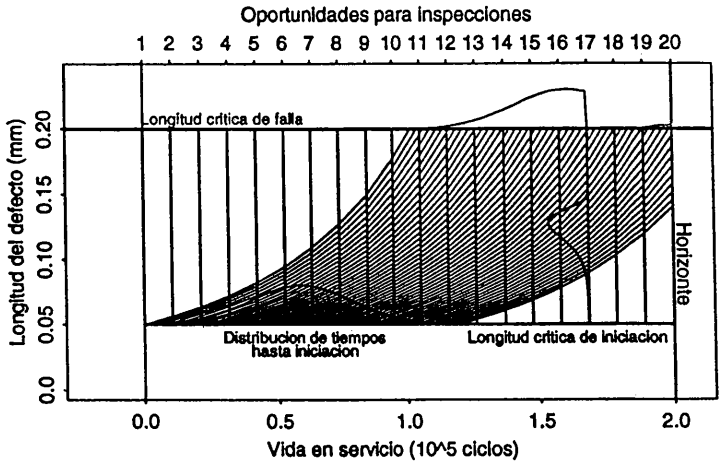


Figura 4. La inspección en 17 rechaza los items a punto de fallar en ese momento truncando la

♦ para cuestiones de modelización, puede resultar de interés la obtención de las distribuciones iniciales (de degradación o de tiempos a iniciación) a partir de la distribución de tiempos de falla, el inverso del problema presentado aquí. Este problema es trivial en el caso de degradación determinística, pero se complica si los parámetros que describen la degradación son aleatorios.

♦ la mayoría de los artículos que tratan de inspecciones en servicio consideran una probabilidad de detección, no una probabilidad de rechazo basada en una medida de nivel de degradación. Si asumimos que se puede medir este nivel, la información obtenida en cuanto al estado de degradación permitirá programar la inspección siguiente en una forma optimizada ajustada al nivel de degradación presente (el desarrollo de las tecnologías de inspecciones no destructivas indican que tal suposición será pronto una realidad).

REFERENCIAS

Jorgenson, D.W. McCall, J.J. and Radner, R. (1967). *Optimal Replacement Policy*, Rand McNally & Company, Chicago.

Barlow, R.E. and Proschan, F. (1975). *Statistical Theory of Reliability and Life Testing Probability Models*, New York: Holt; Rinehart and Winston, Inc.

Anders, G.J. (1990). Probability Concepts in Electric Power Systems, chapter IX, *John Wiley & sons*, New York.

Garrigoux, C.G. and Meeker, W.Q. (1994). *Assessing the effect of In-Service Inspections on the Reliability of Degrading Components*, (Recent Advances in Life Testing and Reliability), N. Balakrishnan, de., CRC Press. (In press).

Becker, R.A., Chambers, J. M., and Wilks, A.R. (1988). *The New S Language*, Wadsworth & Brooks/Cole, Pacific Grove, CA.

Kitagawa, H. and Hisada, T. (1977). *Reliability Analysis of Structures Under Periodic Non-Destructive Inspection*, Pressure Vessel Technology, Pt 1, ASME, New York.

Harris, D.O. and Lim, E.Y. (1983). *Applications of a Probabilistic Fracture Mechanics and Fatigue Methods: Applications for Structural Design and Maintenance*, ASTM STP 798 (J.M. Bloom and J.C. Ekvall, Eds.), American Society for Testing and Materials, *Philadelphia*, pp. 19-41

Yang, J.N. and Chen, S. (1985a). Fatigue Reliability of Structural Components Under Scheduled Inspection and Repair Maintenance, Probabilistic Methods in the Mechanics of Solids and Structures, *Proceedings IUTAM Symposium, Stockolm 1984, Springer*, New York, pp. 559-568.

¿ Qué es el Análisis de Observaciones Repetidas?

LETICIA GRACIA MEDRANO

IIMAS-UNAM, México

1. INTRODUCCIÓN

Las mediciones repetidas son observaciones de una misma característica que se hacen varias veces, lo que las distingue de otras observaciones es que: 1) la misma variable se observa en una unidad más de una vez y que éstas no son independientes como en el análisis de regresión y 2) más de una unidad está involucrada en el estudio así que no forman una simple serie de tiempo.

Las mediciones repetidas se presentan en todos los campos científicos: Agricultura, Demografía, Ingeniería, Medicina, etc. Las observaciones son tomadas bajo diversas condiciones y el objetivo es determinar la influencia de estas condiciones sobre la variable respuesta, la *variación sistemática*. El interés principal está puesto en como la distribución de las respuestas cambia bajo diferentes condiciones. Cualquier parámetro desconocido que especifique la forma de esta distribución sería considerado como un parámetro de ruido.

Los modelos para mediciones repetidas tienen que contemplar principalmente dos aspectos:

1. Los dos tipos de dependencia estocástica que se dan en la misma unidad observacional
   - homogeneidad de las respuestas de una unidad / heterogeneidad a través de las unidades.
   - distancia en tiempo o espacio entre las respuestas de una unidad.
2. El tipo de variable que es medida,
   - datos continuos
   - datos categóricos
   - datos de duración o supervivencia

La unidad de observación se define como el objeto sobre el cuál más de una medición interdependiente se hace, sin importar si ocurren todas las mediciones de manera simultánea o de manera sucesiva.

En el estudio de las mediciones repetidas se consideran tres características en las unidades:

1. Las unidades son muestreadas de manera que las respuestas son independientes entre las unidades.
2. Todas las respuestas de una unidad generalmente estarán más relacionadas entre sí que con las otras respuestas de otras unidades. Parte de esta variabilidad podrá ser explicada por las covariables que diferencían a cada unidad. Y lo que resta de esta variabilidad es considerado una dependencia estocástica.
3. Cuando una espacio contínuo está involucrado, las respuestas más cercanas en una unidad están más correlacionadas.

Para dar una idea de lo importante que es el estudio de la interdependencia estocástica entre las respuestas de una unidad, se da el siguiente ejemplo: Si la variable respuesta es una variable binaria (despierto o dormido). Considerando que un individuo está dormido el 30% del tiempo. Si las respuestas fueran consideradas como independientes, la probabilidad de que un individuo

estuviera dormido en cuatro periodos consecutivos es: 0.3 x 0.3 x 0.3 x 0.3 = 0.0081 Pero si se considera que la probabilidad de estar dormido depende del estado previo y suponiendo una probabilidad de transición es de 0.9 se tiene que la probabilidad de estar dormido es: 0.3 x 0.9 x 0.9 x 0.9 = 0.2187.

2. MODELO DE DISTRIBUCIÓN NORMAL

El primer modelo es para *mediciones cuantitativas* en los reales y que provienen de una distribución normal (o que pueden ser transformadas a ella como el caso de la lognormal). Dado que en las mediciones repetidas existe una dependencia estocástica la distribución normal multivariada, sin especificar a la matriz de covarianzas ha sido utilizada para el análisis de mediciones repetidas. Suponiendo que se tienen el **mismo número** R **de observaciones** para cada unidad, $R(R + 1)/2$ parámetros deben estimarse para la matriz de covarianzas. Si N el número de unidades observadas no es suficientemente grande respecto a R el modelo no podrá ser estimado. Si Y_{NxR} es la matriz de respuestas y M_{NxR} la matriz de medias entonces se tiene que $Y \sim NMV(M, I_N \otimes \Sigma)$ Para el caso en que no se tengan tratamientos o covariables que cambien con la respuesta se tiene un MANOVA común, donde las variables explicativas sólo cambian de un grupo de unidades a otro. Para el caso donde hay diferencias en las unidades el modelo es: $E[Y] = M = XB$ donde X_{NxC} es la matriz de diseño que describe las C condiciones del experimento y B_{CxR} es una matriz de parámetros desconocidos . En este modelo Σ se deja sin especificar, y se tiene un total de $R(R+1)/2 + C$ x R parámetros desconocidos. Un modelo como este no permite relacionar a las respuestas repetidas a través de un modelo de localización.

Pothoff y Roy (1964) generalizan el modelo de MANOVA y lo que proponen es : $E[Y] = XBZ$. Aquí Y y X son las matrices respuesta y de diseño respectivamente y B_{PxC} es una matriz de parámetros y Z_{PxR} es una matriz de covariables que cambian con la respuesta de una unidad . Z generalmente contiene P-1 polinomios ortogonales que describen como cambia la respuesta de una unidad en el tiempo.

Por ejemplo en un experimento de crecimiento animal para estudiar las diferencias en efecto de diferentes tipos de alimento, X describiría las diferencias de tratamiento, Z sería un polinomio en el tiempo y B daría el perfil de crecimiento promedio para cada tipo de alimentación.

De nuevo Σ queda sin especificar, y tiene que estimarse por separado a través de: $\left(\hat{\Sigma}_0 = \left(Y - X\hat{B}Z\right)'\left(Y - X\hat{B}Z\right)/N\right)$.

3. COMPONENTES DE VARIANZA

Hasta ahora se tiene la dependencia estocástica entre las respuestas en una unidad, pero sin tomar en cuenta la distancia entre las observaciones. Una simplificación con respecto a la matriz de covarianzas sin especificar, es modelarla directamente de manera que todas las respuestas de una unidad estén *equiespaciadas*.

Este modelo contempla que la covarianza es constante para todos los pares en una unidad. Todas las respuestas están igualmente relacionadas, pero entre unidades son independientes.

La matriz de covarianza puede escribirse entonces como: $\sum \psi^2 I_R + J_R' \delta J_R$. En esté caso la varianza consta de dos partes: ψ^2 la variación de las respuestas en la misma unidad; más δ una componente adicional de variación a través de las unidades. Esta última también corresponde a la covarianza constante entre las respuestas de una unidad, por lo que δ no requiere ser positiva. Cuando se tiene un diseño razonablemente balanceado, estos parámetros ψ^2 y δ pueden ser estimados através del ANOVA clásico. Utilizando el modelo "completo" para la estimación de δ y ψ y un modelo "simple" para los estimadores de localización. Reduciéndose notablemente la cantidad de parámetros.

4. MODELO DE EFECTOS ALEATORIOS

Otra forma de modelar la estructura de la dependencia estocástica es el modelo de efectos aleatorios, éste enfatiza la variación de las respuestas a través de las unidades en vez de la homogeneidad de éstas en una unidad.

En un ANOVA de dos factores el modelo es: $E[Y_{ik}|\lambda_i] = \mu + \lambda_i + \beta_k, \lambda_i$ describe las diferencias entre unidades y β_k las condiciones de respuesta en cada unidad. Suponiendo que las unidades son muestra aleatoria de una población más grande, λ_i describiría las diferencias promedio entre las unidades, y podría considerarse que tenga una distribución aleatoria, por ejemplo una N(0, δ) independiente de la distribución de $Y_{ik}|\lambda_i$ que es N(0, ψ^2). En forma matricial el modelo queda: $E[Y|\Lambda] = J_N^t BZ + \Lambda J_R$

Donde J_N y J_R son vectores de unos, B_{1xP} es un vector de parámetros, Z tiene la primer columna de unos para la media y el resto son ceros y unos de las condiciones de la unidad. Λ es un vector de N parámetros aleatorios. La esperanza marginal de la respuesta, (luego de integrar) es: $E[Y] = J_N' BZ$ así que Λ es un parámetro que no aparece en el modelo final. Este modelo supone correlación constante entre las respuestas de una unidad. Tiene el mismo modelo de localización y estructura estocástica que el anterior, sólo que aquí δ requiere ser no negativa por ser una varianza; en este sentido es menos general.

5. MODELO DE COEFICIENTES ALEATORIOS

Elston y Grizzle (1962) proponen un modelo en el que las respuestas son de la forma: $E[Y|\Lambda] = XBZ + \Lambda V$ donde $Y|\Lambda \sim NMV(XBZ + \Lambda V, I_N \otimes \Psi)$ y $\Lambda \sim NMV(0, I_N \otimes \Delta)$.

La distribución marginal que se obtiene para Y es: $NMV(XBZ, I_N \otimes (\Psi + V' \Delta V))$ donde Y_{NxR}, X_{NxC}, B_{CxP}, Z_{PxR}, Λ_{NxS}, V_{SxR}, Δ_{SxS}, y Ψ_{RxR}. Siendo N el número de individuos, C las condiciones del diseño a través de las unidades, P las condiciones de diseño dentro de las unidades, R el número de repeticiones en una unidad y S el número coeficientes aleatorios por unidad.

Generalmente las respuestas son consideradas condicionalmente independientes, en ese caso $\Psi = \psi^2 I_R$. La mayoría de las veces Z y V son distintas, V puede contener un subconjunto de renglones de Z ($S \leq P$). La innovación de este modelo consiste en que los parámetros que describen el cambio de las respuestas en una unidad puede tener una distribución aleatoria. Se utiliza más cuando variables explicativas importantes hacen falta. El principal problema de este modelo surge de la pregunta ¿ por qué unos coeficientes del modelo de localización para la media son aleatorios y otros no?

Si V es un vector de unos este modelo y el de efectos aleatorios coinciden, si $\Delta = 0$ y Ψ arbitrario se recupera el modelo de Potthoff y Roy. En este caso la matriz de covarianza es estimada por descomposición de varianza del modelo completo, mientras que los parámetros son estimados del modelo más simple.

6. OTROS MODELOS

Hasta aquí los modelos presentados tratan de modelar el hecho de que las respuestas de una unidad están más cercanas entre si que con las respuestas de otras unidades, pero existen muchos estudios de mediciones repetidas que involucran al tiempo (o espacio) y la evolución de las respuestas es de especial importancia. Si las mediciones repetidas se dan en tiempos discretos y equiespaciados sin valores faltantes y suponiendo estacionariedad se pueden utilizar modelos de correlación serial.

Dado que generalmente se tienen series cortas un modelo AR(1), en muchas ocasiones es suficiente. Se considera que la distribución para las R respuestas de una unidad es $Yi \sim NMV\left(B_i' Z_i, \Sigma\right)$, donde $B_i' Z_i$ es el modelo de localización que depende sólo de las covariables actuales. La matriz de covarianza se escribe como:

$$\sum = \frac{\xi}{\left(1-\rho^2\right)} \begin{pmatrix} 1 & \rho & \cdots & \rho^{R-2} & \rho^{R-1} \\ \rho & 1 & \cdots & \rho^{R-3} & \rho^{R-2} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \rho^{R-1} & \rho^{R-2} & \cdots & \rho & 1 \end{pmatrix}$$

Existen otros modelos para los casos donde las respuestas dependen tanto del estado actual como de los estados previos, tal es el caso *del modelo de dependencia de estado.*

Se han desarrollado modelos que relajan el supuesto de estacionariedad de segundo orden, por ejemplo puede ocurrir que la varianza en la respuesta en una curva de crecimiento aumente junto con la respuesta. Al modelo autoregresivo no estacionario se le conoce como *modelo de antedependencia.* Los modelos pueden irse complicando tratando de conjuntar los modelos que manejan la heterogeneidad entre unidades y los modelos autoregresivos. Cuando se manejan mediciones repetidas que no son equiespaciadas se utilizan los *modelos lineales dinámicos.*

Estos son algunos de la gran variedad de modelos que pueden utilizarse para el análisis de mediciones repetidas cuantitativas.

REFERENCIAS

Elston,R.C. y Grizzle,J.F.(1962) Estimation of time response curves and their confidence bands. *Biometrics* **18**, 148-159.

Lindsey,J.K.(1993) *Models for repeated measurements*, Oxford University Press.

Potthoff, R.F. y Roy,S.N.(1964) A generalized multivariate analysis of variance model useful specially for growth curves problems. *Biometrika* **51**, 313-326.

Análisis Bayesiano Conjugado del Proceso de Galton-Watson

EDUARDO GUTIÉRREZ PEÑA y MANUEL MENDOZA

IIMAS-UNAM, México *ITAM, México*

1. INTRODUCCIÓN

De acuerdo con el modelo de Galton-Watson (Karlin y Taylor 1975) para describir el crecimiento de poblaciones, los individuos se reproducen independientemente con la misma distribución de descendencia $P(X = i) = \pi_i$, $i = 0,...,k$. A partir de Z_0 ancestros originales, la evolución de la población puede describirse a través de $\{Z_n;\ n = 0,1,...\}$ donde Z_n es el número de individuos en la generación n, o con más detalle, por medio de $\{Z_{nj};\ j = 0,...k;\ n = 0,1,...\}$ con Z_{nj} el número de individuos en la generación n que dieron origen a j descendientes. La distribución básica es discreta con vector de probabilidades $\boldsymbol{\pi} = (\pi_0,...,\pi_k)'$. Este modelo ha sido considerado por diversos autores (Harris 1963, Jagers 1975, Karlin y Taylor 1975, entre otros). En particular, $\{Z_n;\ n = 0,1,...\}$ es un proceso de Markov con probabilidades de transición $P(Z_n = i \,|Z_{n-1} = j, \boldsymbol{\pi}) = p(i,j|\boldsymbol{\pi})$ donde $p(i,j|\boldsymbol{\pi})$ es la convolución de orden j de la distribución $\boldsymbol{\pi}$. Si $Z_n = 0$ para alguna n, la población se extingue ya que $Z_r = 0$ para $r \geq n$ y una característica de interés es la probabilidad de extinción $q = P(Z_n = 0,\ \text{alguna}\ n = 1,2,...|\boldsymbol{\pi})$. En particular, interesa el caso $q = 1$, (extinción segura). Se sabe (Jagers 1975, pág. 22) que q es el punto fijo más pequeño de la función generadora de probabilidades de la distribución $\boldsymbol{\pi}$. Aún cuando $\boldsymbol{\pi}$ sea conocido, q no puede determinarse analíticamente y es vital la caracterización alternativa en términos de la media de reproducción (Harris 1963). Sea $m = m(\boldsymbol{\pi}) = \sum_{i=0}^{k} i\pi_i$, la media de reproducción del proceso. Si $\pi_1 \neq 1$, entonces $m > 1$ si y sólo si $q < 1$. La media m es más fácil de calcular que la probabilidad q y da lugar a una clasificación del proceso. Este es subcrítico, crítico o supercrítico si ocurren respectivamente las condiciones $m < 1$, $m = 1$ o bien, $m > 1$. El caso $\pi_1 = 1$ se puede ignorar en tanto que el proceso no involucra incertidumbre. Así, en términos de extinción, interesa m. El proceso se describe completamente sólo a través de $\boldsymbol{\pi}$, pero pueden analizarse m y algunos otros parámetros como, por ejemplo, la varianza $\sigma^2 = \sigma^2(\boldsymbol{\pi}) = \sum_{i=0}^{k} (i - m(\boldsymbol{\pi}))^2$

2. INFERENCIA FRECUENTISTA

En un proceso de Galton-Watson, a partir de n generaciones es posible producir inferencias sobre la población completa. El tema es reciente pero existen diversas contribuciones (Dion 1974, Dion y Keiding 1977, Basawa y Prakasa-Rao 1980, Pérez-Abreu 1987, Maki y McDunnough 1989, González y Pérez-Abreu 1991, Prakasa-Rao 1992, entre otros). El problema de ha sido abordado desde una perspectiva frecuentista y se han producido estimadores para $\boldsymbol{\pi}$, m y q vía máxima verosimilitud. Si se cuenta con $\mathbf{Z}_n = \{Z_{ij};\ j = 0,...k;\ i = 0,...,n\}$, se obtienen las estimaciones

$\hat{\pi}_j = \frac{Y_{nj}}{Y_n}$, $j = 0,\ldots,k$, y $\hat{m} = \frac{(Y_{n+1} - Z_0)}{Y_n}$ donde $Y_n = \sum_{j=0}^{k} Y_{nj}$ y $Y_{nj} = \sum_{i=0}^{n} Z_{ij}$

representan respectivamente, el número total de individuos en la población hasta la generación n, y el número de éstos individuos con exactamente j descendientes. Es interesante comprobar que

$$Z_i = \sum_{j=0}^{k} Z_{ij} \qquad y \qquad Z_{i+1} = \sum_{j=0}^{k} jZ_{ij}$$

De esta manera, la información $\{\boldsymbol{Z}_i \; ; \; i = 1,2,\ldots n\}$, determina el valor de Z_{n+1}, el tamaño de la generación n+1 y, en consecuencia, el valor de Y_{n+1}. En cualquier caso, basta la pareja $(Y_n, \; Y_{n+1})$ para estimar m y se puede proceder a la clasificación del proceso sustituyendo $\hat{m}$ por m en el criterio descrito en la sección 1. Para medir la fiabilidad de este criterio es necesario, sin embargo, conocer la distribución de m. Al respecto sólo se cuenta con resultados que establecen, en el caso supercrítico, la normalidad asintótica de m. No existen caracterizaciones para tamaños de muestra finitos y aun para calcular cualquier probabilidad con el modelo *normal* asintótico es necesario el valor desconocido de σ^2 (aunque es posible estimarlo a partir de $\mathbf{Z}_n$). Desde esta perspectiva, no es posible probar la hipótesis de extinción segura $(q = 1)$ del proceso. En la siguiente sección se considera, como alternativa, un análisis Bayesiano de m que elimina estos problemas.

3. ANÁLISIS BAYESIANO

Si hasta la generación n del proceso se cuenta con la información $\mathbf{Z}_n$, entonces verosimilitud está dada por

$$L(\boldsymbol{\pi}|\mathbf{Z}_n) \; \propto \; \prod_{j=0}^{k} \pi_j^{Y_{nj}} \tag{1}$$

para todo vector $\boldsymbol{\pi}$ cuyas entradas sean positivas y sumen a 1. Desde una perspectiva Bayesiana, es necesario asignar una distribución inicial para el vector $\boldsymbol{\pi}$. Por facilidad, se puede recurrir a la familia conjugada correspondiente (De Groot 1970, Cap. 9)

$$p(\boldsymbol{\pi}) \propto \prod_{j=0}^{k} \pi_j^{\alpha_j - 1} \tag{2}$$

con $\alpha_j > 0$, $j = 0,\ldots,k$. Esto significa asignar una distribución inicial $Dirichlet(\boldsymbol{\pi}|\boldsymbol{\alpha})$ para el vector $\boldsymbol{\pi}$. Es fácil comprobar que

$$E(\pi_j|\mathbf{Z}_n) = \frac{\beta_j}{\beta}, \quad j = 0,\ldots,k, \qquad Var(\pi_j|\mathbf{Z}_n) = \frac{\beta_j(\beta - \beta_j)}{\beta^2(\beta+1)}, \quad j = 0,\ldots,k$$

$$\text{y} \qquad Cov(\pi_r, \pi_s|\mathbf{Z}_n) = -\frac{\beta_r \beta_s}{\beta^2(\beta+1)}, \quad r \neq s, \qquad \text{donde} \quad \beta = \sum_{j=0}^{k} \beta_j .$$

Además, la moda está dada por el vector $\boldsymbol{\pi}^M = (\pi_1^M,\ldots,\pi_k^M)^t$, con $\pi_j^M = (\beta_j - 1)/\{\beta - (k+1)\}$, $j = 1,\ldots,k$. Vale la pena destacar que esta distribución final tiene todas estas propiedades para cualquier tamaño de muestra. Si además, el proceso es supercrítico, la distribución final del vector $\boldsymbol{\pi}$ converge a una *normal* si n tiende a infinito.

Para este propósito conviene observar que la verosimilitud (1), es la misma que se obtendría a partir de una muestra aleatoria *multinomial* de tamaño Y_n. Por lo tanto, si Y_n tiende a infinito, la normalidad asintótica para la distribución final de π queda establecida (Bernardo y Smith 1994, Sec. 5.3). Es importante insistir en que se requiere que Y_n tienda a infinito, no sólo que lo haga n. Naturalmente, esto sólo puede ocurrir en el caso de un proceso supercrítico. Ahora bien, para clasificar el proceso basta con la distribución de m, que está bien definida para cualquier caso. Los dos primeros momentos de m se pueden calcular en forma exacta recurriendo a su estructura lineal. De hecho, si se definen $\boldsymbol{\mu} = \boldsymbol{\mu}(\boldsymbol{\beta})$ y $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}(\boldsymbol{\beta})$ como el vector de medias y la matriz de varianzas y covarianzas de la distribución final de $\boldsymbol{p}$, se tiene que

$$E(m|\mathbf{Z}_n) = \boldsymbol{\lambda}^t\boldsymbol{\mu} = m(\boldsymbol{\mu}) \quad \text{y}$$
$$Var(m|\mathbf{Z}_n) = \boldsymbol{\lambda}^t\boldsymbol{\Sigma}\boldsymbol{\lambda} = \sigma^2(\boldsymbol{\mu})/(\beta+1) \tag{3}$$

donde $\boldsymbol{\lambda} = (1,...,k)^t$. Para obtener la distribución completa de m, se puede considerar una transformación $\boldsymbol{\varphi} = \boldsymbol{\varphi}(\boldsymbol{\pi})$ tal que $\boldsymbol{\varphi} = (m, \boldsymbol{\tau}^t)^t$ con τ un vector de dimensión k-1. A partir de la final de $\boldsymbol{p}$ puede entonces obtenerse la final de $\boldsymbol{\varphi}$ y marginalizando, la final de m. En otras palabras, si se obtiene $p(\boldsymbol{\varphi}|\mathbf{Z}_n)$, la distribución de m puede obtenerse como

$$p(m|\mathbf{Z}_n) = \int p(\boldsymbol{\varphi}|\mathbf{Z}_n)\, d\boldsymbol{\tau}. \tag{4}$$

No es fácil, sin embargo, encontrar una transformación que permita este cálculo analítico de $p(m|\mathbf{Z}_n)$. Una alternativa es obtener una aproximación vía simulación. Reescribiendo (4) como

$$p(m|\mathbf{Z}_n) = \int p(m|\boldsymbol{\tau}, \mathbf{Z}_n)\, p(\boldsymbol{\tau}|\mathbf{Z}_n)\, d\boldsymbol{\tau}$$

y si se cuenta con una muestra $\{\boldsymbol{\tau}_1,....,\boldsymbol{\tau}_N\}$ del modelo $p(\boldsymbol{\tau}|\mathbf{Z}_n)$, (4) se puede aproximar (*e.g.* Bernardo y Smith 1994, Sec. 5.5.5) por

$$p(m|\mathbf{Z}_n) \approx \frac{1}{N}\sum_{i=1}^{N} p(m|\boldsymbol{\tau}_i, \mathbf{Z}_n). \tag{5}$$

Esta aproximación es precisa si se elige un valor de N suficiente grande y la elección de τ debe facilitar la simulación de muestras de $p(\boldsymbol{\tau}|\mathbf{Z}_n)$. Este es el caso si $\boldsymbol{\tau} = (\pi_1,...,\pi_{k-1})^t$ puesto que $p(\boldsymbol{\tau}|\mathbf{Z}_n)$ es una distribución *Dirichlet* con parámetro $(\beta_1,...,\beta_{k-1},\beta_0+\beta_k)$. De hecho, si se simula una muestra $\{\boldsymbol{\pi}_1,....,\boldsymbol{\pi}_N\}$ de $p(\boldsymbol{\pi}|\mathbf{Z}_n)$, se obtiene la información necesaria para (5) así como muestras de cualquier otra característica de interés del proceso. Este procedimiento se ilustra en el ejemplo de la sección 4. Desde otro punto de vista, es posible utilizar los dos primeros momentos y una forma paramétrica simple para aproximar $p(m|\mathbf{Z}_n)$. Una alternativa es el modelo *normal*. Recurriendo al criterio de mínima divergencia logarítmica, (Bernardo y Smith 1994, Sec. 3.4.3) la mejor aproximación *normal* es la que tiene las mismas media y varianza. Otra posibilidad, sugerida por algunos ejemplos, es un modelo *gama* con los mismos dos primeros momentos. Finalmente, la normalidad asintótica de π, cuando ocurre, induce otra aproximación *normal* (asintótica) para m con

media $m(\pi^M)$ y varianza $\sigma^2(\pi^M)/[\beta-(k+1)]$. Por lo que respecta al problema de predicción, es decir a las inferencias sobre Z_r, para $r > n+1$, se conocen los dos primeros momentos de la distribución condicional de Z_r dado $\boldsymbol{Z}_n$ y el vector π (Jagers 1975, Sec. 2.2). Así, $E(Z_r|\boldsymbol{Z}_n,\pi) = m^{r-(n+1)} Z_{n+1}$ $\forall m$ y

$$Var(Z_r|\boldsymbol{Z}_n,\pi) = Z_{n+1}\,\sigma^2\,\{m^{r-(n+2)}(m^{r-(n+1)}-1)\}/(m-1)$$

cuando $m \neq 1$, mientras que $Var(Z_r|\boldsymbol{Z}_n,\pi) = Z_{n+1}\,(r-(n+1))\sigma^2$ cuando $m = 1$. Desafortunadamente, no existe una expresión completa para $p(Z_r|\boldsymbol{Z}_n,\pi)$. En todo caso, para la predictiva $p(Z_r|\boldsymbol{Z}_n)$ se pueden aproximar su media y varianza, a partir de $\{\pi_1,.....,\pi_N\}$, como $E(Z_r|\boldsymbol{Z}_n) \approx N^{-1}\sum_{i=1}^{N} E(Z_r|\boldsymbol{Z}_n,m_i)$ para la media y

$$Var(Z_r|Z_n) \approx N^{-1}\sum_{i=1}^{N} Var(Z_r|Z_n,m_i,\sigma_i^2) + N^{-1}\sum_{i=1}^{N} E(Z_r|Z_n,m_i)^2 - \left[N^{-1}\sum_{i=1}^{N} E(Z_r|Z_n,m_i)\right]^2$$

para la varianza en donde $\{m_1,.....,m_N\}$ y $\{\sigma_1^2,.....,\sigma_N^2\}$ son las muestras de m y σ^2 inducidas por $\{\pi_1,.....,\pi_N\}$. Finalmente, $p(Z_r|\boldsymbol{Z}_n)$ puede aproximarse a través de una distribución con media $E(Z_r|\boldsymbol{Z}_n)$ y varianza $Var(Z_r|\boldsymbol{Z}_n)$.

4. EJEMPLO

Se simularon 10 generaciones con $k = 5$ y $\pi = (0.50,0.05,0.05,0.05,0.05)'$. En consecuencia, $m = 1.2$ y $\sigma^2 = 1.76$ (el proceso es supercrítico). Los datos son $(Y_{10,0}, Y_{10,1}, Y_{10,2}, Y_{10,3}, Y_{10,4}, Y_{10,5}) = (137, 238, 29, 29, 24, 24)$. Hasta la generación 10, se tiene $Y_{10} = 481$ mientras que $Y_{11} = 600$. Por máxima verosimilitud se tiene $\hat{m} = 1.245$. Por otra parte, la media y varianza de la final de m son 1.253 y 0.00369 y para la aproximación asintótica, los valores son 1.237 y 0.00359.

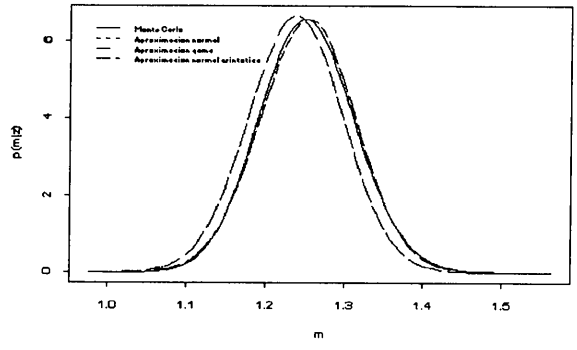


Fig. 1. *Densidad posterior de m*

De la distribución final de m se simularon 20,000 muestras. En la Figura 1 se presentan esta distribución y las aproximaciones normal (N), gama (G) y normal asintótica (A). Las aproximaciones N y G son más precisas que la aproximación A. La probabilidad del evento

$m>1$ es cercana a 1, de forma que el proceso se puede clasificar sin problemas como supercrítico.

5. CONCLUSIONES

El problema de producir inferencias sobre la media de reproducción m de un proceso de Galton-Watson, está directamente relacionado con la clasificación del proceso. Desde una perspectiva frecuentista, las inferencias sobre m, más allá de su estimación puntual, enfrentan dificultades y aún los resultados asintóticos son de cuestionable utilidad. Con un enfoque Bayesiano, no es fácil encontrar una expresión analítica para la final de m aún en el caso de una inicial conjugada. Sin embargo, es posible obtener una aproximación de esa distribución final, vía simulación, arbitrariamente precisa. Se sugiere una aproximación normal pero, cualquier caso, se requiere un estudio más profundo en esa dirección. En relación al problema de predicción, se propone una aproximación, de nuevo vía simulación, para los dos primeros momentos de la distribución predictiva relevante. Los resultados se pueden generalizar, en su mayor parte, al caso en que la distribución inicial pertenece a una familia de mezclas de conjugadas.

REFERENCIAS

Basawa, I.V. y Prakasa-Rao, B.L.S. (1980). *Statistical Inference for Stochastic Processes.* Academic Press : New York.

Bernardo, J.M. y Smith, A.F.M. (1994). *Bayesian Theory.* Wiley : Chichester.

DeGroot, M.H. (1970). *Optimal Statistical Decisions.* McGraw-Hill : New York.

Dion, J.P. (1974). Estimation of the Mean and the Initial Probabilities of the Branching Processes. *J. Appl.Prob.* **11**, 687-694.

Dion, J.P. y Keiding, N. (1978). Statistical Inference in Branching Processes. In : *Branching Processes* (Joffe, A. y Ney, P. eds.) 105-140. Marcel Dekker : New York.

González, F.A. y Pérez-Abreu, V. (1991). Propiedades Asintóticas de los Estimadores de Máxima Verosimilitud en los Procesos Ramificados Bisexuales. *Agrociencia. Mats. Apl. Est. y Comp.* **2**, 115-126.

Harris, T.E. (1963). *The Theory of Branching Processes.* Springer Verlag : Berlín

Jaggers, P. (1975). *Branching Processes with Biological Applications.* Wiley : New York.

Karlin, S. y Taylor, J.M. (1975). *A First Course in Stochastic Processes.* Academic Press : New York.

Maki, E. y McDunnough, P. (1989). What Can or Can't Be Estimated in Branching and Related Processes?. *Stoch. Process. Appl.* **31**, 307-314.

Pérez-Abreu, V. (1987). Los Procesos Ramificados como Modelo para Detectar Brotes de Epidemias de una Enfermedad Contagiosa : Aspectos Estadísticos. *Memorias del Segundo Foro de Estadística Aplicada.* UNAM : México.

Prakasa-Rao, B.L.S. (1992). Nonparametric Estimation for Galton-Watson type Processes. *Stat. Prob. Letters* **13**, 289-293.

Ineficiencia de la Carta p para Tamaños de Subgrupo Grande: Diagnóstico y Alternativas

HUMBERTO GUTIÉRREZ PULIDO y OSVALDO CAMACHO CASTILLO

Univ. de Guadalajara, México

1. INTRODUCCIÓN

La carta p muestra las variaciones en la proporción de artículos defectuosos de un proceso y es bastante utilizada para tener un control de tal proporción. Los límites de la carta se calculan bajo el supuesto de distribución binomial, por lo que:

$$Limites\ de\ control = \bar{p} \pm 3\sqrt{\frac{\bar{p}(1-\bar{p})}{n}}$$

donde $\bar{p}$ es la estimación de la proporción p, de artículos defectuosos en el proceso, que junto con n (el tamaño de subgrupo) son parámetros.

La carta de individuales es un diagrama para variables de tipo contínuo en donde se grafican directamente las mediciones individuales. Sus límites de control están dados por :

$$\bar{X} \pm 3(\bar{R} / d_2),$$

donde la constante d_2=1.128, y $\bar{R}$ es el promedio de los rangos móviles de orden 2.

La amplitud de los límites de control de una carta p es inversamente proporcional a la raíz cuadrada del tamaño de muestra o subgrupo n, por lo que cuando n es muy grande los límites se estrechan. Así que proporciones con pequeñas desviaciones con respecto al promedio de artículos defectuosos caen fuera de los límites. Ésta puede llegar al extremo de que ningún punto caiga dentro de los límites; evidentemente en estos casos, la carta p resulta de nula utilidad práctica. Veamos dos casos reales.

Ejemplo 1. En una fábrica de artículos de plástico inyectado se tiene el problema de la "rebaba" en las piezas. Con el propósito de detectar causas especiales de variación se implementa una carta p para la variación en la proporción de piezas con rebaba. Se toman datos de los lotes (n=5000) del producto principal; la proporción de artículos defectuosos en 24 lotes se muestra a continuación:

0.172	0.190	0.226	0.186	0.176	0.202	0.180	0.170
0.222	0.160	0.192	0.178	0.196	0.252	0.192	0.248
0.258	0.230	0.190	0.156	0.194	0.220	0.216	0.236

La carta p para estos datos, tomando en cuenta el tamaño de lote de 5000, aparece en la figura 1. En la que 15 de 24 puntos están fuera de los límites de control. Por lo que esta carta no permite diferenciar los cambios importantes de los que no lo son. Al ser el subgrupo muy grande (5000), los límites de control en la carta p resultan sumamente estrechos (0.185,0.219), de tal forma que una variación de menos de 2 % del promedio (0.202), que es una variación usual en los procesos fabriles, aparecerá como una causa especial, sin embargo en la práctica tal variación difícilmente será importante. Por lo que en estos casos la carta p es de poca utilidad.

TABLA 1
Ejemplo 2

Pedido	Tamaño	Proporción
1	78297	0.14240
2	77112	0.00000
3	746460	0.00705
4	51600	0.19767
5	96600	0.04762
6	204008	0.00004
7	629475	0.01918
8	670056	0.03251
9	93440	0.00000
10	6217	0.13141
11	48000	0.28125
12	35805	0.10627
13	112088	0.02602
14	112088	0.00986
15	1760540	0.01794
16	1650510	0.06674
17	227616	0.05885
18	110088	0.01374
19	1007380	0.03262
20	113220	0.05772
21	548960	0.06304
22	89472	0.08413
23	132300	0.01486
24	132300	0.01109
25	457380	0.09103
26	323100	0.00337
27	828000	0.12802
28	58416	0.00731
29	58416	0.11002
30	6070	0.05189
31	188384	0.00890

Ejemplo 2. En una empresa dedicada a la producción de etiquetas. Se reciben pedidos que desde algunos miles hasta millones de etiquetas. Se planteó la necesidad de evaluar la variación de la proporción de etiquetas defectuosas en la revisión final, por lo que se implementó una carta p con datos de 31 lotes (tabla 1).

La carta p con límites variables para estos datos. Se aprecia que todos los puntos (31) están fuera de los límites de control (figura 3). Si se usan límites promedios, éstos irían de 0.042 a 0.044, por lo que una variación respecto al promedio (0.043) de más de una milésima en el valor de p, se registraría como una causa especial.

Aun para los valores más usuales de (n,p) la aplicación de las pruebas estándar (Gutiérrez, 1992) a la carta p, genera una mayor cantidad de falsas alarmas para ciertos valores de (n,p) que las se dan bajo normalidad (Camacho y Gutiérrez, 1995).

2. DIAGNÓSTICO

Por medio de un estudio de la potencia de la prueba "un punto fuera de los límites de control" para diferentes valores de (n,p), determinaremos los valores de n que hacen que la carta p sea de poca utilidad práctica, debido a la estrechez de sus límites. Ya que otra forma de ver el problema es que la potencia de la prueba aumenta tanto que detecta cambios muy pequeños. En la figura 5 se representan gráficamente algunos de estos resultados donde se puede apreciar la manera en que crece la potencia de la prueba 1 aplicada del lado superior, conforme n aumenta. Se ve que con tamaños de n grande aun cambios pequeños en la proporción de defectuosos tienen una alta probabilidad de ser detectados por la carta. Además la potencia es mayor cuando p es pequeña.

Con base en la figura 5, y considerando que en un proceso con una proporción pequeña de defectuosos es de interés detectar cambios pequeños en tal proporción, esto se puede lograr con valores de n también pequeños, por lo que el problema de la estrechez de los límites se agudizará conforme p sea más pequeña. Si por el contrario se tiene un proceso con una proporción grande de artículos defectuosos, comunmente sólo es de interés detectar

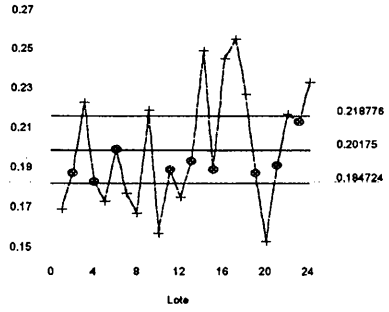


Fig. 1. Carta p

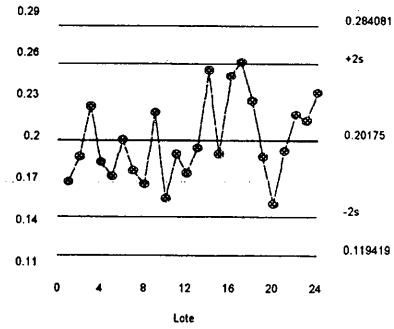


Fig. 2. Carta de individuales

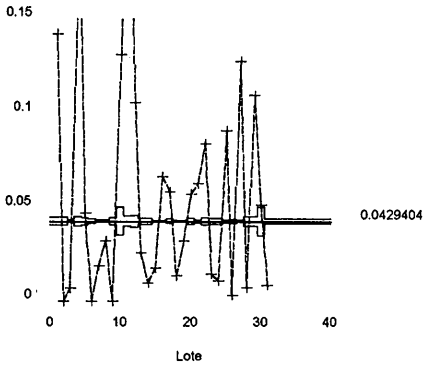


Fig. 3. Carta p

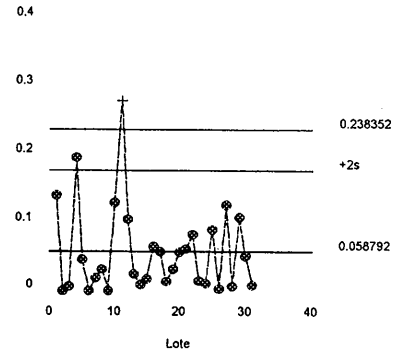
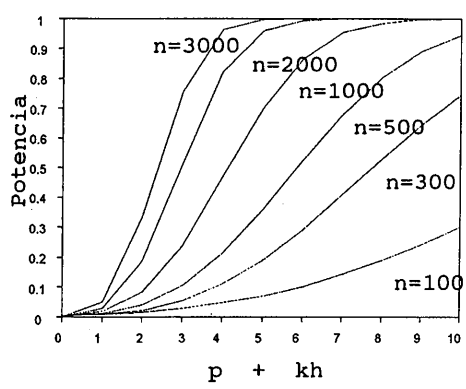
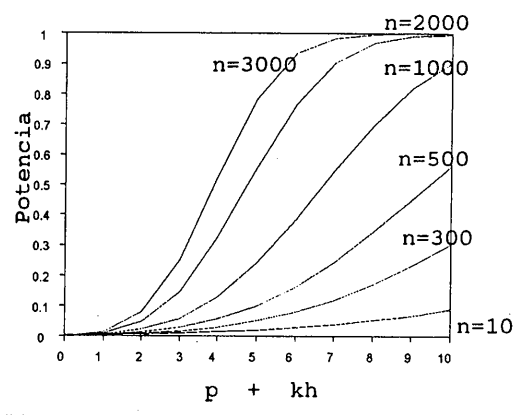


Fig. 4. Carta de individuales



(a) Potencia con p=0.05



(b) Potencia con p=0.15

Figura 5. Potencia para detectar un cambio de p + kh; con h=0.005, y k=0,1,2,...,10.

cambios moderados y grandes, pero el problema es que con tamaños de n del orden de miles no sólo ese tipo de cambios van a ser detectados, sino también los pequeños; por lo que el problema de n grande también estará presente en este caso (p grande).

3. ALTERNATIVAS

Una posibilidad para analizar la proporción de artículos defectuosos con subgrupo muy grande, es usar una carta de individuales para variable aleatoria $X_i=p_i$, que sin considerar el valor de n con que se obtuvo p_i, tendría una distribución simétrica y unimodal aproximada por la distribución normal. Con lo anterior sólo se detectan variaciones grandes en la proporción de artículos defectuosos, lo que se corrige aplicando las ocho pruebas para causas especiales aumentando la potencia (ver Gutiérrez, 1992).

La carta de individuales para las proporciones del ejemplo 1 (figura 2) se aprecia que es un modelo más útil que el de la carta p. En el ejemplo 2 se tiene una situación similar (figura 4). En este caso se elimina el punto fuera del límite para establecer los límites de control definitivos.

Bajo la suposición de que aún en un muestreo con n grande, el modelo binomial describe la realidad de tal proceso, y de que la carta p no funciona debido al incremento de la potencia, se hizo un estudio Monte Carlo en el que para diferentes combinaciones de (n,p) se simularon grupos de proporciones (20), en los que se estimó la desviación estándar por medio del rango móvil entre las sucesivas proporciones del grupo. Con los resultados se comparó esta estimación con el valor de la desviación estándar de las proporciones de un modelo binomial (tabla 2).

TABLA 2

Desviación estándar del modelo binomial σ_p *vs desviación estándar estimada de rangos móviles* $\hat{\sigma}_{p_i}$.

n	p=0.005		p=0.05		p=0.20	
	σ_p	$\hat{\sigma}_{p_i}$	σ_p	$\hat{\sigma}_{p_i}$	σ_p	$\hat{\sigma}_{p_i}$
500	0.00315	0.00015	0.00975	0.00044	0.01789	0.00081
1000	0.00223	0.00015	0.00689	0.00066	0.01265	0.00101
4000	0.00110	0.00037	0.00345	0.00112	0.00632	0.00203
10000	0.00071	0.00055	0.00218	0.00181	0.00400	0.00321
25000	0.00045	0.00045	0.00138	0.00140	0.00253	0.00255

Con valores pequeños de n la desviación estándar del modelo binomial, σ_p, es mayor que la estimada por simulación con rangos móviles, $\hat{\sigma}_{p_i}$. Conforme n crece la segunda estimación se acerca a la del modelo binomial; por ejemplo con n=25000, los valores de ambas desviaciones son muy similares. Por lo que si un proceso funciona bajo el modelo binomial, entonces la carta de individuales tendrá que ser más potente que la carta p para detectar cambios.

De esta manera cuando se tiene n grande y la carta de individuales es un buen modelo y no así la carta p (como en los dos ejemplos); es debido a que el proceso no funciona bajo un modelo binomial; si fuera el caso, la carta p tendría mejor desempeño. Por lo que el

problema más que de potencia, es de que el modelo binomial con n grande supone menor variabilidad que la que ocurre en la realidad y requiriéndose entonces otro modelo que pueda predecir la variaciones y el desempeño de la proporción de defectuosos.

Con base en lo anterior, y de acuerdo con el razonamiento expuesto al inicio de la presente sección, cuando la carta p no sea de utilidad debido a que las variación de las proporciones de defectuosos de un proceso son mayores que la pronosticada por el modelo binomial (como probablemente ocurrirá cuando n sea muy grande), una alternativa útil es llevar una carta de control de individuales para las proporciones, que será un instrumento para detectar cambios importantes y tendencias en el proceso.

Aun cuando el proceso tenga una variación más o menos próxima al modelo binomial, la carta p no resultará útil por lo alto de la potencia de la prueba, lo que hace que aun pequeñas variaciones caigan fuera de los límites. En estos casos es mejor modelar la variación de las proporciones con la alternativa propuesta.

4. CONCLUSIÓN

En los procesos en los que los límites de la carta p son significativamente más estrechos que la variación ordinaria de las proporciones de artículos defectuosos (lo que se espera que ocurra con tamaños de subgrupo grande); una alternativa útil es analizar las proporciones mediante una carta de individuales, estimando la desviación estándar en el estudio inicial con los rangos entre las sucesivas proporciones.

REFERENCIAS

Camacho Castillo, O. y H. Gutiérrez Pulido (1995). Estudio de la significancia de las pruebas para detectar causas especiales de variación en las cartas p y np. *Revista de Estadística*, vol. VII, **número 9**, pag. 84-94.

Gutiérrez Pulido, H. (1992). *Control Total de Calidad*. Ed. Edug, Guadalajara.

Lucas, J.M. and M.S. Saccucci (1990). Exponentially weighted moving average control schemes: properties and enhancements. *Technometrics*, vol. 32, **No. 1**.

Montgomery, D.C. (1991). *Introduction to Statistical Quality Control*, second edition. Wiley, Singapore.

Roes, K.C.B.; R.J.M.M. Does and Y. Schurink (1993). Shewhart-type control charts for individual observations. *Journal of Quality Technology*, vol. 25, **No. 3**.

Análisis de Medidas Repetidas para Datos Categóricos

CLAUDIA LARA PEREZ SOTO y MA. DEL REFUGIO RIVERA RENDON

Bayer de Mexico S. A. de C.V., México

1. RESUMEN

En los ensayos clínicos se presentan situaciones donde se realizan mediciones a través del tiempo para cada paciente, en donde la variable de respuesta está medida en escala nominal u ordinal.

Una forma adecuada de analizar este tipo de datos es por medio del análisis de medidas repetidas para datos categóricos.

En este trabajo se presenta una aplicación de esta metodología en la industria farmacéutica en un estudio abierto, prospectivo y longitudinal que mide la eficacia de un fármaco determinado por medio de pruebas psicológicas.

El orden en el que se presentan las ideas es el siguiente: descripción del ensayo clínico, planteamiento de los modelos, ejemplo para uno de los reactivos de una de las pruebas evaluadas, generalización y conclusiones.

2. DESCRIPCIÓN DEL ENSAYO CLÍNICO

La nimodipina es un compuesto calcioantagonista del grupo de las dihidropiridinas que posee acción vasodilatadora a nivel cerebral. Existen estudios que documentan que la nimodipina es efectiva en la supervivencia de células isquémicas cerebrales y puede tener efectos favorables en el tratamiento de la Enfermedad de Alzheimer (EA) y en la Demencia Multiinfarto (DM).

El objetivo primario es evaluar la eficacia y tolerabilidad de la nimodipina en el tratamiento de pacientes afectados de EA y DM.

Serán incluidos pacientes de ambos sexos mayores de 45 años afectados de EA y/o con DM en condiciones de ver y oír.

Serán excluidos pacientes con infarto miocardico reciente (4 meses), insuficiencia cardiaca descompensada, historia de epilepsis, insuficiencia hepática grave, insuficiencia renal, psicosis esquizofrénicas y afectivas, frecuente ingesta de bebidas alcohólicas.

El tratamiento consiste en la ingesta de nimodipina 30 mg. 3 veces al día durante 12 semanas.

La eficacia se evaluará por medio de una serie de pruebas psicológicas tendientes a medir ciertos aspectos relacionados con la conducta de los pacientes, manifestaciones cognotivas y la capacidad de llevar a cabo las actividades de su vida cotidiana.

Las mediciones se realizaron al inicio del tratamiento y a las semanas 4, 8 y 12.

Nos centraremos en el RAGS-E (Relative Assessment Geriatric Scale) que es una escala en la cual hay observaciones de los parientes sobre la sintomatología global del paciente. Este formato consta de 21 preguntas y es llenado por el médico con el familiar.

Las categorías para cada reactivo son: Para Nada, Algo, Moderadamente y Bastante.

3. MODELOS TEÓRICOS

El objetivo principal es hacer énfasis en el uso de los modelos Log-Lineales como una alternativa para estudiar los procesos de cambio.

Se considera que se tiene la variable categórica Y a dos tiempos; Y_1 y Y_2 con l categorías cada una, generando una tabla de contingencia de $l \times l$ ($l \geq 2$) para individuos (ver Tabla 1).

TABLA 1

Y_1 / Y_2	a(1)	b (2)	...	i (l)	Total Renglón
a (1)	f_{11}	f_{12}	...	f_{1l}	$F_{1\cdot}$
b (2)	f_{21}	f_{22}	...	f_{2l}	$F_{2\cdot}$
i (l)	f_{l1}	f_{l2}		f_{ll}	$F_{l\cdot}$
Total Columna	$F_{\cdot 1}$	$F_{\cdot 2}$	...	$F_{\cdot l}$	$F_{\cdot\cdot}$

Sea f_{ij} la frecuencia observada y F_{ij} la esperada para algún modelo, ambas para la celda (i,j). Sea $F_{i\bullet}$ y $F_{\bullet j}$ las distribuciones marginales para los tiempos 1 y 2 respectivamente.

El modelo log-lineal es el siguiente:

$$\log(F_{ij}) = \lambda + \lambda_{1(i)} + \lambda_{2(j)} + \lambda_{12(ij)}$$

El modelo de independencia en el análisis de medidas repetidas es difícil que se cumpla pues se espera que existan cambios en el tiempo.

Las hipótesis a probar son las siguientes:

i) *Homogeneidad Marginal* (HM)

$$H_0: \quad F_{i\cdot} = F_{\bullet j}$$

en donde el modelo log-lineal sería

$$\log(F_{ij}) = \lambda + \lambda_{1(i)} + \lambda_{1(j)} + \lambda_{12(ij)}$$

con la restricción de $\lambda_{i\cdot} = \lambda_{\cdot i}$.

ii) *Cuasi-Simetría* (CS)

$$H_0: \quad \lambda_{12(ij)} = \lambda_{12(ji)}$$

el modelo log-lineal para esta prueba es

$$\log(F_{ij}) = \lambda + \lambda_{1(i)} + \lambda_{2(j)} + \lambda_{12(ij)}$$

donde se considera que

$$\lambda_{12(ij)} = \lambda_{12(ji)} \quad \text{y} \quad \sum_i \lambda_{1(i)} = \sum_j \lambda_{2(j)} = \sum_i \lambda_{12(ij)} = 0.$$

iii) *Simetría* (S)

$$H_0: \lambda_{1(i)} = \lambda_{2(i)} \quad y \quad \lambda_{12(ij)} = \lambda_{12(ji)} \quad \forall \quad i \neq j$$

y su modelo log-lineal

$$\log(F_{ij}) = \lambda + \lambda_{1(i)} + \lambda_{1(i)} + \lambda_{12(ji)}$$

s. a. $\lambda_{12(ij)} = \lambda_{12(ji)}$ y $\sum_i \lambda_{1(i)} = \sum_i \lambda_{12(ij)} = 0$

Es importante probar estas hipótesis porque al rechazarlas se tendrían las siguientes implicaciones:

Para (i) cualquier cambio que tenga Y sobre el tiempo se verá afectado en las distribuciones marginales.

Para (ii) la medida de asociación $\theta_{ij(i'j')}$ serían diferentes a través del tiempo.

Para (iii) las frecuencias de las celdas también cambiarán con respecto al tiempo.

Es importante notar que estas pruebas están relacionadas, por ejemplo; el modelo para probar S implica que la HM se cumpla. Otro tipo de relación es que el probar HM y CS es equivalente a probar S. Esta última relación se utiliza para la prueba condicional de HM, gracias a esto, para las estadísticas de prueba se puede establecer la siguiente igualdad

$$T(HM) = T(S) - T(CS)$$

en donde $T(\cdot)$ es la estadística de prueba el modelo.

Ejemplo. Se consideró un reactivo de la prueba RAGS-E, el cual mide hasta qué punto el paciente muestra inestabilidad en el funcionamiento mental (ej. memoria y habilidad de concentración buena un día y pobre al siguiente), para los tiempos basal y fin del tratamiento (semana 12), (ver Tabla 2).

Al probar la hipótesis de HM se puede observar que en los valores marginales para los tiempos basal y fin de tratamiento, la distribución de los sujetos cambia.

De manera intuitiva, la hipótesis de S tampoco se cumple ya que los sujetos se acumulan en la parte inferior izquierda de la tabla.

Los modelos anteriormente presentados fueron probados dando los resultados que se presentan en la Tabla 3.

TABLA 2

Basal / Final	**Para Nada**	**Algo**	**Moderadamente**	**Bastante**	**Total Renglón**
Para Nada	**53**	**8**	**5**	**6**	**72**
Algo	**22**	**23**	**6**	**1**	**52**
Moderadamente	**17**	**11**	**17**	**1**	**46**
Bastante	**9**	**4**	**5**	**14**	**32**
Total Columna	**101**	**46**	**33**	**22**	**202**

TABLA 3

Modelos	G. L.	Estadística de prueba χ^2	Valor -p
Independencia	9	85.402	0.0000
Homogeneidad Marginal	3	14.95	0.0019
Simetría	6	19.62	0.0323
Cuasi- Simetría	3	4.67	0.1976

Se probó el modelo de independencia que como era de esperarse se rechaza. Los modelos de HM y S también se rechazan, lo que nos muestra que el número de sujetos que se clasificaba en alguna categoría en el tiempo inicial, cambió al final del tratamiento, indicando que la administración del fármaco tuvo un efecto significativo en la inestabilidad en el funcionamiento mental.

4. GENERALIZACIÓN

Los modelos anteriores pueden ser aplicados a tablas de contingencia K-dimensionales, incrementando el número de términos a analizar, lo que conduce a particionar la variabilidad en el modelo a probar. Si analizamos el modelo log-lineal para la tabla de $I \times I \times I$, el factor λ_{123} no se considera porque lo que interesa es detectar el momento donde ocurre el cambio, esto sucede cuando alguno de los factores λ_{12}, λ_{13} o λ_{23} es significativo en el modelo.

5. CONCLUSIONES

Los modelos para tablas de contingencia pueden ser adaptados para análisis de cambio.

Estos modelos deben ser a lo más de 5 dimensiones, pues con K mayor, es muy probable que existan celdas con frecuencias 0. Dependiendo de la hipótesis que se rechaze, se puede saber en qué parte de la tabla se da el cambio.

REFERENCIAS

Bishop, Fienberg & Holland.(1975). *Discrete Multivariate Analysis: theory and practice*. Cambridge, Mass.: MIT Press.

Clogg, Eliason & Grego.(1990). Models for Analysis of Change in Discrete Variables. *Statistical Methods in Longitudinal Research*. Volume **2**. Academic Press.

Everitt. (1977). *The Analysis of Contingency Tables*. Chapman & Hall.

Koch & Landis. (1977). A General Methodology for Analysis of Experiments with Repeated Measurement of Categorical Data. *Biometrics* **33**, 133-158.

SAS/STAT® User's Guide, Vol **1**, Version 6: CATMOD Procedure. (1990). *SAS Institute Inc.*

Descomposición de la Interacción en Tablas de Doble Entrada

IGNACIO MÉNDEZ R

IIMAS-UNAM, México

1. INTRODUCCIÓN

En la investigación es frecuente que la información numérica se presente en tablas de doble entrada. Se considera una observación por celda, el modelo aditivo es muy usado, dado que no es posible la estimación de todos los términos de interacción. Esto obedece a la falsa conceptualización de que la interacción es de "todo o nada". Sin embargo, es factible que la interacción se reduzca a una o una cuantas celdas. Esta situación se confunde totalmente con la ocurrencia de observaciones atípicas. En este trabajo se plantean algunas formas para identificar la fuente de la interacción u observaciones discordantes del modelo aditivo y también para modelar esa situación.

2. MODELOS Y ESTRATEGIA CONSIDERADAS

2.1. Modelo aditivo con variables indicadoras. El modelo general es : $y_{ij} = \mu_{ij} + \varepsilon_{ij}$. La media de la celda ij en forma aditiva : $\mu_{ij} = \mu + \alpha_i + \beta_j$, donde y_{ij} es la medición en el elemento de la celda con nivel i del factor A y el nivel j del factor B. μ_{ij} es la media de la celda ij . Se supone que el término aleatorio tiene distribución N(0,1), además que α_i y β_j son los efectos principales de los factores de clasificación.

2.2. Identificación de celdas que interactuan (discordantes). Para esto se usa la detección de observaciones discordantes. También se elaboran gráficas (q-q) para normalidad de los residuos, observaciones muy alejadas de la recta esperada se consideran discordantes. Otra técnica utiliza la mediana de las tetradas que involucran a cada celda, donde estas tetradas son las diferencias entre renglones para una misma columna o viceversa: $y_{ij} - y_{i'j} - (y_{ij'} - y_{i'j'})$. $\forall ii,jj$. Bradu y Hawkins (1982) consideran que las celdas con medianas notoriamente mayores que el resto son indicativas de celdas discordantes con el modelo aditivo. Johnson y Graybill (1972) y también Bradu y Gabriel (1978) desarrollan pruebas estadísticas sobre esas tetradas para la nulidad simultánea de sus esperanzas. Daniel (1978) y también Gentleman y Wilk (1975) utilizan la idea de que una o dos celdas que provocan interacción o que son discordantes con el modelo aditivo producen un patrón específico en los residuos. Como otra posibilidad Gabriel (1978) y Mandel en trabajos anteriores, proponen la descomposición espectral de la matriz de residuos. Algo no señalado por Gabriel pero si en Milliken y Johnson (1989) Vol. 2, es que a partir de los vectores propios se puede sugerir la identificación de aquellas hileras y columnas que se comportan diferente que el resto. Una variante empírica que propongo es usar los "componentes principales " de hileras y columnas, los que también tienden a ser muy diferentes para las celdas discordantes, lo que también identifica celdas que no se ajustan al modelo aditivo. El uso de gráficas de los datos y de los residuos, como análisis descriptivo inicial, da una idea de las celdas interactuantes (o con observaciones discordantes).

2.3. Modelo con variables indicadoras para observaciones discordantes. Algunos como Johnson y Graybill (1972), proponen eliminar del análisis la celda (o celdas discordantes).

considero esto erróneo, ya que es preferible el uso de modelos como $\mu_{ij} = \mu + \alpha_i + \beta_j + \delta_{1i(ab)} + ..+\delta_{pi(a'b')}$. Los términos adicionales δ^S son los coeficientes que multiplican a i(ij), variables indicadoras de celdas que no se ajustan al modelo aditivo. Barnett y Lewis (1994) proponen un modelo con una observación discordante que es semejante al anterior.

2.4. Modelos de Mandel. Mandel (1995) y en trabajos anteriores utiliza el ajuste del modelo aditivo, en particular los efectos de hilera y de columna estimados para usarlos de diversas maneras, como parte de las variables independientes en el modelo. Además también propuso el modelo de descomposición espectral $\mu_{ij} = \mu + \alpha_i + \beta_j + \Sigma_k \theta_k u_{ki} v_{kj}$, donde los θ k son parámetros que corresponden a los valores propios de la descomposición espectral de la matriz de residuos, y u_{ki} , v_{kj} son los vectores propios de hileras y columnas. El modelo lineal en hileras y ij = $a_i + b_i.c_j + \varepsilon_{ij}$, equivale de acuerdo a Milliken y Johnson (1989) al modelo: $y_{ij} = \mu + a_i + b_j + b_j.\alpha_i + \varepsilon_{ij}$, donde a_i y b_j son los efectos de hilera y columna en forma categórica y α_i. es el efecto de hilera en forma continua. El modelo lineal en columnas es y ij = bj + aj.ci + ε_{ij}. Estos se pueden expresar en forma canónica. Así, el lineal en hileras es igual a: $y_{ij} = \bar{y}_i + b_i(\bar{x}_j - \bar{x}) + d_{ij}$, donde $\bar{y}_i$ es la media de hilera y $\bar{x}_j$ es la de la columna j , $\bar{x}$ la media general. Además b_i es el coeficiente de regresión de los datos de la hilera i sobre las medias $\bar{x}_j$ y d_{ij} son los residuos de la regresión. Es importante señalar que si se tienen ordenadas al origen iguales para todas las regresiones, se tiene el modelo de Tukey de "un grado de no aditividad". Si a_i y b_j son categóricas y α_i, β_j son numéricas , el modelo de Tukey es : $y_{ij} = \mu + a_i + bj + \theta.\alpha_i.\beta_j + \varepsilon_{ij}$. Es importante señalar que Mandel introduce términos adicionales a los efectos principales que resultan ortogonales a estos.

2.5. Modelo de contrastes ortogonales. El uso de contrastes ortogonales para efectos principales que capturen los niveles donde está la celda o celdas que parecen producir la interacción, por supuesto, tantos como los grados de libertad. Así, la interacción se descompone en productos de los contrastes ortogonales para los efectos principales. El modelo propuesto es $y_{ij} = \mu + \alpha_i + \beta_j + \Sigma_k c_{ki}.\ c_{kj} + \varepsilon_{ij}$, donde los c^S son los contrastes y se hace una selección de ellos, para incluir en el modelo solo aquellos que son muy significativos y producen residuos bien comportados.

2.6. Bloques a posteriori. El modelo de variables indicadoras para las celdas discordantes (o interactuantes) equivale a usar bloques de tamaño uno, a posteriori, para cada celda. Como una extensión empírica de este modelo, propongo de modo exploratorio, el uso de bloques de cualquier tamaño que agrupen las celdas con residuos "semejantes". Estos bloques se pueden encontrar aplicando una técnica de formación de conglomerados con los valores de los residuos del modelo aditivo. El número de conglomerados, es decir, de bloques a posteriori, dependerá de los grados de libertad del error (que no resulten muy pequeños) y de las propiedades estadísticas, como el CME y la normalidad de los residuos obtenidos al ajustar un modelo con efectos principales y con los bloques a posteriori propuestos. La idea puede parecer inconveniente por depender el modelo final (los bloques) del primer ajuste como modelo aditivo, por lo que hay un riesgo de modelar la aleatoriedad presente y producir un sobreajuste. Sin embargo, se propone como un método exploratorio y para ser comparado con el ajuste de otros modelos. Además prácticamente todos los modelos propuestos por Mandel y los autores citados, también tienen el mismo inconveniente.

2.7. Criterios para selección de modelos. Se usan la magnitud del cuadrado medio del error, CME; el error estándar de los efectos principales; la amplitud y magnitud de los residuos; y la cercanía de ellos a la normalidad, medida como la cercanía a uno en el valor de p para la prueba de Shapiro-Wilks aplicada a los residuos estandarizados. Además también la apariencia de las gráficas de residuos y los valores de D de Cook.

3. APLICACIÓN DETALLADA A UN EJEMPLO

Se toman los datos de un ejemplo que aparece en el libro de Mandel (1995) y el artículo de Johnson y Graybill (1972). Son datos del rendimiento o la producción de trigo bajo 5 niveles de fósforo y 3 de nitrógeno. Los tratamientos son mezcla de fertilizantes agregados a macetas.

3.1. Análisis exploratorio inicial.

Las gráficas con una línea para cada renglón indicaron con bastante claridad que la hilera 1 y la columna 1 son las que interactuan.

3.2. Modelo aditivo.

Al ajustar el modelo aditivo, es claro en los residuos que la celda 1,1 es discordante. El residuo estandarizado es de 2.579, el cual es significativo al 1%, según tabla XXXVII de Barnett y Lewis (1994). Adicionalmente el valor de p en la prueba de Shapiro-Wilks (S-W) es de .2138. Se puede concluir que hay un alejamiento de la normalidad muy probablemente causado por una observación que interactua o es discordante.

3.3 Modelo con variable indicadora.

En el artículo de Johnson y Graybill (1972), se detecta el mismo residuo como discordante y se propone el análisis eliminándolo. Así, el cuadrado medio del error pasa de 32331 con 8 g.l. del aditivo a 6228 con 7 g.l., con una p en S-W para los residuos estandarizados de .71, mucho mejor ajuste. Sin embargo, se pierde la información de esa unidad. Como una propuesta alternativa se usa un modelo con una variable indicadora para esa observación. El ajuste en términos de los residuos es exactamente el mismo que eliminando la observación discordante. Sin embargo, hay un cambio en las medias ajustadas de los efectos principales y sus errores estándar. En la Tabla 1 se presenta una comparación de las estimaciones de las medias de n y p (solo tres niveles de p) con sus errores estándar. Además, desde un punto de vista biológico es mejor el modelo con la indicadora que la eliminación del dato discordante , ya que su información no se pierde, solo se matiza.

3.4. Modelos de Mandel.

3.4.1. Vectores propios y componentes principales.

Al efectuar el ajuste del modelo de descomposición espectral, para los niveles (0,45,90) de n las u_i, los vectores propios son respectivamente: (-.16245 , .09552, .066943) y para p en sus niveles (0,22,45,90,180), los valores de v_j son (-.05487, -.00597 .02749 ,.023254, .010104). Al usar el producto u_i por v_j como una covariable, se obtiene un ajuste muy bueno, con un CME de 10771, con 7 g.l. pero residuos mal comportados, con una p en S-W de .48. Se considera que hay alejamiento de la normalidad. Como una variante heurística del modelo anterior, se propone usar como una covariable al producto de el primer componente principal de hileras con el de columnas. El modelo es semejante al de Mandel. Este modelo produjo un CME de 648 con 7 g.l., por lo que es mucho mas eficiente que el de

Tukey. Sin embargo, produce alejamiento fuerte de la normalidad en los residuos estandarizados, con una p en s-w de .16 . 3.4.2 moldelos lineal por columna y viceversa. El ajuste del modelo lineal por hilera, produce una reducción drástica del CME, y residuos muy cercanos a la normalidad , con un CME de 347 con 4 g.l. y una p en S-W de .43. Los ajustes del modelo de Tukey (un g.l. de no aditividad) y el linear por columna, son muy semejantes, con CME de 3269 y 3749, respectivamente, y valores de p en S-W de .63 y .47. Esto seleccionaría como aceptable el modelo lineal por hilera.

3.4.2 Contrastes ortogonales.

Se plantean contrastes que involucren a la hilera y columna 1, y otros que no lo hagan. Así, para los niveles de n los contrastes son (-2,1,1) y (0,1.-1) llamados n1 , n2. Para fósforo se tiene como f1 , (-4,1,1,1,1), y como f2 , f3, y f4 respectivamente (0,-1.-1,1,1) , (0,-1,1,-1,1) y (0,1,-1,-1,1). Así, una partición ortogonal de la interacción se efectúa con los contrastes que resultan del producto de cada uno de los de n con cada uno de los de f. Con el supuesto de normalidad y bajo hipótesis de no interacción, estos tienen distribución normal independiente, se eliminan los efectos mas significativos y que producen alejamiento de la normalidad del grupo de contrastes. Además se corre el ajuste de varios modelos con diversos contrastes en la interacción, buscando bajo CME y cercanía a la normalidad de los residuos. Con este proceso, resulta como un modelo adecuado el siguiente: $y_{ij} = \mu + n1 + n2 + f1. + f2 + f3 + f4 + n1.f1 + n1.f4 + \varepsilon_{ij}$. Este modelo produce un CME de 1986 y un valor de p en S-W de .996.

3.5 Bloques a posteriori.

A partir de los residuos del modelo aditivo, se aplico un análisis de conglomerados con distancia de Ward y se identificaron tres, el modelo con ellos tiene propiedades aceptables, un CME de 1986 con 6 g.l. y residuos con una p en S-W de .996 .

4. COMPARACIÓN DE LOS DIVERSOS MODELOS

En las Tablas 1 y 2 se presentan los principales resultados para el ajuste de los modelos y las medias estimadas de efectos principales. Se presentan algunos otros modelos no discutidos, como una referencia sobre el valor del estimador de la varianza del error vía el CME. En general se reafirma la conclusión de que el modelo de contrastes es bastante adecuado. Otro modelo que parece bueno fue el de tres bloques a posteriori. Ambos tienen buenas propiedades estadísticas.

5. CONCLUSIONES

El modelo lineal en hileras produce un buen ajuste pero no es muy interpretable. Con la eliminación del dato discordante con modelo aditivo también hay buen ajuste, pero se pierde la información de esa celda. En cambio, si se usa una variable indicadora de esa celda, el ajuste es el mismo pero no se pierde la información para los efectos principales, como un buen modelo interpretable, que no pierde información y produce un buen ajuste se tiene el de contrastes ortogonales. En general se recomienda explorar varios modelos, como los comentados en este escrito, siempre que el modelo aditivo presente un ajuste inadecuado.

TABLA 1

Resumen de Modelos.

Modelo	R^2	G. L.	CME	P (S-W)	Residuos	D Max.
Aditivo Aditivo	.94	8	32331	.21	-205 a 339	.831
Indic.(1,1)	.989	7	6228	.68	-86 a 23.7	.614
Lineal Por Col	.995	6	3749	.79	-68.5 a 89	
Lineal Por Hil	.9997	4	347	.89	-21.2 a 17.9	23.21
Concurr.(Tukey)	.995	7	3269	.63	-69.1 a 88	
Aditivo Sin Col 1	.959	6	6714	.53	-86.6 a 123.7	
Aditivo Sin Hil 1	.999	4	888	.97	-36.4 a 36.4	
Aditivo Sin H Yc 11	.999	3	80	.73	-8 a 8	
Contrastes	.997	6	1986	.996	-51 a 51	.873
3blq. Posteriori	.997	6	1986	.996	-51 a 5	.873
Vectores Propios	.998	7	11484	.48	-208 a 114.3	1.52
Comp. Principal.	.9989	7	648	.16	-42.5 a 40.2	1.09

TABLA 2

Medias Ajustadas y Errores Estándar para Efectos Principales.

	n					
Modelo	Medias			Errores Estándar		
	0	45	90	0	45	90
Aditivo	2586.8	2923.2	2871.4	80.4	80.4	80.4
Indicad 1,1	2502.1	2965.5	2923.7	38.1	36.0	36.0
Vector.Propi	2586.8	2923.2	2871.4	5.1	5.1	5.1
Comp.Princ.	2586.8	2923.2	2871.4	11.4	11.4	11.4
Contrastes	2586.8	2923.2	2871.4	19.9	19.9	19.9
3blq Poster	2642.4	2996.8	2945.0	22.7	44.8	44.8
Mandel	2586.8	2923.2	2871.4	25.6	25.6	25.6
Linear Por Hil	2586.8	2923.2	2871.4	8.3	8.3	8.3
Linear Por Col	2586.8	2923.2	2871.4	27.4	27.4	27.4

	p					
Medias				E.E.		
	0	45	180	0	45	180
Aditivo	1852.3	3084	3224	103.8	103.8	103.8
Indica 1,1	1683.0	3126	3266	53.9	46.1	46.1
Vec.Propios	1852.3	3084	3224	61.9	61.9	61.9
Comp.Princ.	1852.3	3084	3224	14.7	14.7	14.7
Contrastes	1852.3	3084	3224	25.7	25.7	25.7
3blq. Poster	1925	3186.7	3254	31.5	42.7	56.1
Mandel	1852.3	3084	3224	33	33	33
Lineal Por Hil.	1852.3	3084	3224	10.8	10.8	10.8
Lineal Por Col	1852.3	3084	3224	35.4	35.4	35.4

REFERENCIAS

Bradu And Hawkins (1982) Location Of Multiple Outliers In Two-Way Tables, Using Tetrades. *Technometrics*. **24**:103'108.

Bradu D. And K.R. Gabriel (1974) Simultaneous Statistical Inference On Interactions In Two-Way Analysis Of Variance. *JASA 69*:428-436.

Bradu D. And K.R. Gabriel (1978) The Biplot As A Diagnostic Tool For Models Of Two-Way Tables. *Tecnometrics* **20**:47-68.

Cox. D.R. (1984) Interaction. *Int. Stat.Rev.* **52**:1-31

Daniel C. (1978) 'Patterns In Residuals In The Two-Way Layout". *Technometrics* **20**:386-395

Gabriel K.R. (1978) 'Least Squares Approximation Of Matrices By Additive And Multiplicative Models" *J.R.Statis.Soc*.B **40**:186-196.

Gentleman J.F. And M.B. Wilk (1975) Detecting Outliers In A Two-Way Table *: 1* Statistical Behavior Of Residu als *Tecnometrics* **17**:1-14.

Johnson D.E. And F.A. Graybill (1972) Estimation Of σ In A Two-Way Classification Model With Interaction *Jasa* **67**:388-394.

Mandel J. (1995) *Analysis Of Two-Way Layouts* Chapman And Hall.

Milliken G.A. And D.E. Johnson (1989) *Analysis Of Messy Data* Vol 2 Van Nostrand Reinhold. New York.

Barnett V. And T. Lewis (1994) *Outliers In Statistical Data.* John Wiley And Sons

Una Comparación de Tres Métodos de Clasificación

LUIS ENRIQUE NIETO B. y MARIO CORTINA-BORJA

ITAM, México

1. INTRODUCCIÓN

Un problema que aparece comúnmente en la práctica es el de construir una clasificación con k clases basada en una muestra aleatoria de una variable aleatoria multivariada. Para esta muestra se conoce la clase de la cual proviene cada observación. El propósito de esta clasificación es el de encontrar una regla de asignación que permita decidir a qué clase pertenece un nuevo elemento de la población.

En este trabajo comparamos tres métodos de clasificación multivariada: a) Método discriminante de Fisher (Mardia et al. 1988); b) Arboles de clasificación (CART) (Breiman et al. 1984); c) Método de medianas generalizadas.

Los datos utilizados en este ejercicio fueron tomados de la base de datos SIMM90 recopilada por el Consejo Nacional de Población (CONAPO, 1994) a partir del XI Censo de Población y Vivienda, 1990. La población total corresponde a los 2403 municipios del país. Las variables consideradas para construir la clasificación fueron: 1) % de la población mayor de 15 años sin primaria completa (S/PRI); 2) % de ocupantes en vivienda sin drenaje ni excusado (S/EXC); 3) % de ocupantes en vivienda con piso de tierra (PISOT); 4) % de población ocupada con ingreso menor a dos salarios mínimos. Para definir a las clases en la población nos basamos en el grado de marginación (muy alto, alto, medio, bajo y muy bajo) definido por CONAPO a partir de su índice de marginación.

Nuestro objetivo es predecir el grado de marginación municipal a partir de un subconjunto de cuatro variables empleadas por CONAPO, utilizando el modelo de análisis discriminante clásico y el modelo de un árbol de clasificación no paramétrico.

Para construir la clasificación obtuvimos una **muestra de aprendizaje** que consistió de 311 municipios. Dicha muestra se obtuvo utilizando muestreo estratificado de mínima varianza. Los tamaños de muestra en cada estrato fueron:

MA--34; A -- 108; M -- 58; B -- 95; MB -- 16.

Estos tamaños de muestra se calcularon a partir de un intervalo de confianza de 95% para la media de cada una de las variables; El tamaño de muestra final se tomó como el mayor de los tamaños para cada variable.

Primeramente mostramos, mediante pruebas no paramétricas, que los supuestos de normalidad y de homogeneidad de matrices de varianza-covarianza no se satisfacen para estos datos. Con esta muestra se calcularon las funciones discriminantes lineales y el árbol de clasificación correspondientes. Los estimadores obvios del error de clasificación construidos a partir de las predicciones de los modelos para los datos muestrales son sumamente sesgados. Además de calcular estimadores del error de clasificación basados en métodos de remuestreo sobre el conjunto de aprendizaje, comparamos los métodos de clasificación por su desempeño en la predicción del grado de marginación para el resto de los municipios del país.

A continuación describimos los métodos de clasificación que deseamos comparar.

2. MÉTODO DISCRIMINANTE DE FISHER

Este método busca clasificar y separar k subpoblaciones definidas por p variables por medio de ecuaciones lineales que definen a los vectores discriminantes. La idea básica para construir estas ecuaciones es encontrar al vector que maximice el cociente entre la variabilidad entre clases y la variabilidad dentro de cada clase. Una vez encontrado este primer vector discriminante se obtiene otro vector que maximice al mismo cociente sujeto a que su covarianza con el primer vector discriminante sea 0. Este proceso se continúa hasta completar $s = min\ (k\text{-}1,\ p)$ vectores ortogonales. Este método supone que los grupos de los que provienen las observaciones tienen una matriz de varianza-covarianza común. Es frecuente que este supuesto no se satisfaga. Por otra parte, las pruebas usuales de homogeneidad para matrices de varianza-covarianza suponen que los datos provienen de distribuciones normales multivariadas, lo cual puede tampoco satisfacerse en la práctica.

Puede verse que el primer vector que maximiza el cociente de variabilidades es el primer vector propio de la matriz $\Sigma^{-1}B$ donde Σ es la matriz de varianza-covarianza dentro de las clases y B es la matriz de varianza-covarianza entre las clases. El método supone que $\Sigma_1 = \Sigma_2 = \cdots = \Sigma_k = \Sigma$, es decir, que las k clases tienen variabilidad homogénea.

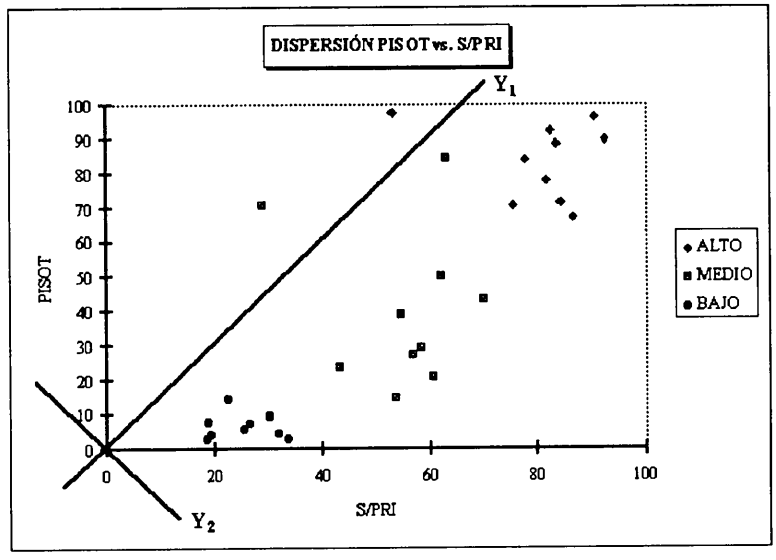


Fig.1

El método de Fisher es intuitivamente fácil de entender porque se basa en la idea de clasificar en términos de combinaciones lineales de las p variables que explican la variabilidad de los datos dentro de cada clase. Para asignar un nuevo elemento a una clase se proyecta su valor original en el espacio definido por los s vectores discriminantes y se le asocia la clase correspondiente al centroide más cercano en este espacio.

El siguiente ejemplo muestra una aplicación de este método en dos dimensiones (S/PRI y PISOT) con 3 grupos (A, M, B) para una muestra de 30 municipios.

En la Figura 1 aparecen los datos graficados en el espacio original junto con los s= 2 vectores discriminantes. La Figura 2 muestra los datos proyectados sobre los dos vectores discriminantes.

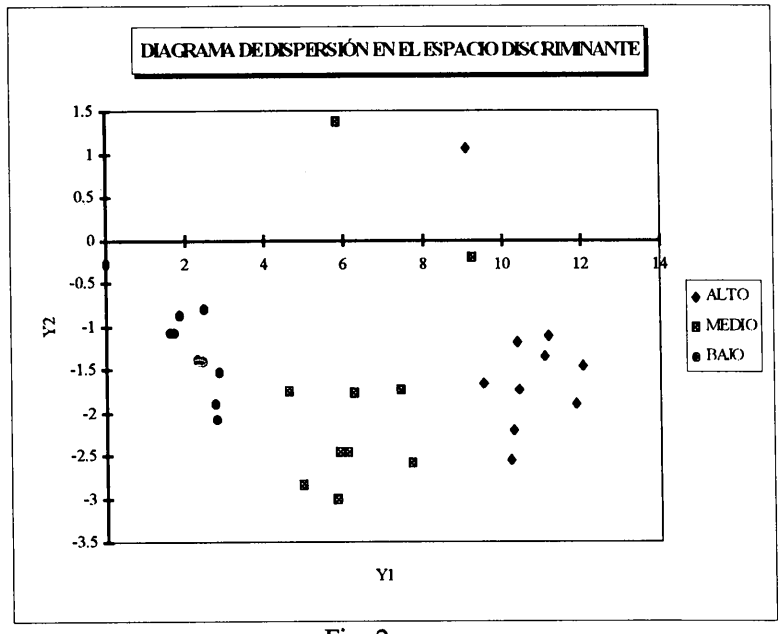


Fig. 2

3. ARBOLES DE CLASIFICACIÓN (CART)

Un método no paramétrico para construir clasificaciones es el conocido por CART (*Classification and Regression Trees)* y fue propuesto en 1984 por Breiman, Friedman, Olshen y Stone. Este método construye clasificaciones basadas en particiones binarias en los rangos de las variables. Las particiones se obtienen en términos del grado de discriminación que hacen las variables respecto a los grupos definidos en la muestra. A pesar de su gran potencial para aplicaciones, el método ha tenido relativamente poca difusión.

CART abarca dos formas posibles de clasificar: una es teniendo una clasificación previa de los individuos (i.e. una variable respuesta categórica) y la otra se basa en una variable respuesta continua relacionada con las otras p variables mediante un modelo de regresión. En el primer caso se tiene un árbol de clasificación y en el segundo un árbol de regresión.

Para construir estos árboles, CART se basa en biparticiones del conjunto total de datos con que se cuenta (la muestra de aprendizaje). Cada grupo resultante se biparte hasta llegar a un cierto tope. Cada partición está definida en términos de una de las p variables explicativas. Cada subgrupo generado por las biparticiones es un nodo del árbol. En cada nodo se tiene a las observaciones que cumplen (o no) la característica que define a una partición binaria. Un nodo es terminal cuando no es conveniente continuar realizando más particiones. El criterio para decidir cuándo seguir partiendo se basa en una medida de impuridad que se relaciona con la devianza del modelo de regresión. A cada nodo terminal se le asocia con un valor de la variable respuesta. En el caso de que esta sea categórica, el

valor corresponde a la clase de la mayor parte de los individuos que están en ese nodo terminal; en el caso de un árbol de regresión, el valor del nodo terminal es el promedio de los valores de la variable respuesta para los individuos del nodo.

Un árbol de clasificación construido por CART asigna a un nuevo individuo siguiendo las particiones binarias del árbol que se formó hasta encontrar el nodo terminal en el que se aloja. La clase asignada al nuevo individuo es la clase de ese nodo terminal.

La clasificación resultante con CART es equivalente a dividir el espacio original de las variables explicativas mediante hiperplanos ortogonales. El número de estos hiperplanos puede ser muy grande (pero finito). Esto contrasta con el método de Fisher que utiliza solamente $s = min\ (g-1,\ p)$ vectores discriminantes ortogonales. Además, CART pude hacer uso de variables categóricas como parte de las variables explicativas para construir la clasificación, mientras que el método de Fisher supone que las variables son continuas. Finalmente, el método de Fisher impone a los datos un modelo de homogeneidad en las matrices de varianza covarianza internas. CART no hace supuesto alguno respecto a la estructura de los datos.

Las Figuras 3 y 4 muestran un ejemplo de aplicación de CART con los datos utilizados en las Figuras 1 y 2. La Figura 3 muestra el árbol de clasificación resultante y la Figura 4 muestra el conjunto de particiones en el espacio original.

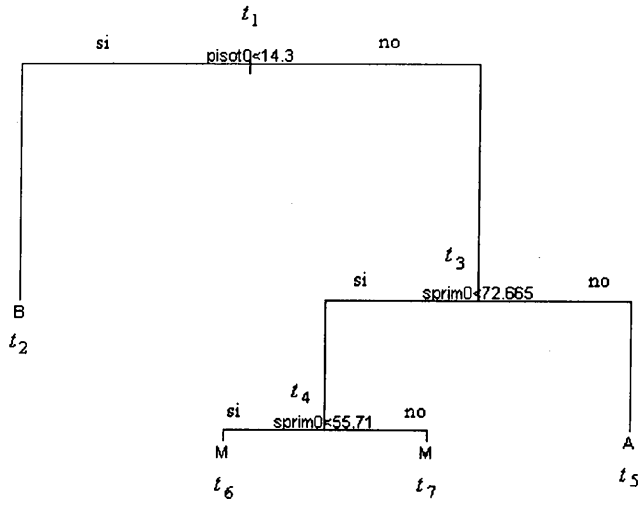


Fig. 3

4. MÉTODO DE MEDIANAS GENERALIZADAS

Un método de clasificación muy sencillo es el siguiente: considere el valor θ que minimiza la siguiente expresión: $\sum_{i=1}^{n} \|X_i - \theta\|$. Para el caso univariado, θ coincide con la mediana muestral. Por analogía, podemos definir a una mediana multivariada como el valor θ que minimiza esta suma considerando una norma en el espacio multivariado. En este ejemplo utilizamos la norma $\mathscr{L}_2$. La mediana muestral θ se obtiene utilizando el método de Newton-Raphson para minimizar la suma de distancias (Bedall y Zimmermann, 1978).

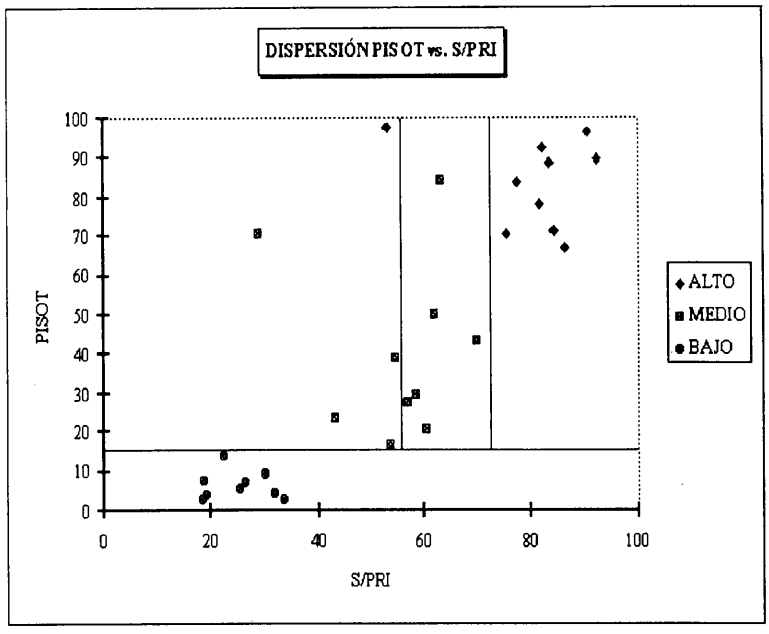


Fig. 4

Este método es robusto frente a la presencia de observaciones discrepantes. Para cada clase obtuvimos la mediana utilizando la muestra de aprendizaje. La forma en que este método asigna un nuevo elemento a determinada clase consiste en simplemente encontrar la mediana generalizada de la muestra de aprendizaje más cercana en términos de distancia euclidiana a este nuevo elemento.

5. CONCLUSIONES

Evaluamos el comportamiento de cada método de dos formas: en primer lugar obtuvimos un estimador del error de clasificación reclasificando a los elementos de la muestra de aprendizaje. Esto da un estimador insatisfactorio, puesto que los datos con los que se construyó la clasificación son ahora utilizados para evaluarla. Cabe mencionar que este método de estimación da una estimación sesgada del error de clasificación.

En este caso contábamos con la población completa y con el valor de la variable respuesta para cada elemento, por lo que la evaluación final de los métodos se realizó clasificando a los 2092 municipios que no entraron en la muestra de aprendizaje. Esta es una forma de encontrar el verdadero valor del error de clasificación para cada uno de los métodos.

En la Figura 5 se encuentran graficados los dos primeros vectores discriminantes, ya que en ellos se concentra el 99.84% de la variabilidad total. La Figura 6 contiene al árbol de clasificación obtenido para la muestra de aprendizaje.

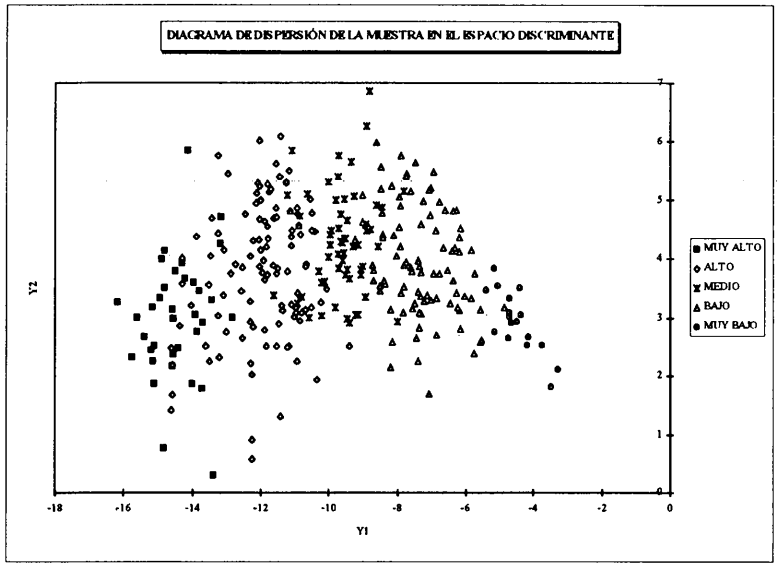


Fig. 5

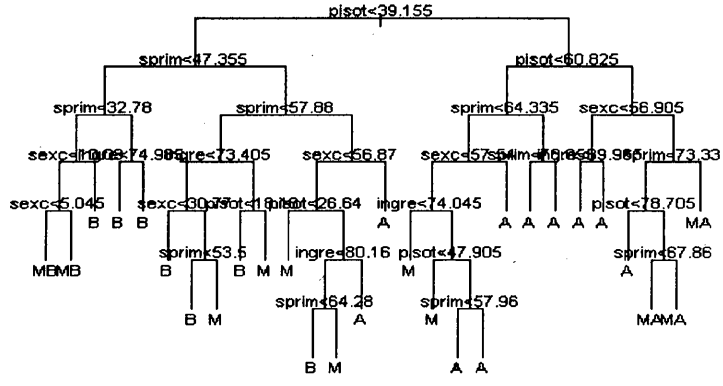


Fig. 6

Los valores del estimador de resustitución $R(d)$ y los valores reales del error de estimación $R^*(d)$ para los tres métodos considerados fueron:

	$R(d)$	$R^*(d)$
MÉTODO DE FISHER	19.61%	21.79%
ARBOLES DE CLASIFICACIÓN	9.97%	24.56%
MEDIANAS GENERALIZADAS	23.15%	24.08%

Los errores de clasificación calculados para el resto de la población son similares para los tres métodos.

REFERENCIAS

Bedall, F.K. y Zimmermann, H. (1978) Algorithm AS 143: The Mediancentre. *Applied Statistics*, **31**, 169-173.

Breiman, L., Friedman, J.H., Olshen, R. y Stone C. (1984) *Classification and Regression Trees.* Wadsworth International Group,

Consejo Nacional de Población (1994) *Sistema automatizado de información sobre la marginación en México.* CONAPO, México DF.

Mardia, K.V., Kent, J.T. y Bibby, J.M.(1988) *Multivariate Analysis*, Academic Press, San Diego.

Pruebas de Significancia en Factoriales No-replicados Usando Graficación Semi-normal

JORGE OLGUÍN

IIMAS-UNAM, México

1. INTRODUCCIÓN

Durante la última década ha habido un creciente interés por desarrollar métodos para el análisis de experimentos factoriales fraccionados y otros diseños ortogonales que por realizarse sin réplica se han denominado *factoriales no-replicados*.

Estos diseños se usan principalmente cuando se puede suponer la inexistencia de interacciones de varios órdenes por lo que mediciones de las mismas estarían solamente en función del error experimental. Parte de la estrategia entonces, es utilizar un diseño cuyos contrastes evalúen conjuntamente (confundan) los efectos de interés con dichas interacciones. Por lo tanto, para la elección de un diseño adecuado son necesarios, tanto un conocimiento previo del fenómeno bajo estudio, como una planeación cuidadosa. De otro modo, los patrones de confusión podrían disminuir considerablemente la utilidad del experimento. La principal dificultad para el análisis de los experimentos factoriales no replicados es la ausencia del estimador de la varianza del error experimental, estimador que en otros diseños se obtiene con base en las réplicas del experimento.

Por otra parte, debido a la cantidad de factores involucrados, sobre todo en las etapas iniciales de una investigación, los investigadores utilizan con frecuencia diseños con un alto grado de saturación en los que cada contraste individual es potencialmente activo. Sin embargo, con base en la experiencia, se ha reconocido que en aplicaciones industriales con frecuencia se cumple lo que se conoce como *esparcidad de efectos*, de acuerdo con la cual en un factorial no replicado normalmente se espera que solamente una pequeña proporción de efectos sean activos.

Extendiendo las ideas de Daniel (1959) y Zahn (1975) a continuación se presenta un método general para realizar pruebas de significancia de contrastes en factoriales fraccionados y otros diseños ortogonales utilizando la gráfica semi-normal.

2. MÉTODO PROPUESTO

Sean $Y_1, \ldots, Y_m$ contrastes ortogonales obtenidos de un factorial no replicado con $n = m + 1$ observaciones y supóngase que los contrastes han sido calculados de modo que todos tengan la misma varianza. Bajo las suposiciones usuales de los modelos de análisis de varianza $Y_1, \ldots, Y_m$ son variables aleatorias independientes normalmente distribuidas con medias $\mu_1, \ldots, \mu_m$ y varianza desconocida σ^2. Se sabe que la mayor parte de las medias son cero, por lo que el problema relevante es inferir sobre la base de un conjunto de valores observados de $Y_1, \ldots, Y_m$ cuales de las medias son diferentes de cero (si las hay).

Sean $V_{1:m}, \ldots, V_{m:m}$ las estadíisticas de orden de los contrastes en valor absoluto y sean $W_{1:m}, \ldots, W_{m:m}$ los valores esperados de las estadísticas de orden de una muestra de tamaño m de la distribución seminormal estándar. El procedimiento utiliza secuencialmente las siguientes estadísticas de prueba:

$$T\left(k,b(m)\right) = \frac{V_{k:m}}{S\left(b(m),m\right)}, \quad k = m, m-1, \ldots, b(m)+1 \tag{1}$$

donde

$$S\left(b(m),m\right) = \frac{\sum_{i=1}^{b(m)} V_{i:m} W_{i:m}}{\sum_{i=1}^{b(m)} W_{i:m}^2} \tag{2}$$

Nótese que (2) es la pendiente de la regresión hacia el origen usando los puntos $(W_{i:m}, V_{i:m})$, $i = 1, \ldots, b_{(m)}$.

El primer paso del proceso de detección consiste en comparar el valor de $T_{(m,b(m))}$ con cierto valor crítico $c_{m,b(m)}$ el cual controla, a un cierto nivel γ, la probabilidad de obtener al menos un falso positivo en un experimento en el que todas las medias de los contrastes son cero (experimento nulo). Se tiene entonces que, bajo la suposición de que todos los contrastes son nulos, $c_{m,b(m)}$ es tal que

$$\Pr\left\{T\left(m-1,b(m)\right) < c_{m-1,b(m)}\right\} = 1-\gamma \tag{3}$$

para cierto valor (nivel de significancia) γ.

Si la estadística $T_{(m,b(m))}$ es menor que el valor crítico $c_{m,b(m)}$, todos los contrastes son declarados nulos y el procedimiento termina. De otro modo, el contraste mayor (en valor absoluto) es declarado significativo y el procedimiento continúa con la comparación de $T_{(m-1,b(m))}$ con el valor crítico $c_{m-1,b(m)}$ el cual, bajo la suposición de que los restantes m-1 contrastes son nulos, es tal que

$$Pr\left\{T\left(m-1,b(m)\right) < c_{m-1,b(m)}\right\} = 1-\gamma \tag{4}$$

y así sucesivamente. El procedimiento termina cuando una estadística de prueba es menor que su valor crítico correspondiente, o bien, cuando se alcanza la estadística $T_{(b(m),b(m))}$. En consecuencia el procedimiento supone que no habría más de m-$b(m)$ contrastes activos en un experimento. Por otra parte, la presencia de más de m-$b(m)$ contrastes activos, contaminaría el valor de (2) lo que reduciría la potencia del método. De aquí la importancia de seleccionar las valores de b(m) de manera realista.

Un análisis de más de 100 experimentos reales tomados de la literatura estadística y de otras fuentes realizado por Olguín (1994), sugiere tomar $b(m) \approx 0.6 \times m$. A este caso particular en adelante se le denominará método HP.

Nótese que la computación de (2) requiere de los valores esperados de las estadísticas de orden de muestras de la distribución semi-normal estándar (EOS). Debido a que las (EOS) han sido tabuladas solamente para un número reducido de tamaños de muestra, para la aplicación de HP éstas fueron obtenidas aplicando para esta distribución una fórmula general presentada por David (1970). Los detalles aparecen en el Olguín (1994).

Se han obtenido mediante simulación valores críticos con PER[1] de 0.05, 0.20 y 0.40 (los sugeridos por Daniel) para 10 tamaños de diseños. Estos valores críticos así como una macro en MINITAB para la utilización de este método se pueden adquirir solicitándolos al autor.

[1] PER es la probabilidad de declarar uno o más contrastes como activos cuando los m contrastes de un experimento son nulos.

2.1 Ejemplo

Para ilustrar el uso de este método considérese el ejemplo reportado por Grove y Davis (1992) en el cual se utilizó un diseño factorial fraccionado 2_{IV}^{8-4} para investigar los efectos de 8 factores a dos niveles sobre las rugosidades resultantes en la manufactura de cubiertas para guanteras de automóviles.

En la Tabla 1 se muestran los valores absolutos de los contrastes en orden de magnitud (estadísticas de prueba), los valores esperados de las estadísticas de orden de la distribución semi-normal estándar (EOS) para m = 15, los 8 factores abreviados con las letras A . . . H, y el patrón de confusión incluyendo interacciones hasta orden 2 ya que las de mayor orden se consideraron nulas por los expertos responsables del experimento.

La Figura 1 muestra la gráfica semi-normal incluyendo líneas para realizar pruebas de significancia con valores de PER = 0.40, 0.20 y 0.05. Las líneas se construyen como sigue. Se obtiene la pendiente de la regresión hacia el origen de las 9 menores estadísticas de orden sobre sus correspondientes (EOS); ésta es S(9,15) = 0.635. Entonces, para cada valor de PER los valores críticos correspondientes se multiplican por 0.635. Estos puntos se grafican sobre las 6 mayores EOS y finalmente los puntos se unen con líneas para aumentar el impacto visual.

El proceso de detección se realiza en la gráfica de derecha a izquierda. Para el nivel de significancia deseado, el contraste mayor en valor absoluto (i.e. la mayor estadística de orden) se compara con la línea correspondiente. Si éste está sobre la línea, se declara significativo y se procede con el siguiente. Cuando se encuentra la primera estadística de orden debajo de la línea (o cuando ésta ha terminado) el contraste correspondiente y todos los de magnitud inferior son declarados como ``no significativos". En este ejemplo claramente los efectos principales B (Temperatura del fundido) y C (Temperatura del molde) son claramente activos. Utilizando el criterio más relajado (PER=0.40) habría otros tres contrastes que serían declarados como activos. El patrón de confusión en la tabla 1 indica que la 13a estadística de orden incluye el efecto de interacción BC. El hecho de que B y C han sido considerados activos, aumenta la probabilidad de que la interacción BC también lo sea.

Tabla 1. Ejemplo de cubiertas para guantera (Grove and Davis, 1992)

Número de orden	Contrastes en valor absoluto	Valores EOS para m = 15	Factores y patrón de confusión
1	0.038	0.079	Apartura de boquilla (D)
2	0.088 †	0.158	AH+BD+CE+FG
3	0.088	0.239	AD+BH+CF+EG
4	0.263 †	0.322	AB+CG+DH+EF
5	0.338 †	0.407	Recorrido de inyección (A)
6	0.338 †	0.496	2^a velocidad de inyección (E)
7	0.388	0.589	1^a velocidad de inyección (F)
8	0.438	0.688	Punto de cambio (G)
9	0.438 †	0.794	AE+BF+CH+DG
10	0.563 †	0.910	AF+BE+CD+GH
11	0.988 †	1.040	Fuerza de cierre del molde (H)
12	1.113	1.191	AC+BG+DF+EH
13	1.163 †	1.376	AG+BC+DE+FH
14	2.438 †	1.625	Temperatura del molde (C)
15	2.963 †	2.052	Temperatura del fundido (B)

† Estos valores corresponden a contrastes de signo negativo.

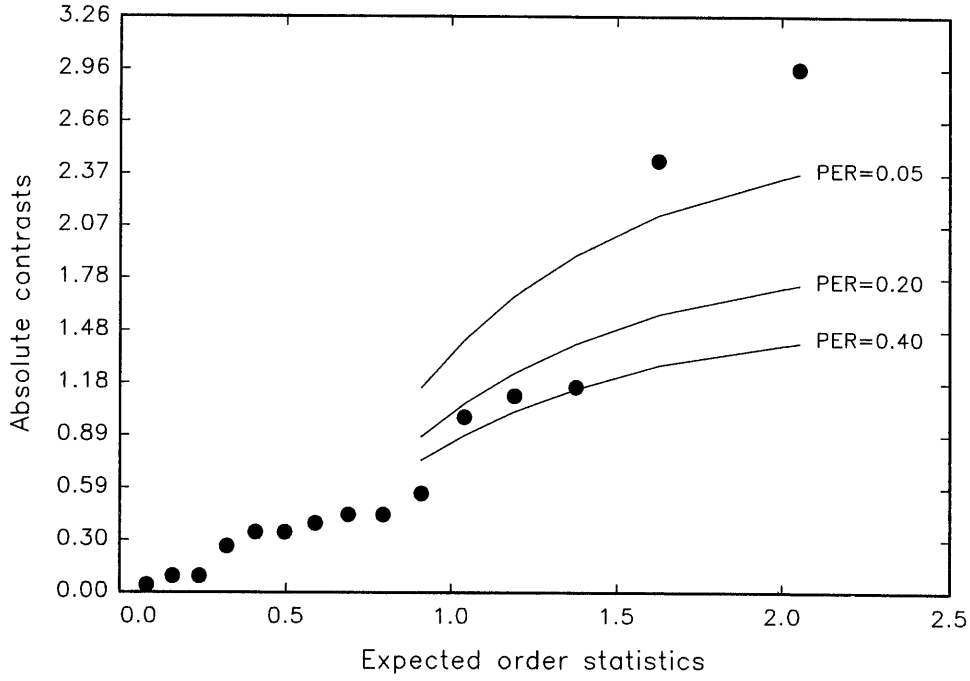


Figura 1. Gráfica semi-normal con líneas "**HP**" de significancias para el ejemplo de cubiertas para guantera.

REFERENCIAS

Daniel, C. (1959) Use of Half Normal Plots in Interpreting Factorial Two-Level Experiments. *Technometrics*, **1**, 311--341.

Grove, D. M. and Davis, T. P. (1992) *Engineering, Quality and Experimental Design*. Essex, UK: Longman.

Olguín, J. (1994). *The Analysis of Unreplicated Factorial Experiments*. PhD Thesis, University of London, University College London, August, 1994.

Zahn, D. A. (1975). Modifications and Revised Critical Values for the Half-Normal Plot, *Technometrics*, **17**, 184--200.

Bondad de Ajuste para la Distribución Levy

F. O'REILLY y R. RUEDA

IIMAS, UNAM, México

1. INTRODUCCIÓN

Sea $(X_1, X_2, ..., X_n)$ una muestra aleatoria de $F(x)$ de la cual sólo se supone que es absolutamente contínua. Se desea probar

$$H_0 = F(x) = F(x;\theta) \text{ con } \theta \in \Theta .$$

Para probar H_0, se construye una funcional del llamado proceso empírico,

$$\xi_n(x) = \sqrt{n}\left\{ F_n(x) - F\left(x, \hat{\theta}_n\right)\right\}_{x \in R}. \tag{1}$$

Bajo condiciones usuales de regularidad para la familia descrita por H_0, este proceso converge a un proceso Gaussiano de media cero y cierta función de covarianza, que se relaciona con la de un Browniano "atado".

Las funcionales cuadráticas del proceso empírico son de la forma

$$y_n^2 = \int_R \xi_n^2(x)\omega(x)dx , \tag{2}$$

con $\omega(x)$ alguna función de peso; y para la obtención de su distribución asintótica, se utiliza la convergencia del proceso empírico al proceso Gaussiano ya mencionado y el hecho de que la funcional cuadrática es una función contínua y por ello converge a la distribución que tiene la misma funcional pero aplicada al proceso límite.

En lugar de estudiar los procesos empíricos con "tiempo" dado por $x \in \mathbf{R}$, es útil "reetiquetar el tiempo" a través de $t = F(x, \theta)$; pero siendo θ desconocido, se utiliza el estimador $\hat{t} = F\left(x, \hat{\theta}_n\right)$, que al aumentar n converge a t.

Con ello, $\xi_n(x)$ tiene un proceso asociado que toma exactamente el mismo valor; este proceso es

$$\xi_n^*(\hat{t}) = \sqrt{n}\left\{ \frac{\#F\left(x_i;\hat{\theta}_n\right) \le \hat{t}}{n} - \hat{t} \right\}_{\hat{t} \in (0,1)} \tag{3}$$

y la funcional

$$y_n^2 = \int_0^1 \xi_n^2\left(F^{-1}\left(\hat{t},\hat{\theta}_n\right)\right)\omega\left(F^{-1}\left(\hat{t},\hat{\theta}_n\right)\right)dF^{-1}\left(\hat{t},\hat{\theta}_n\right), \tag{4}$$

que para $\omega(x) = f\left(x,\hat{\theta}_n\right)$, resulta en

$$W_n^2 = \int_0^1 \xi_n^{*2}(t)dt, \tag{5}$$

y para $\omega(x) = f\left(x;\hat{\theta}_n\right)\left(F\left(x;\hat{\theta}_n\right)\left(1-F\left(x;\hat{\theta}_n\right)\right)\right)^{-1}$, resulta ser

$$A_n^2 = \int_0^1 \frac{\xi_n^{*2}(t)dt}{t(1-t)}. \tag{6}$$

La teoría asintótica para funcionales cuadráticas, se estudia para el nuevo proceso para el cual se tiene que

$$\xi_n^*\left(\hat{t}\right) \Rightarrow \xi^*(t), \tag{7}$$

donde $\left\{\xi^*(t)\right\}$ es un proceso Gaussiano de media cero y función de covarianza $\rho(s,t) = E(\xi^*(s),\xi^*(t))$; entonces

$$W_n^2 = \int_0^1 \xi_n^{*2}(t)dt \overset{D}{\approx} \sum_{j=1}^{\infty} \lambda_j z_j^2, \tag{8}$$

con $\left\{\lambda_1,\lambda_2,...\right\}$, los eigenvalores de la ecuación

$$\lambda h(t) = \int_0^1 h(s)\rho(s,t)dt, \tag{9}$$

y z_1, z_2, ..., z_n variables aleatorias independientes normales estándar.

Si lo que se utiliza es a la A_n^2 de Anderson-Darling, algo similar ocurre.

La identificación de la función de covarianza del límite del proceso empírico aparece en Durbin (1973); usando su notación (para θ escalar, el caso vectorial es similar), sea $I(\theta)$ la información de Fisher por unidad muestral y sea $g(s,\theta) = \frac{\partial}{\partial\theta}F(x,\theta)$, con $s = F(x,\theta)$ y $g(t,\theta)$, con $F(y,\theta)$; entonces,

$$\rho(s,t) = s \wedge t - st - g(s,\theta)I^{-1}(\theta)g(t,\theta). \tag{10}$$

2. DISTRIBUCIÓN LEVY

La función de densidad Levy con una escala σ está dada por

$$Le(x,\sigma) = f(x;\sigma) = \sqrt{\frac{\sigma}{2\pi x^3}}\exp\left\{-\frac{\sigma}{2x}\right\}, \quad x > 0, \tag{11}$$

y la función de distribución por

$$F(x;\sigma) = 2\left[1-\Phi\left(\sqrt{\frac{\sigma}{x}}\right)\right], \tag{12}$$

donde Φ denota a la función de distribución de una normal estándar.

De acuerdo a (10), puede mostrarse que

$$\rho(s,t) = s \wedge t - st - 2\phi\left(\Phi^{-1}(1-s/2)\right)\Phi^{-1}(1-s/2)\phi\left(\Phi^{-1}(1-t/2)\right)\Phi^{-1}(1-t/2), \qquad (13)$$

con $s = F(x;\sigma)$ y $t = F(y;\sigma)$.

3. RELACIÓN CON LA DISTRIBUCIÓN GAMMA

Se sabe que X ~ Le(x; σ) si y sólo si $Y = \frac{1}{X} \sim \frac{1}{\sigma}U$, con U ~ $X^2_{(1)}$.

La densidad (ya parametrizada) para Y es

$$f_Y(y;\sigma) = \frac{\sigma^{1/2}}{\Gamma(1/2)\sqrt{2}} y^{-1/2} e^{\frac{-\sigma y}{2}} \qquad (14)$$

Una muestra (X_1, X_2, ..., X_n) de la Levy con σ desconocido, es equivalente a una muestra (Y_1, Y_2, ..., Y_n) de la gamma con parámetros $\alpha = \frac{1}{2}$ y $\beta(= 2/\sigma)$ desconocido, haciendo $Y_i = 1/X_i$, para cada i.

Denótese a las distribuciones empíricas de las muestras (X_1, X_2, ..., X_n) y (Y_1, Y_2, ..., Y_n) por $F_n^X(x)$ y $F_n^Y(y)$ y a las distribuciones por $F_X(x;\sigma)$ y $F_Y(y;\sigma)$ respectivamente, entonces

$$F_n^X(x) = 1 - F_n^Y(1/x) \quad y \quad F_X(x;\sigma) = 1 - F_Y(1/x;\sigma), \qquad (15)$$

por lo que si $\xi_n(x)$ y $v_n(y)$ denotan a los correspondientes procesos empíricos se sigue que

$$\xi_n(x) = -v_n(1/x). \qquad (16)$$

De lo anterior, si se utiliza una funcional cuadrática, al hacer el cambio de variable y = 1/x se tendrá

$$\int_0^\infty \xi_n^2(x)\omega(x)dx = \int_0^\infty v_n^2(y)\omega(1/y)\frac{1}{y^2}dy. \qquad (17)$$

En el caso $\omega(x) = f_X(x;\hat{\sigma}_n)$, se tiene

$$f_X(1/y;\hat{\sigma}_n)\frac{1}{y^2} = f_Y(y;\hat{\sigma}_n) = \omega_Y(y),$$

ya que

$$\hat{\sigma}_n = \frac{n}{\sum_{i=1}^n \frac{1}{x_1}} = \frac{n}{\sum_{i=1}^n y_i}.$$

En el caso de la A_n^2, también es cierto que

$$\omega(1/y)\frac{1}{y^2} = \omega_Y(y).$$

Así, para las estadísticas Cramér Von-Mises y Anderson-Darling, se cumple que

$$\int_0^\infty \xi_n^2(x)\omega(x)dx = \int_0^\infty v_n^2(y)\omega_Y(y)dy. \tag{18}$$

La función de covarianza en (13) y la reportada en Lockhart y Stephens (1983) son iguales debido a (16). Algunos cuantiles de la distribución asintótica para A_n^2 y W_n^2, tomados de Lockhart y Stephens (1983) se muestran en la siguiente tabla.

	0.25	0.10	0.05	0.025	0.01	0.005
W_n^2	0.132	0.205	0.265	0.328	0.412	0.484
A_n^2	0.803	1.19	1.50	1.82	2.27	2.61

4. RELACIÓN CON LA DISTRIBUCIÓN GAUSSIANA INVERSA

Sea Z una variable Gaussiana Inversa con parámetros (μ, λ). La densidad de Z en la parametrización usual es

$$f_Z(z;\mu,\lambda) = \sqrt{\frac{\lambda}{2\pi z^3}}\exp\left\{-\frac{\lambda(z-\mu)^2}{2\mu^2 z}\right\} \qquad \mu>0, \lambda>0.$$

Si se considera la parametrización $\theta = \lambda/\mu$ y $\sigma = \lambda$, entonces

$$f_Z(z;\theta,\sigma) = \sqrt{\frac{\sigma}{2\pi z^3}}e^{\theta}\exp\left\{-\frac{\theta^2 z}{2\sigma}-\frac{\sigma}{2z}\right\}. \tag{19}$$

Si θ→0, la densidad converge a la Levy con parámetro σ,

$$Le(z;\sigma) = \sqrt{\frac{\sigma}{2\pi z^3}}\exp\left\{-\frac{\sigma}{2z}\right\}.$$

La distribución $Le(z; \sigma)$ es el límite en un punto frontera del espacio paramétrico original $(0, \infty) \times (0, \infty)$; pero si se considera la extensión al espacio $[0,1) \times (0,1)$, entonces la distribucion Levy es un caso particular de la Gaussiana Inversa.

Debido a lo anterior, la distribución (19) es miembro de la familia exponencial no regular, ya que el conjunto de valores del parámetro en los que la densidad es propia, no forma un abierto. En este punto, $\theta = 0$, hay un cambio cualitativo importante, pues la distribución deja de tener momentos de orden mayores o iguales a un medio, y se convierte en una ley estable con cola pesada ($\propto z^{-3/2}$) en lugar de una cola con caída exponencial.

En O'Reilly y Rueda (1992), se identificó a la distribución asintótica del proceso empírico de la Gaussiana Inversa y se demostró como la correspondiente función de covarianza dependía sólo de θ, y al hacer θ tender a cero, apareció la función de covarianza dada en

(13), e identificada en la literatura como la asociada a una Gamma (α, β) con α conocida e igual a 1/2; en otras palabras, la Ji-cuadrada de un grado de libertad y con escala desconocida, que es el recíproco de la Levy con parámetro σ; y que coincide por el resultado (16) de la sección 3, como ya se había mencionado.

Con el objeto de explorar el parecido entre una Gaussiana Inversa (θ, σ) con θ muy chico y una Levy con parámetro σ, se llevaron a cabo unos estudios de potencia simulando 1000 muestras de la distribución Gaussiana Inversa para valores de n = 10(10)100 y viendo que proporción de las veces eran detectadas al 5%, como no provenientes de la Levy. Los resultados indicaron que resulta fácil tomar a una Gaussiana Inversa por una Levy (si $\theta < 2^{-6}$) aún para muestras relativamente grandes (n = 100), sabiendo por otro lado las repercusiones que podría tener tal equivocación si se trata con eventos relacionados con valores extremos.

REFERENCIAS

Durbin, J. (1973). Distribution theory for tests based on the sample distribution function. *Reg. Conf. Ser. Appl. Math.* 9. Philadelphia: SIAM.

Lockhart, R.A. y Stephens, M.A. (1983). *Goodness-of-fit statistics with estimated shape parameters.* Tech. Report. Department of Mathematics and Statistics, Simon Fraser University.

O'Reilly, F.J. y Rueda, R. (1992). Goodness of fit for the inverse Gaussian distribution. *Can. J. Statist.* **20**, 387-397.

Análisis de Regresión de Gini

BLANCA R. PEREZ S.,
UAM, Iztapalapa, México

SERGIO DE LOS COBOS S.
UAM, Iztapalapa, México

y

MIGUEL A. GUTIERREZ A.
UAM, Azcapotzalco, México

1. INTRODUCCIÓN

El método de los mínimos cuadrados se utiliza comúnmente para estimar el valor esperado ($E(Y|X=x)$).

El método se sustenta en los siguientes supuestos:

1. Existe una relación lineal entre la variable aleatoria independiente (o explicativa) X y el valor esperado de la variable dependiente, $E(Y|X=x)$.
2. Los errores son independientes e idénticamente distribuidos y no correlacionados con la variable independiente.
3. Adicionalmente se supone que los errores se distribuyen como una normal con media 0 y varianza σ^2.

Dado el modelo de regresión lineal simple:

$$E(Y \mid X=x) = \alpha + \beta x$$

el estimador de mínimos cuadrados de β es

$$\hat{\beta} = \frac{\sum (y_i - \bar{y})(x_i - \bar{x})}{\sum (x_i - \bar{x})^2}$$

$\hat{\beta}$, así calculada, es sensible a valores extremos, lo que se puede observar al considerar las pendientes entre cada observación (x_i, y_i) y el vector de promedios $(\bar{x}, \bar{y})$.

$$m_i = \frac{y_i - \bar{y}}{x_i - \bar{x}},$$

porque $\hat{\beta}$ se puede escribir como promedio ponderado

$$\hat{\beta} = \sum m_i w_i = \sum \frac{y_i - \bar{y}(x_i - \bar{x})^2}{x_i - \bar{x} \sum (x_i - \bar{x})^2}$$

donde $w_i = \dfrac{(x_i - \bar{x})^2}{\sum (x_i - \bar{x})^2}$ $\quad \left(w_i \geq 0 \text{ y } \sum w_i = 1\right)$.

Obsérvese que las pendientes de los puntos (x_i, y_i) donde $|x_i - \bar{x},|$ es "grande" pesan mas en el promedio que las pendientes de los otros puntos.

Hay dos posibilidades para obtener un estimador menos sensible a valores extremos:

1. Utilizar pesos menos desbalanceados como

$$w_i = \frac{1}{n} \quad \text{o} \quad w_i = \frac{|x_i - \bar{x}|}{\sum |x_j - \bar{x}|}$$

2. Utilizar otra medida de dispersión para los errores de observación.

Las dos posibilidades fueron exploradas por Olkin y Yitzhaki (1992), en base a sus resultados, se propone un método alternativo de obtener los estimadores de la función de regresión, esto se hace en las siguientes secciones.

2. DESVIACIÓN MEDIA O DE GINI

La diferencia media de una variable aleatoria (va) X con respecto a δ es $E(|X - \delta|)$.

La diferencia media de X alcanza su mínimo valor cuando $\delta = M$ (la mediana de los datos) y se conoce como desviación media de Gini, $E(|X - M|)$.

Por otro lado, la diferencia media de Gini de las variables aleatorias, X_1 y X_2, es el valor esperado $E(|X_1 - X_2|)$

Si $X \sim N(0, \sigma^2)$, entonces $E(|X - M|) = \sqrt{2\sigma} / \sqrt{\pi}$.

Y si $X_1, X_2 \sim N(0, \sigma^2)$ son variables independientes, entonces $E(|X_1 - X_2|) = 2\sigma / \sqrt{\pi}$.

Por lo que en el caso que las variables sean normales, la diferencia media, la desviación media y la varianza, son medidas de dispersión equivalentes.

3. REGRESIÓN DE GINI

Dos son las rectas que se pueden estimar utilizando la diferencia media de Gini como medida de dispersión:

1. La recta que minimiza la expresión

$$f(\alpha, \beta) = \sum |\varepsilon_i| = \sum |y_i - \alpha - \beta x_i|.$$

2. La recta que minimiza la expresión

$$g(\beta) = \sum_{i>j} |\varepsilon_i - \varepsilon_j| = \sum |y_i - y_j - \beta(x_i - x_j)|.$$

En el primer caso $f(\alpha, \beta)$, es bivaluada; y en el segundo caso $g(\beta)$ es univaluada.

Proposición 1 $f(\alpha, \beta)$ y $g(\beta)$ son funciones convexas.

Demostración: Considere los vectores $f(\alpha_1, \beta_1)$ y $f(\alpha_2, \beta_2)$ el número real λ, $(0 \le \lambda \le 1)$.

$$f(\lambda(\alpha_1, \beta_1) + (1-\lambda)(\alpha_2, \beta_2)) = \sum |y_i - \lambda\alpha_1 - (1-\lambda)\alpha_2 - (\lambda\beta_1 - (1-\lambda)\beta_2)x_i|$$

$$\le \lambda \sum |y_i - \alpha_1 - \beta_1 x_i| + (1-\lambda) \sum |y_i - \alpha_2 - \beta_2 x_i| = \lambda f(\alpha_1, \beta_1) + (1-\lambda) f(\alpha_2, \beta_2)$$

por lo que se demuestra que $f(\alpha, \beta)$ es convexa.

Ahora considere los números β_1, β_2 y λ, $(0 \le \lambda \le 1)$, por la misma razón

$$g(\lambda(\beta_1) + (1-\lambda)(\beta_2)) \le \lambda g(\beta_1) + (1-\lambda) g(\beta_2).$$

Por ser $g(\beta)$ y $f(\alpha, \beta)$ funciones convexas, entonces alcanzan su valor mínimo en uno de los vértices de su gráfica.

Proposición 2. El problema

$$\boxed{\text{minimizar } g(\beta) = \sum|\varepsilon_i - \varepsilon_j| = \sum|y_i - y_j - \beta(x_i - x_j)|}$$

es equivalente al problema de programación lineal:

$$\boxed{\text{Minimizar } H(\beta, \gamma) = \gamma \text{ sujeto a las } 2^{n(n-1)/2} \text{ restricciones } \sum_{i<j} \pm(\varepsilon_i - \varepsilon_j) \le \gamma,}$$

las condiciones son resultado de tener todas las posibles combinaciones de los signos (+) y (−) en los sumandos.

Proposición 3. El problema

$$\boxed{\text{minimizar } f(\alpha, \beta) = \sum|\varepsilon_i| = \sum|y_i - \alpha - \beta x_i|}$$

es equivalente al problema de programación lineal:

$$\boxed{\text{Minimizar } H(\alpha, \beta, \gamma) = \gamma \text{ sujeto a las } 2^n \text{ restricciones } \sum_{i<j} \pm\varepsilon_i \le \gamma,}$$

El número de restricciones dificulta utilizar el método simplex.

Proposición 4. Los vértices de $g(\beta)$ coinciden con la solución de las ecuaciones:

$$\varepsilon_i - \varepsilon_j = 0$$

para todo $i \neq j$. Esto implica que

$$\varepsilon_i - \varepsilon_j = y_i - y_j - \beta(x_i - x_j) = 0$$

y que

$$\beta_{ij} = \frac{y_i - y_j}{x_i - x_j}$$

los vértices de la gráfica de la función coinciden con la pendiente entre dos puntos muestrales.

Proposición 5. Los vértices de $f(\alpha, \beta)$ coincide con la solución de los sistemas de ecuaciones:

$$\begin{pmatrix} 1 & x_i \\ 1 & x_j \end{pmatrix} \begin{pmatrix} \alpha \\ \beta \end{pmatrix} = \begin{pmatrix} y_i \\ y_j \end{pmatrix}$$

La solución del sistema es:

$$\beta_{ij} = \frac{y_i - y_j}{x_i - x_j}$$

y

$$\hat{\alpha}_{ij} = y_i - \hat{\beta} x_i = y_j - \hat{\beta} x_j.$$

La solución coincide con la recta que pasa por los puntos (x_i, y_i) y (x_j, y_j).

El vértice donde se alcanza el mínimo valor (de g o de f) se puede encontrar por aproximaciones sucesivas, siempre buscando una ruta de descenso de acuerdo al siguiente algoritmo:

- Seleccionar un sistema de ecuaciones que definen un vértice. De este vértice parten $2m$ vectores direccionales (m = 1,2), determinados por cada una de las ecuaciones del sistema.
- Escoger, entre los $2m$ vectores el vector de dirección de máximo descenso.
- Escoger el vértice que se encuentra en la dirección del vector de máximo descenso donde la función tiene el mínimo valor.
- Repetir el proceso a partir de este vértice, hasta encontrar un vértice donde los vectores direccionales no desciendan mas.

El algoritmo tiende rápidamente a la solución.

En cualquiera de los casos, los vértices que minimizan la función objetivo no necesariamente son únicos.

4. EL ESTIMADOR DE β COMO PROMEDIOS PONDERADOS

Si la solución es única e igual a:

$$\hat{\beta} = \frac{y_{i_0} - y_{j_0}}{x_{i_0} - x_{j_0}}$$

se puede ver como un promedio ponderado

$$\sum m_{ij} w_{ij}$$

con $w_{i_0,j_0} = 1$ y $w_{i,j} = 0, \quad \forall\ i,j \neq i_0, j_0$

Si la solución no es única, (son dos o mas vértices los que minimizan la función objetivo, V es el conjunto de estos vértices), $\hat{\beta}$ se toma como el promedio ponderado

$$\hat{\beta} = \sum m_{ij} w_{ij}$$

con

$$w_{ij} = \begin{cases} \sum \frac{|x_i - x_j|}{|x_k - x_l|} & \text{si } (x_i, x_j) \text{ y } (x_k, x_l) \text{ se relacionan con los elementos de } V \\ 0 & \text{en otro caso} \end{cases}$$

De este modo, sólo cuentan las pendientes que minimizan la diferencia media de Gini.

5. REGRESIÓN MÚLTIPLE DE GINI

La diferencia media de Gini puede utilizarse para ajustar un modelo de regresión lineal múltiple. La solución se encuentra en uno de los vértices de la función, si un solo vértice minimiza la función; o en la combinación convexa de dos o mas vértices de la función

$$\sum |\varepsilon_i - \varepsilon_j| = \sum |y_i - y_j - \beta_1 (x_{i1} - x_{j1}) - \beta_2 (x_{i2} - x_{j2}) - \ldots - \beta_m (x_{im} - x_{jm})|$$

o de la función

$$\sum|\varepsilon_i| = \sum|y_i - \alpha - \beta_1 x_{i1} - \beta_2 x_{i2} - \ldots - \beta_m x_{im}|$$

de manera semejante a como se especificó en el caso de regresión lineal simple.

6. PRUEBAS A LOS SUPUESTOS DEL MÉTODO DE LOS MÍNIMOS CUADRADOS

La linealidad de la variable dependiente con respecto a las variables independientes se prueba considerando la desigualdad

$$\sum|y_i - y_j| - |\hat{\beta}| \sum|x_i - x_j| \leq \sum|\varepsilon_i - \varepsilon_j|$$

lo cual implica que en un ajuste perfecto se cumpla la ecuación:

$$\frac{|\hat{\beta}| \sum|x_i - x_j|}{\sum|y_i - y_j|} = 1$$

El supuesto de normalidad se prueba comparando el estimador de varianza de los métodos de mínimos cuadrados con el estimador de varianza de la regresión de Gini.

El supuesto de independencia entre ε y X se prueba considerando la "cercanía" entre (ε, X).

REFERENCIAS

Olkin, I. and Yitzhaki, S. (1992) Gini Regression Analysis, *International Statistical Review*, **60**, 2, pp 185-192.

Uso de Estimación Tipo Ridge para Reducir Sesgo en Regresión Logística

GUSTAVO RAMÍREZ VALVERDE y JANET C. RICE

Colegio de Postgraduados, México *Tulane University, USA.*

1.INTRODUCCIÓN

El Modelo de Regresión Logística (MRL) es un tipo de regresión para variables respuesta binaria, este modelo puede escribirse como: $\pi_i = (1+e^{\underline{x}_i\underline{\beta}})^{-1}$, donde $\pi_i = p_i = p(\, y_i=1|\, \underline{x}_i\,)$, $\underline{\beta}=(\beta_0\ \beta_1\ ...\ \beta_p)^T$ es un vector de parámetros desconocidos y $\underline{x}_i = (1\ x_1\ x_2\ ...\ x_p)$ es un vector con los valores de p variables explicativas asociadas a la observación i. El método de estimación mas usado en el MRL es el de Máxima Verosimilitud, sin embargo, el estimador de máxima verosimilitud (EMV) no existe cuando la matriz de información estimada es singular (Lesaffre y Marx, 1993). Existen dos posibles fuentes de singularidad en la matriz de información estimada, la primera de ellas es cuando existen dependencias lineales entre las variables explicativas, esto es, cuando la matriz $X=(\underline{x}_1\ \underline{x}_2\ ...\ \underline{x}_n\ )^T$ es singular (esta situación es conocida como exacta colinealidad entre las variables explicativas. La segunda ocurre cuando se presenta la condición de completa (quasi-completa) separación (Albert y Anderson, 1989; Santer y Duffy, 1986), esto ocurre si existe un hiperplano en el espacio R^p tal que todos los valores $\underline{x}_i$ correspondientes a las observaciones con valores $y_i = 1$ están en un lado del hiperplano, y aquellas con valores $y_i = 0$ en el otro (quasi-completa separación se presenta cuando además se tienen al menos una observación de cada tipo en el hiperplano). Esta última condición se va a denotar como exacta MV-colinealidad. La matriz de información tiende a estar mal condicionada a medida que se acerca a alguna de las situaciones arriba descritas. Numerosos trabajos se han presentado describiendo los efectos del mal condicionamiento de la matriz de información estimada en la precisión de el EMV(Marx y Smith, 1990, Shaeffer, Roi y Wolfe, 1984; Shaeffer, 1979, 1983 y 1986; Eissfield y Sereika, 1991) o en el poder de la prueba de wold (Hauck y Donner, 1977). En este trabajo (sección 2) se muestra el efecto negativo que tiene cierto tipo de mal condicionamiento (MV-colinealidad) en el sesgo de el EMV, en la sección 3 se realizó un estudio de simulación para comparar algunos estimadores alternativos con el EMV.

2. EFECTO DE LA MV-COLINEALIDAD EN EL SESGO DEL EMV

Sea Ω el espacio de posibles realizaciones de $\underline{Y}$, una muestra aleatoria de tamaño n obtenida con la matriz diseño X, donde la matriz de información estimada está mal condicionada y el EMV existe. Defínase a $c_{i1} \in \Omega$ como la realización simétrica de $c_{i2} \in \Omega$ (i = 1, 2, ..., N) si todas las observaciones y_j (j = 1, 2, ..., n) $\in c_{i1}$ con $y_j = 1$ corresponden a observaciones con $y_j = 0$ de c_{i2} y todas las observaciones y_j (j = 1, 2, ..., n) $\in c_{i1}$ con $y_j = 0$ corresponden a observaciones con $y_j = 1$ de c_{i2} . Dos realizaciones simétricas tienen el mismo valor absoluto de los parámetros estimados (el MVE), pero presentan diferente signo, y por lo tanto tienen el mismo número condición de sus respectivas matrices de información estimadas. Sea c_i (i = 1, 2,..., N) el conjunto formado por las realizaciones simétricas c_{i1} y c_{i2} con N el número de pares de realizaciones simétricas en Ω, entonces, 2N es la cardinalidad de Ω.

Sin perder generalidad, se supondrá una sola variable explicativa y que c_{i1} (i=1,2,...,N) corresponde a la realización que tiene el mismo signo que el verdadero parámetro y sea $\hat{\beta}_i$ el

EMV de b obtenido con la realización c_{i1}, entonces, $p(c_{i1}|c_i) >>> p(c_{i2}|c_i)\ \forall\ i = 1,2, ..., N$. El sesgo esperado dado que la matriz de información esta mal condicionado y el EMV existe es:

$$E\left(\hat{\beta}_i - \beta | y \in \Omega\right) = E\left[E\left(\hat{\beta}_i - \beta | y \in c_1 | c_i \in \Omega\right)\right]$$

$$= E\left\{\left(\hat{\beta}_i - \beta\right)\left[p\left(\underline{y} = c_{i1} | \underline{y} \in c_i\right) - p\left(\underline{y} = c_{i2} | \underline{y} \in c_i\right)\right] | c_i \in \Omega\right\} - E\left[2\beta p\left(\underline{y} = c_{i2} | \underline{y} \in c_i\right)\right\} | c_i \in \Omega\right]$$

Cuando existe MV-colinealidad se tiene que $p(c_{i1}|c_i) >>> p(c_{i2}|c_i)\ \forall\ i = 1,2, ..., N$ y $\hat{\beta}_i$ tiende a infinito, entonces, $p(\underline{y} = c_{i2} | \underline{y} \in c_i) \to 0$, $p(\underline{y} = c_{i1} | \underline{y} \in c_i) - p(\underline{y} = c_{i2} | \underline{y} \in c_i) \to 1$ y $\hat{\beta}_i - \beta \to \infty$, entonces, El sesgo esperado condicionado a la presencia de MV-colinealidad y la existencia del EMV tiende a infinito a medida que $\hat{\beta}_i$ se aproxima a infinito, situación que se presenta con la MV-colinealidad.

3. SIMULACIÓN

Un estudio de simulación fue realizado con el fin de estimar la magnitud del efecto de la MV-colinealidad en el sesgo del EMV y comparar comportamiento de algunos estimadores alternativos. En este artículo se consideraron 4 estimadores, el primero es el EMV, el segundo construído para disminuir el sesgo, este estimador será llamado TAYLOR (Anderson y Richardson, 1979; Schaeffer, 1983; Copas, 1988) corresponde al valor obtenido con $\hat{\beta}_T = \hat{\beta} - b$, donde $\hat{\beta}$ es el EMV y b es el sesgo aproximado y es calculado por: $b = -\frac{I}{2}\hat{i}^{-1} \sum_{i=1}^{n} \underline{x}_i \left(\frac{1}{2} - \hat{\pi}_i\right)\hat{\pi}_i\left(1 - \hat{\pi}_i\right)\underline{x}_i^T \hat{i}^{-1}\ \underline{x}_i$, donde $\hat{i}^{-1}$ es el inverso de la matriz de información evaluada a $\underline{\beta} = \hat{\beta}$ el EMV de $\underline{\beta}$ y $\hat{\pi}_i$ es el EMV de π_i.

Los otros dos estimadores son tipo Ridge (Schaeffer, 1979; Shaeffer, Roi, y Wolfe, 1984) . El estimador Ridge es calculado con la ecuación: $\hat{\beta}_R = \left(X^T \hat{V} X + kI\right)^{-1} X^T \hat{V} X \underline{\hat{\beta}} = \left(\hat{i} + kI\right)^{-1} \hat{i}\hat{\beta}$, donde I es una matriz idéntica de tamaño p+1 y $k > 0$ es una constante llamada parámetro ridge. Shaeffer (1979, 1986) propuso diferentes valores para k, en este artículo se consideraron dos de los estimadores propuestos por Schaeffer, el primer valor del parámetro ridge es $k_1 = (p+1)/\hat{\beta}^T\hat{\beta}$ dando el primer estimador tipo Ridge (RTE1), el segundo estimador (RTE2) más conservador que el anterior (en relación al ajuste que hace sobre el EMV) se construye con $k_2 = 1/\hat{\beta}^T\hat{\beta}$.

Los casos simulados tuvieron 2 variables explicativas y un tamaño de muestra de n = 25. Los factores bajo estudio fueron: a)Correlación entre variables explicativas. Dos diferentes correlaciones fueron usadas (r = 0.95662 y r = 0.99572), dando un número condición para la matriz X^TX de 45.106 y 466 respectivamente, .b) Dirección de la colinealidad. Dos ángulos (0° y 90°)entre el parámetro y el eigenvector asociado con el menor eigenvalor de la matriz X^TX (llamado dirección de la colinealidad) y c) Tamaño de la norma del vector de parámetros. Dos diferentes tamaños fueron usados $|\underline{\beta}| = 1$ o 2. Las situaciones estudiadas se resumen en todas las combinaciones entre los factores estudiados. Una vez que la matriz diseño y el parámetro $\underline{b}$ es determinado, los valores de π_i fueron calculados de acuerdo a la situación simulada.

La condición de MV-colinealidad depende de la variable respuesta, por lo que, para la determinación de su presencia se calculó el número condición de la matriz de información estimada (k_I). La simulación fue llevada a cabo hasta que se obtuvieron 1000 ensayos con k_I mayor que 100 (situaciones con MV-colinealidad) si el número condición de la matriz X^TX (k_X) era 45.106 (r = .95662) y para las situaciones con k_X de 466 (r = .99572) hasta que se obtuvieron 1000 ensayos con k_I > 1000 (situaciones con MV-colinealidad). El número de ensayos donde el EMV no existe no fueron contabilizados, entonces los resultados están condicionados a la existencia de el EMV.

Los resultados de la simulación para las situaciones con r = 0.99572, $|\beta| = 2$ y cero grados de colinealidad son presentados en la figura 1, y se puede observar que en los casos con MV-colinealidad el EMV mostró sesgos medios y error cuadráticos medios (ECM) mucho mayores que los casos sin MV-colinealidad. El sesgo medio en las situaciones sin MV-colinealidad oscila entre -0.18 y 0.173 y en los casos con MV-colinealidad (k_I >1000) que van de -12.837 a 13.152, el MSE en los casos con k_I < 1000 oscilan entre 0.22 y 27.17, y en los casos con k_I >1000 van de 0.82 a 450.81. Cuando k_I >1000, el estimador Taylor y los dos estimadores Ridge tuvieron menor sesgo medio y MSE que los que tuvo el EMV, siendo el estimador TAYLOR el mejor en ambos sentidos. Cuando k_I < 1000, el estimador Taylor y los dos estimadores Ridge sobrecorrigieron el sesgo medio del EMV, siendo malas opciones para reducir el sesgo, sin embargo, todos ellos tuvieron menor ECM que el EMV. Todas las demás situaciones con cero grados de colinealidad presentan en general las mismas tendencias arriba descritas (Esta información no se presenta, pero esta disponible con el primer autor).

La figura 2 presenta los resultados obtenidos en condiciones similares a la figura 1 con excepción de la dirección de la colinealidad, la figura 2 es con 90° . En la figura 2 se puede observar que los casos con MV-colinealidad tienen al igual que con 0° el sesgo medio y el ECM mucho mayores que las que no tienen MV-colinealidad Cuando k_I >1000, los dos estimadores Ridge tuvieron menores sesgos medios y ECM que los que tuvo el EMV, pero el estimador TAYLOR sobrecorrigio el sesgo medio del EMV, siendo este muy grande, incluso mayor que el EMV. siendo la peor opción en ambos sentidos. Cuando k_I < 1000, el estimador TAYLOR sobrecorrige el sesgo del EMV , pero tiene menor ECM., sin embargo , los dos estimadores Ridge tuvieron menor sesgo medio y ECM que el EMV, pero existe una ligera tendencia a sobrecorregir el sesgo de algunos parámetros; los estimadores Ridge tuvieron menor ECM que el estimador TAYLOR. Las demás situaciones con 90° de dirección de la colinealidad presentaron en general las mismas tendencias.

4. CONCLUSIÓN

En todas las situaciones estudiadas, el sesgo medio del EMV condicionado a la presencia de MV-colinealidad fue mucho mayor que el EMV condicionado únicamente en la existencia del EMV, con lo que se verifica que la MV-collinealidad incrementa el sesgo medio del EMV. El estimador TAYLOR es una buena opción para reducir el sesgo cuando el parámetro presenta 90° de colinealidad, pero es mala selección cuando se tiene 0° , desafortunadamente el parámetro y por consiguiente la dirección de la colinealidad son desconocidos, por lo que no es posible determinar cuando el estimador TAYLOR es apropiado. Los estimadores Ridge son una buena opción para reducir el sesgo y el ECM en presencia de MV-colinealidad, mostrando mejores resultados RTE2, sin embargo el problema para escoger el mejor k permanece irresuelto.

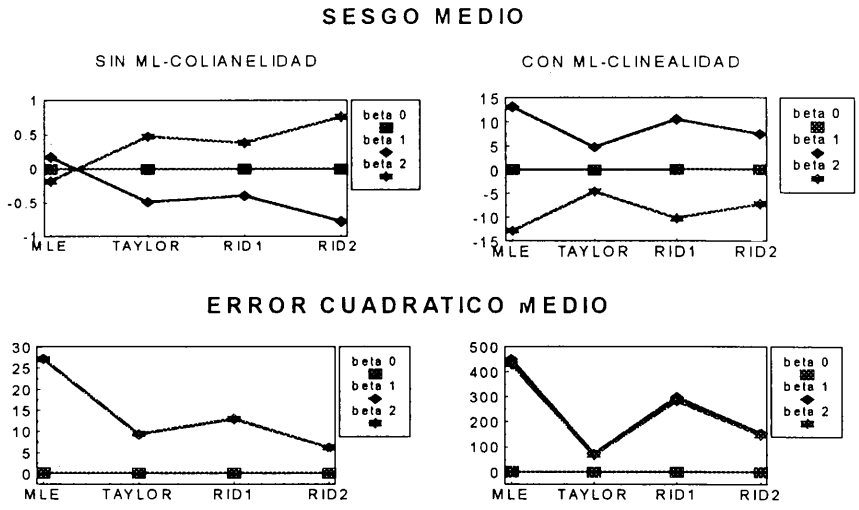


Fig. 1. Gráficos presentando los resultados en cuanto a sesgo medio y MSE de los casos con r = 0.99572 y 0 grados de colinealidad.

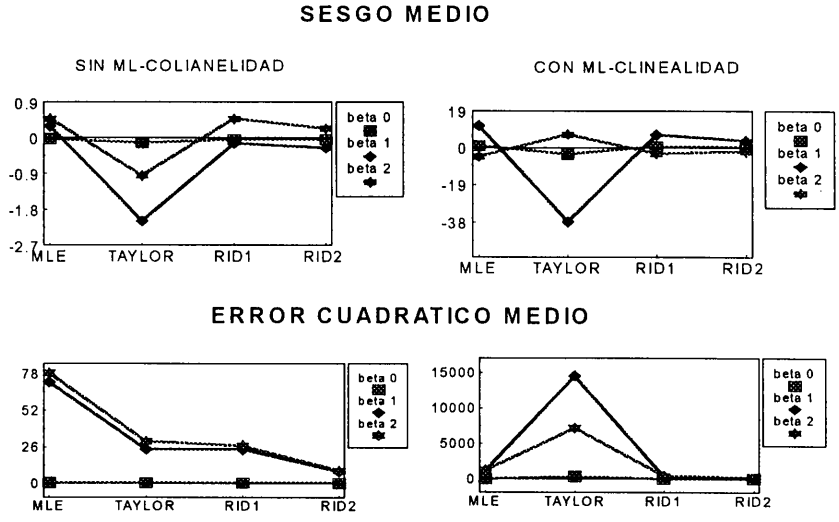


Fig 2. Gráficos presentando los resultados en cuanto a sesgo medio y MSE de los casos con r = 0.99572 y 90 grados de colinealidad.

REFERENCIAS

Albert, A. and J. A. Anderson. 1984. On the Existence of Maximum Likelihood Estimates in Logistic Regression Model. *Biometrika* **71**:1-10.

Anderson, J. A. and S. C. Richardson. 1979. Logistic Discrimination and Bias Correction in Maximum Likelihood Estimation. *Technometrics 21*:71-78.

Copas, J. B. 1988. Binary Regression Models for Contaminated Data. *Journal of the Royal Statistical Society*,Series B **50**:225-65.

Cox, D. R., D. V. Hinkley. 1974. Theoretical Statistics. London: Chapman and Hall. *Biometrics* **40**:1117-1123.

Hauck, W. W. and A. Donner. 1977. Wald's Test as Applied to Hypotheses in Logit Analysis. *Journal of the American Statistical Association*. **72**:851-3.

Lesaffre, E. and B. D. Marx. 1993. Collinearity in Generalized Linear Regression. *Communications in Statistics Theory and Methods* **22**:1933-52.

Marx, B. D. and E. P. Smith. 1990. Principal Component Estimation for Generalized Linear Regression. *Biometrika* **77**:23-31.

Santner, T. J. and D. E. Duffy. 1986. A Note on A. Albert and J.A. Anderson's Conditions for the Existence of Maximum Likelihood Estimates in Logistic Regression Models. *Biometrika* **73**:755-758.

Schaefer, R. L. 1986. Alternative Estimators in Logistic Regression when the Data are Collinear. *Journal of Statistical Computations and Simulations* **25**:75-91.

1983. Bias correction in Maximum Likelihood Logistic Regression. *Statistics in Medicine* **2**:71-78.

1979. "Multicollinearity and Logistic Regression." Doctoral Dissertation. University of Michigan.

Schaeffer, R. L., L. D. Roi and R. A. Wolfe. 1984. A Ridge Logistic Estimator. *Communications in Statistics Theory and Methods* **13**:99-113.

Weissfeld, L. A. and S. M. Sereika. 1991. A Multicollinearity Diagnostic for Generalized Linear Models. *Communications in Statistics Theory and Methods* **20**:1183-98.

Selección de Indicadores de Actividad Biológica Mediante Algoritmos Genéticos

ROGELIO RAMOS QUIROGA
CESAR GUERRA SALCEDO
GRACIELA GONZÁLEZ FARIAS
MANUEL VALENZUELA RENDÓN

ITESM, Campus Monterrey, México

1. INTRODUCCIÓN

En la industria farmacéutica, la investigación y el desarrollo de nuevos medicamentos juegan un papel crucial para la supervivencia de la compañía. Típicamente, en las etapas iniciales de desarrollo se tiene un compuesto químico con características de actividad biológica promisorias, pero para el cuál se tienen posibles niveles de potencia o de toxicidad que no son adecuados, consecuentemente, series de compuestos químicos son producidos para su evaluación; es precisamente en esta fase que las técnicas de QSAR (quantitative structural-activity relationships) proveen a los químicos con herramientas para identificar características de un compuesto asociadas con la actividad biológica. Potencialmente la química computacional es capaz de generar miles de descriptores o características de un compuesto; este hecho complica el problema de determinar cuales son los descriptores importantes.

El objetivo del presente trabajo consiste en implementar una solución, mediante la utilización de algoritmos genéticos, para el problema de caracterizar un conjunto de descriptores que determinen un alto grado de actividad biológica. Presentaremos un análisis de un conjunto de observaciones obtenidas por Selwood *et al* (1990) el cuál contiene información de 53 propiedades físicas (descriptores) de 31 compuestos químicos. En este caso las técnicas de Regresión Lineal Múltiple no pueden ser implementadas debido a la dimensionalidad del problema; además, es frecuente que en las fases iniciales de exploración, se cuente con bioensayos cuyos resultados indican una respuesta binaria en términos de su actividad, esto es, indica solamente si un compuesto es activo o no lo es.

Una solución al problema de selección propuesta por McFarland y Gans (1986), es llamada CSA (Cluster Significance Analysis) y se basa en el concepto de "concentración de parámetros" desarrollado por Magee (1983). La idea básica consiste en fijarnos en un conjunto específico de descriptores, si éstos son importantes para describir el nivel de actividad de un compuesto, entonces se esperaría que los compuestos activos formen un grupo /cluster) muy compacto. CSA mide la evidencia que respalda la importancia de los descriptores considerados. Este tipo de comportamiento de los compuestos activos ha sido descrito como "agrupamiento asimétrico" por Dunn y Wold (1978).

Otros métodos de análisis presentados en la literatura son, entre otros, el propuesto por Selvood *et al* (1990) y el de Wikel y Dow (1993). En un problema con N compuestos y K descriptores, cada compuesto puede considerarse como un punto en $\mathcal{R}^k$. Selwood y colaboradores utilizaron un mapeo no-lineal para proyectar puntos de $\mathcal{R}^k$. en el plano y de ese modo tener una herramienta visual para determinar la importancia de los K descriptores considerados. Wikel y Dow por otra parte, utilizaron un enfoque de redes neuronales para identificar descriptores relevantes a actividad biológica.

2. METODOLOGÍA

El método para calcular probabilidades de asociación usado en CSA es una parte importante en nuestra propuesta de solución. Los datos de Selwood *et al* forman una matriz de orden 31 × 53 donde cada renglón representa un compuesto químico especificado por 53 descriptores. De los 31 compuestos 15 son considerados "activos". Entonces el grupo de activos puede visualizarse como una nube de 15 puntos en un espacio de 53 dimensiones. Se obtiene una medida de compactación de los activos, calculando su distancia cuadrática promedio, esto es, si c_1, c_2, ..., c_{15} son vectores rengón en $\mathcal{R}^{53}$ entonces su distancia cuadrática promedio es:

$$MSD_a = \frac{\sum_{i>j} \|c_i - c_j\|^2}{n(n-1)/2} \qquad (1)$$

donde n = 15. Para calcular la significancia de esta medida y determinar si puede ser debida a una asociación aleatoria, se calculan los MSD's de todos los posibles grupos de 15 compuestos tomados del total de 31 y se calcula su p-valor:

$$p = p\text{-}valor \equiv \frac{\#\text{de MSD's} \leq MSD_a}{\binom{31}{15}} \qquad (2)$$

El cálculo de este p-valor implica el cálculo de poco más de 300 millones de MSD's lo cual es computacionalmente prohibitivo. McFarland y Gans propusieron estimar este p-valor tomando una muestra de M grupos del total de ${}_{31}C_{15}$ donde el tamaño de muestra M se determina considerando una distribución binomial con parámetros M y p y tomando M de modo que su intervalo de confianza pra p tenga una longitud pequeña especificada de antemano. Para evaluar un conjunto arbitrario de d descriptores se efectúa el procedimiento anterior pero ahora los compuestos serán vectores renglón en $\mathcal{R}^d$.

El método que proponemos en este trabajo, consiste en implementar un algoritmo genético que guíe la búsqueda de un conjunto de descriptores con un p-valor lo más pequeño posible. Para ello consideramos que nuestra población está formada por 2^{53} cromosomas, donde cada cromosoma es un vector renglón en $\mathcal{R}^{53}$, formado por unos y ceros. Un 1 en la j-ésima posición significa que el descriptor está siendo considerado. Inicialmente, seleccionamos una muestra de esa población y evaluamos la capacidad predictiva de cada uno de los elementos de la muestra. Esta capacidad estará medida por la significancia del agrupamiento de compuestos activos (p-valores de McFarland y Gans). Los elementos más aptos serán retenidos y a partir de ellos se obtendrá una nueva generación de combinaciones mediante selección, cruza y/o mutaciones iterando el algoritmo hasta convergencia.

Los Algoritmos Genéticos (AG's) son procedimientos de búsqueda y optimización que mimetizan el proceso adaptivo natural de las especies en sistemas naturales; están basados en la suposición de que las mismas leyes que regulan los procesos adaptivos naturales de las especies pueden ser aplicados para resolver problemas de tipo ingenieril (Goldber; 1989; Holland, 1975, Valenzuela, Guerra e Icaza 1991).

Los Algoritmos Genéticos trabajan sobre poblaciones de estructuras que presentan posibles soluciones a un problema en particular. Estas estructuras se llaman cromosomas o individuos (partiendo de la suposición de que un individuo está formado por cromosomas y que los procesos de selección natural actúan sobre cromosomas) y son cadenas formadas de algún alfabeto de baja cardinalidad (generalmente 0, 1).

A partir de una población inicial (que puede ser establecida aleatoriamente) el AG aplica operadores genéticos y evoluciona poblaciones que contengan individuos altamente capaces (esta capacidad de un individuo es la analogía a una función de aptitud que aparentemente la naturaleza aplica a los organismos, los más aptos sobreviven y los menos no).

Los operadores genéticos más utilizados son cruce, selección y mutación.

(a) Cruce. A partir de dos indivduos (padres) es posible generar uno o más individuos, donde cada uno de ellos tendrá características de sus padres (se supone que si son padres altamente competitivos, los hijos resultarán altamente competitivos).

(b) Selección. La selección asigna estocásticamente (Goldber, 1989, Holland, 1975) copias de los individuos más aptos para así formar una generación.

(c) Mutación. Con cierta probabilidad los genes de los cromosomas pueden mutar (0 por 1 y viceversa) con esto se copia el hecho de que en las especies existen mutaciones a lo largo de generaciones.

Para poder aplicar AG's a algún problema en particular, es necesario establecer una representación de los cromosomas y una función objetivo que tomará el rol de la función de evaluación de individuos. Es importante establecer también los puntos de cruce de los padres así como la probabilidad de mutación de los genes.

Se aplicó un Algoritmo Genético Simple (Goldber, 1989) al problema definido en las secciones anteriores, los parámetros del mismo fueron: un punto de cruce con probabilidad 0.89, mutación de genes con probabilidad 0.003 y cruce uniforme (Spears y De Jong, 1991).

Representación de los cromosomas: Los cromosomas para el problema aquí establecido son strings binarios (0 y/o 1) que codifican el descriptor a evaluar dentro de un compuesto, así si los compuestos tienen asociados por k descriptores, el cromosoma tendrá k bits cada uno representando la inclusión de un descriptor o la exclusión del mismo (Guerra, 1991). Esta representación permite manejar un espacio de búsqueda de 2^k. Esta representación es muy simple y facilita el tratamiento del problema, ya que la función objetivo tiene que lidiar únicamente con cromosomas que no están particionados.

La función objetivo: La función objetivo depende del experimento en turno, básicamente es la evaluación con CSA de un cromosoma y puede agregarse penalizaciones dependiendo de la cantidad de elementos considerados (descriptores en el cromosoma) y de la correlación que exista entre los elementos ahí representados.

3. RESULTADOS Y DISCUSIÓN

Los datos completos de Selwood *et al* se encuentran en el material suplementario del trabajo original (Selwood, *et. al*, 1990). Las características generales de las 53 variables involucradas se muestra en el Cuadro 1.

Como se mencionó en la sección anterior cada cromosoma fue evaluado calculando su p-valor, pero debido al tamaño del problema se estimó usando muestras de tamaño M=100,000. La mejor generación del proceso consistió de una combinación de 19 variables de las cuales 5 fueron eliminadas por tener correlaciones altas con otras variables presentes y

cuyos p-valores iniciales eran superiores. Finalmente, en la siguiente etapa se utilizó un proceso de selección de variables usando regresión lineal múltiple dejando dos modelos para fines comparativos, uno de 5 variables y otro de 3. Ver Cuadro 2.

Lo primero que observamos es que el conjunto final de 3 descriptores coinciden con los obtenidos por McFarland y Gans (Wikel y Dow, 1993). El Cuadro 3 muestra una comparación de los modelos finales con 3 variables encontradas en los trabajos de McFarland y Gans (1994), Selwood *et al* (1990), Wikel y Dow, (1993) y el nuestro basado en el algoritmo genético. Cabe mencionar que en todos los enfoques considerados, la última

Cuadro 1. Variables en los datos de Selwood *et al* (1990)

Características	Variables
Electrónicas	ATCH1-ATCH10, ESDL1-ESDL10, NSDL1-NSD10, DIPV_X, DIPV_Y, DIPV_Z, DIP_MOM
Tamaño	VDWVOL, SURF_A, MOFI_X MOFI_Y MOFI_Z PEAX_X, PEAX_Y, PEAX_Z, MOL_WT
del Sustituyente	S8_1DX, S8_1DY, S8_1DZ, S8_1CX, S8_1CY, S8_1CZ, SUM_F, SUM_R
Otras	LOGP, M_PNT

Cuadro 2. Proceso de selección de Variables

Características	Etapa 1: GA	Etapa 2: RLM (5)	Etapa 2: RLM (3)
Electrónicas	ATCH2 ATCH4 ATCH5 DIPV_X DIPV_Z ESDL3 ESDL10 NSDL7	ATCH4 ATCH5 DIPV_X	ATCH4 ATCH5 DIPIV_X
Tamaño	MOFI_X MOFI_Z PEAX_Y	MOFI_X	
del Sustituyente	S8_1DXS8_1CZ SUMR	S8_1CZ	

Cuadro 3. Comparación de modelos con 3 variables

Trabajo	Variables	R^2	$\hat{\sigma}$
McFarland y Gans	ATCH4 ATCH5 DIPV_X	0.69	0.49
Selwood et al	M_PNT LOGP ESDL6	0.55	0.59
Wikel y Dow	ATCH4 LOGP MOFI_X	0.60	0.55
AG	ATCH4 ATCH5 DIPV_X	0.69	0.49

etapa consistió en una aplicación de procedimientos de selección de regresión lineal múltiple. Las dos últimas columnas del Cuadro 3 son

R^2 = coeficiente de determinación múltiple y $\hat{\sigma} = \sqrt{\text{Error Cuadratico Medio}}$

McFarland y Gans (1994) combinaron sus resultados con los obtenidos por Selwood *et al* y Wikel y Dow obteniendo un mejor modelo con 5 descripciones: ATCH_4, ATCH_5, DIPV_X, MOFI_Y y LOGP con

$$R^2 = 0.83 \text{ y } \hat{\sigma} = 0.47.$$

En las figuras 1, 2 y 3 mostramos gráficamente la separación de compuestos producida por los tres conjuntos de descriptores considerados. La figura 1 corresponde al conjunto de descriptores seleccionados mediante el enfoque basado en el algoritmo genético, podemos notar que en esa gráfica es donde se observa la máxima separación entre compuestos activos e inactivos, este hecho era de esperarse puesto que nuestro criterio de selección está basado en CSA el cual distingue conjuntos de descriptores que producen clusters compactos.

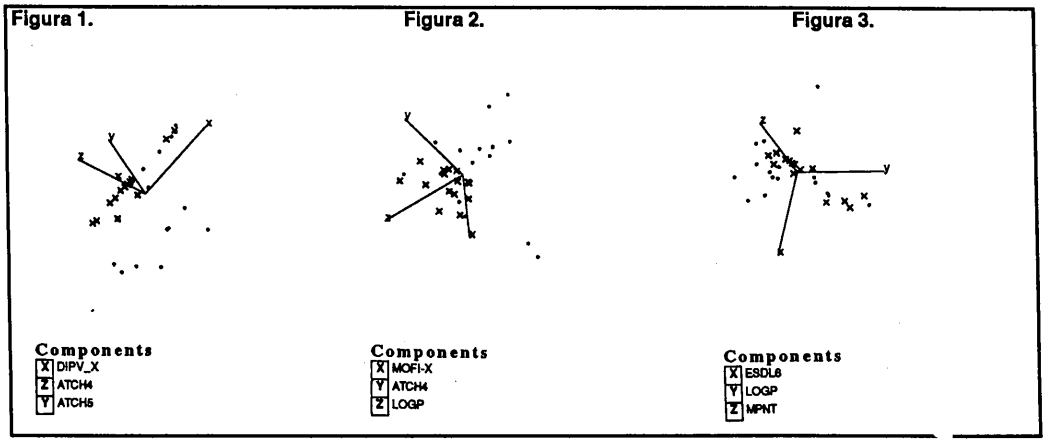


Figura 1. Separación de compuestos mediante DIPV_X, ATCH4 y ATCH5.

Figura 2. Separación de compuestos mediante MOFI_X. ATCH4 y LOGP (Wikel y Dow).

Figura 3. Separación de compuestos mediante ESDL6, LOGP y M_PNT (Selwood *et al*).

x: *Activos* •: *Inactivos*

4. CONCLUSIONES

La implementación de algoritmos genéticos como mecanismo de búsqueda muestra resultados positivos brindando para el conjunto de 53 descriptores y 31 compuestos de Selwood *et al*, un conjunto reducido de 3 descriptores en concordancia con los obtenidos por McFarland y Gans y mejores que los obtenidos por mapeos no lineales por Selwood *et al* y por la implentación de redes neuronales de Wikel y Dow.

McFarland y Gans muestran un modelo excelente de 5 variables, sin embargo, cabe mencionar que es la combinación de 3 enfoques diferentes. Esto implica que para poder obtener dicho modelo deberían implementarse todos los métodos, lo cual es claramente

incosteable. Por otra parte auqnue el modelo da un buen ajuste, contiene un par de variables altamente correlacionadas, cosa que las metodologías propuestas habían tratado de evitar desde las etapas iniciales de la selección de variables.

El algoritmo genético comienza su búsqueda sobre un espacio mayor (los 53 descriptores) razón por la cual se esperaba que diese un resultado apropiado. Se pretende seguir trabajando en la implementación del algoritmo con ciertas restricciones, como podría ser el nivel específico de actividad biológica etc. También se estudiará el comportamiento del algoritmo en la selección de nichos. El procedimiento descrito en este trabajo puede emplearse como una alternativa a los métodos stepwise de selección de variables en el análisis de regresión, tomando como función objetivo alguna combinación de diferentes indicadores de ajuste tales como el coeficiente ajustado de determinación o el cuadrado medio del error.

REFERENCIAS

Dunn W.J. y S. Wold, Structure-carcinogenicity study of 4-nitroquinoline 1-oxides using the SIMCA method of pattern recognition, *Journal of Medicinal Chemistry*, Vol. **21,** No. 10, 1001-1007, (1978).

Goldberg. D.E. *Genetic Algorithms in Search Optimization and Machine Learning*. Addison Wesley, Reading, MA, (1989).

Guerra C.M. S. *Optimización de Aspectos de Diseño Físico de Bases de Datos Utilizando Algoritmos Genéticos*. Tesis de Maestría, ITESM Campus Monterrey, (1991).

Holland J.H.. *Adaptation in Natural and Artificial Systems*. The University of Michigan Press, Ann Argor, MI, (1975).

Magee P.S., Parameter focusing- A new QSAR technique, In: IUAPC Pesticide Chemistry: Human Welfare and the Environment, J. Miyamoto y P.C. Kearny (eds.), *Pergamon Press*: Oxford 251-260, (1983).

McFarland J.W. y D.J. Gans, On the significance of clusters in the graphical display of structure-activity data, *Journal of Medicinal Chemestry*, Vol. **29**, 505-514, (1986).

McFarland J.W. y D.J. Gans, On identifying likely determinats of biological activity in high dimensional QSAR problems, *Quant. Struct. - Act. Relat.*, Vol. **13**, 11-17 (1994).

Selwood D.L., D.J. Livingstone, J.C. W. Comley, A.B. O'Dowd, A. T. Hudson, P. Jackson, Jandu K.S., V.S. Rose y J.N. Stables, Structure-activity relationship of antifilarial antimycin analogues: An multivariate pattern recognition study, *Journal of Medicinal Chemistry*, Vol. **33**, 136-142 (1990).

Spears W.M., K.A. De Jong. On the Virtues of Parameterized Uniform Crossover. *Proceedings of the Four Intl. Conference on Genetic Algorithms*, Morgan Kauffman (1991).

Valenzuela R. M., C. M. Guerra S., J.I. Icaza A. A Genetic Algorithm Approach to Partial Match retrieval Based on Hash Functions. *Proceedings of the IV International Simposium on Artificial Intelligence*, Limusa De., (1991).

Wikel J.H. y E.R. Dow, *Bioorg. Med. Chem. Lett.*, Vol. **3** 645-651 (1993).

Aspectos Computacionales del Filtro de Kalman Robustificado

ROSARIO ROMERA

Universidad Carlos III de Madrid, España

1. INTRODUCCIÓN

El filtro de Kalman proporciona un algoritmo recursivo para el estimador lineal de mínima varianza en el llamado problema lineal cuadrático Gaussiano. Es una herramienta muy utilizada en filtrado y predicción con sistemas dinámicos lineales, que admite también extensión a sistemas no lineales.

El trabajo que se presenta es una expresión del filtro de Kalman que trata simultáneamente dos problemas que aparecen en la aplicación práctica del filtro: (i) La expresión del algoritmo mediante formulación del filtro tipo raíz cuadrada, permite evitar los problemas de inestabilidad numérica que aparecen en la implementación real del filtro de Kalman. Por otra parte el algoritmo desarrollado permite su adaptación para computación en paralelo, que resulta especialmente útil en el supuesto de grandes dimensiones para el vector de estado del sistema. (ii) La contaminación de los datos, es una eventualidad frecuente en los problemas a los que se aplica filtrado de Kalman. Datos atípicos y distribuciones no Gaussianas en sistemas dinámicos lineales motivan el interés por robustificar los procedimientos de filtrado y predicción. Una aproximación basada en M-estimadores ha sido desarrollada por Cipra y Romera (1991) y es tal que proporciona buenos resultados desde el punto de vista práctico.

El algoritmo presentado en esta comunicación contiene ideas que pueden ser trasladadas a otra de las frecuentes aplicaciones del filtro de Kalman, que es el "suavizado" robustificado en este caso.

2. FORMULACIÓN DE LA PROPUESTA BÁSICA DE ROBUSTIFICACIÓN DEL FILTRO DE KALMAN

Considérese el sistema

$$x_{t+1} = F_t x_t + \omega_t \tag{1}$$

$$y_t = H_t x_t + v_t \tag{2}$$

donde x_t es el vector de estado $(n \times 1)$, e y_t es el vector de observaciones $(m \times 1)$.

Las perturbaciones del sistema ω_t y de las observaciones v_t son mutuamente independientes y verifican $E(\omega_t) = 0$, $E(v_t) = 0$, $E(\omega_t, \omega_t') = \delta_{tt} Q_t$ y $E(v_t, v_t') = \delta_{tt} R_t$. Las matrices F_t, H_t, Q_t y R_t se suponen conocidas para cada instante t.

El filtro de Kalman clásico proporciona la estimación del estado para el instante t $(\hat{x}_t^t)$ como combinación lineal de una estimación del estado en el instante $(t\text{-}1)$ $(\hat{x}_t^{t-1})$ y los datos de observaciones hasta el instante $(t\text{-}1)$ (y^{t-1}). El estimador mínima varianza del estado

$\hat{x}_t^t = E[x_t \mid y^t]$ y su matriz de covarianzas $P_t^t = E[(x_t - \hat{x}_t^t)(x_t - \hat{x}_t^t)' \mid y^t]$ donde $y^t = \{y_0, y_1, \ldots, y_t\}$ vienen dados por el algoritmo recursivo $\hat{x}_t^t = \hat{x}_t^{t-1} + P_t^{t-1} H_t' (H_t P_t^{t-1} H_t' + R_t)^{-1} (y_t - H_t \hat{x}_t^{t-1})$ y $P_t^t = P_t^{t-1} - P_t^{t-1} H_t' (H_t P_t^{t-1} H_t' + R_t)^{-1} H_t P_t^{t-1}$.

La predicción del estado del sistema toma la expresión

$$\hat{x}_{t+1}^t = F_t \ \hat{x}_t^t \tag{3}$$

$$P_{t+1}^t = F_t P_t^t F_t' + Q_t. \tag{4}$$

Una expresión análoga para la predicción del estado es la relación recursiva

$$\hat{x}_{t+1}^t = F_t \hat{x}_t^{t-1} + K_t (y_t - H_t \hat{x}_t^{t-1}) \tag{5}$$

$$P_{t+1}^t = F_t P_t^{t-1} F_t' + Q_t - K_t R_{et} K_t' \tag{6}$$

donde $K_t = F_t P_t^{t-1} H_t' R_{et}^{-1}$ es la matriz de ganancia del filtro de Kalman de dimensión $(n \times m)$, y $R_{et} = H_t P_t^{t-1} H_t' + R_t$ es la matriz de covarianzas de las innovaciones de dimensión $(m \times m)$.

Para observaciones contaminadas, las ecuaciones del sistema vienen dadas por

$$x_{t+1} = F_t \ x_t + \omega_t \qquad \omega_t \ \text{iid} \ N(0, Q_t)$$
$$y_t = H_t \ x_t + v_t \qquad v_t \ \text{iid} \ \varepsilon\text{-contaminadas} \ N(0, R_t)$$

siendo las perturbaciones $\{\omega_t\}$ y $\{v_t\}$ mutuamente independientes. Los valores de predicción dados en (3) y (4), se obtienen ahora aplicando algún tipo de robustificación al filtro de Kalman.

La falta de robustez del filtro original, ha sido ampliamente tratada en la literatura especializada, y diversas líneas de aproximación sugeridas, ver Meinhold y Singpurwalla (1989), Peña, y Guttman (1988), Servi, y Ho (1981) entre otros. En Cipra y Romera (1991) se presenta la obtención de una aproximación basada en principios de M-estimación. La eficiencia computacional es notable frente a otros procedimientos. Este trabajo proporciona las expresiones robustas del filtro dadas por

$$\hat{x}_t^t = \hat{x}_t^{t-1} + P_t^{t-1} H_t' [H_t P_t^{t-1} H_t' + R_t^{1/2} W_t^{-1} R_t^{1/2}]^{-1} (y_t - H_t \ \hat{x}_t^{t-1}), \tag{7}$$

$$P_t^t = P_t^{t-1} - P_t^{t-1} H_t' [H_t P_t^{t-1} H_t' + R_t^{1/2} W_t^{-1} R_t^{1/2}]^{-1} H_t P_t^{t-1}, \tag{8}$$

donde los valores de predicción en el instante t para el instante $(t+1)$, $\hat{x}_{t+1}^t = E\left(x_{t+1} \mid y^t\right)$ y $P_{t+1}^t = E\left[\left(x_{t+1} - \hat{x}_{t+1}^t\right)\left(x_{t+1} - \hat{x}_{t+1}^t\right)' \mid y^t\right]$, se construyen según

$$\hat{x}_{t+1}^t = F_t \ \hat{x}_t^t, \tag{9}$$

$$P_{t+1}^t = F_t P_t^t F'_t + Q_t. \tag{10}$$

La expresión $R_t^{1/2}$ denota la matriz raíz cuadrada de R_t y $W_t = \text{diag}\ \{w_{1t}, \ldots, w_{mt}\}$ de dimensión $m \times m$, donde $w_{jt} = \psi_j (s_{jt} - b_{jt} \ \hat{x}_t^{t-1}) / (s_{jt} - b_{jt} \ \hat{x}_t^{t-1})$, y $\psi_1, \ldots \psi_m$ son psi-funciones adecuadas para la robustificación.

$$B_t = \begin{pmatrix} b_{1t} \\ \vdots \\ b_{mt} \end{pmatrix} = R_t^{-1/2} H_t, \qquad S_t = \begin{pmatrix} S_{1t} \\ \vdots \\ S_{mt} \end{pmatrix} = R_t^{-1/2} y_t$$

($R_t^{-1/2}$ es la matriz inversa de $R_t^{1/2}$ y b_{jt} son las $1 \times n$ filas de la matriz B_t).

Las expresiones (7) - (10) pueden reescribirse de forma predictiva como $\hat{x}_{t+1}^t = F_t \hat{x}_t^{t-1} + K_t(y_t - H_t \hat{x}_t^{t-1})$ y $P_{t+1}^t = F_t P_t^{t-1} F_t' + Q_t - K_t R_{et} K_t'$, donde R_{et} es la matriz de covarianzas de las innovaciones con expresión $R_{et} = H_t P_t^{t-1} H_t' + R_t^{1/2} W_t^{-1} R_t^{1/2}$ y K_t es la matriz de ganancia del filtro $K_t = F_t P_t^{t-1} H_t' R_{et}^{-1}$..

En el caso particular $m = 1$ con observaciones y_t escalares y seleccionando como psi-función la función de Huber ψ_H de la forma

$$\Psi_H(z) = \begin{cases} z & |z| \le c \\ c\ \mathrm{sgn}(z) & |z| > c \end{cases}$$

(se comprueba que esta elección de la función psi en el caso de distribuciones normales ε-contaminadas proporciona un estimador robusto que es óptimo en el sentido mini-max) entonces las fórmulas obtenidas anteriormente se pueden reemplazar por las no aproximativas dadas por $\hat{x}_{t+1}^t = F_t \hat{x}_t^{t-1} + r_{et} r_t^{-1/2} \psi_H(r_{et}^{-1} r_t^{1/2} (y_t - h_t \hat{x}_t^{t-1})) K_t$, donde $r_{et} = h_t P_t^{t-1} h_t' + r_t$, y $K_t = r_{et}^{-1} F_t P_t^{t-1} h_t'$. La matriz de covarianzas P_{t+1}^t tiene la expresión $P_{t+1}^t = F_t P_t^{t-1} F_t' + Q_t - r_{et} K_t K_t'$.

3. FORMULACIÓN DEL FILTRO DE KALMAN RAÍZ CUADRADA

La implementación real del filtro de Kalman, requiere en su versión más general un coste operacional de $O(n^3)$ para cada estimación del estado del sistema, siendo n la dimensión del vector estado. Procedimientos algebráicos de descomposición de las matrices tipo Varianzas-covarianzas (P_t^{t-1}) o de Ganancia del filtro ($K(t)$) que aparecen en la llamada *ecuación de Riccati del filtro de Kalman*, cuya expresión está dada en (5) y (6), han permitido desarrollar versiones computacionalmente más eficientes. Un ejemplo de ello son los *filtros de Kalman raiz cuadrada* que utilizando la factorización de matrices semidefinidas positivas, hacen uso de una descomposición del tipo

$$LDL', \tag{11}$$

donde L es una matriz triangular inferior con unos en la diagonal principal y D es matriz diagonal.

En el caso del filtro de Kalman robusto de la sección 1 (ver Cipra y Romera, 1991) la idea básica de esta formulación raíz cuadrada consiste en mantener las matrices P_t^{t-1} y R_{et} en la forma factorizada expresada en (11), i.e. $P_t^{t-1} = L_{pt} D_{pt} L_{pt}'$, $R_{et} = L_{et} D_{et} L_{et}$, donde L_{pt}, L_{et} son matrices triangulares inferiores con unos en la diagonal principal y siendo D_{pt}, D_{et} matrices diagonales.

Los valores de entrada para el algoritmo raíz cuadrada robusto en el instante t vienen dados por las matrices F_t, H_t, R_t, Q_t, el vector de observaciones y_t y el valor de predicción $\hat{x}_t^{t-1}$ para el instante t.

Se construyen matrices

$$U = \begin{pmatrix} I_m & H_t & L_{pt} & 0 \\ 0 & F_t & L_{pt} & I_n \end{pmatrix}$$

de dimensión $(m+n) \times (m+2n)$ (I_m es la $m \times m$ matriz identidad), y

$$V = \begin{pmatrix} R_t^{1/2} W_t^{-1} R_t^{1/2} & 0 & 0 \\ 0 & D_{pt} & 0 \\ 0 & 0 & Q_t \end{pmatrix}$$

es de dimensión $(m+2n) \times (m+2n)$. El procedimiento del algoritmo consiste básicamente en la factorización $UVU' = LDL'$, para el cual se requiere la construcción de las matrices L de dimensión $(m+n) \times (m+2n)$ y D de dimensión $(m+2n) \times (m+2n)$ del mismo tipo al sugerido en (11). Es fácil probar que las matrices L y D tienen la forma

$$L = \begin{pmatrix} L_{et} & 0 & 0 \\ K_t L_{et} & L_{p,t+1} & 0 \end{pmatrix} \tag{12}$$

$$D = \begin{pmatrix} D_{et} & 0 & 0 \\ 0 & D_{p,t+1} & 0 \\ 0 & 0 & D^a \end{pmatrix} \tag{13}$$

donde D^a puede ser una matriz diagonal arbitraria de dimensión $n \times n$.

La salida del algoritmo para el instante t devuelve $\hat{x}_{t+1}^t = F_t \hat{x}_t^{t-1} + (K_t L_{et}) L_{et}^{-1} (y_t - H_t \hat{x}_t^{t-1})$ y $P_{t+1}^t = L_{p,t+1} D_{p,t+1} L_{p,t+1}$, donde las matrices L_{et}, $K_t L_{et}$, $L_{p,t+1}$ y $D_{p,t+1}$ se toman a partir de (12) y (13) (recursivamente estas matrices $L_{p,t+1}$ y $D_{p,t+1}$ son las entradas para la iteración en el instante t+1).

4. ALGORITMO

La descripción del algoritmo para una etapa es la siguiente:

Tiempo t	Entradas:	F_t, H_t, R_t, Q_t
	$t \geq 0$	y_t
		$\hat{X}_t^{t-1}$
		L_{pt}, D_{pt}
	Salida:	$\hat{X}_{t+1}^t$
	$t > 0$	$L_{p(t+1)}, D_{p(t+1)}$ y se calcula P_{t+1}^t

Los valores de predicción se obtienen según:

- Construir las matrices A, $(m+n)\times(m+2n)$, y D, $(m+2n)\times(m+2n)$,

$$A=\begin{pmatrix} I & H_t L_{pt} & 0 \\ 0 & F_t L_{pt} & I \end{pmatrix}, \qquad D=\begin{pmatrix} R_t^{1/2} W_t^{-1} R_t^{1/2} & 0 & 0 \\ 0 & D_{pt} & 0 \\ 0 & 0 & Q_t \end{pmatrix}$$

- Obtener A^* y D^* a partir de las expresiones siguientes

$$A^*=\begin{bmatrix} L_{et} & 0 & 0 \\ K_t L_{et} & \mathrm{L}_{\mathrm{p}\ t+1} & 0 \end{bmatrix}, \qquad D^*=\begin{bmatrix} D_{et} & 0 & 0 \\ 0 & \mathrm{D}_{\mathrm{p}\ t+1} & 0 \\ 0 & 0 & \mathrm{D}_{\mathrm{a}} \end{bmatrix}$$

siendo las dimensiones de A^* y D^* $(m+n)\times(m+n+n)$ y $(m+n+n)\times(m+n+n)$ respectivamente, D_a matriz diagonal arbitraria de dimensión $n\times n$, se comprueba la relación $ADA' = A^*D^*A^{*\prime}$ que proporciona los valores del siguiente paso:

- Calcular $\hat{x}_t^{t+1} = F_t \hat{x}_t^{t+1} + (K_t L_{et}) L_{et}^{-1}(y_t - H_t \hat{x}_t^{t-1})$, donde $K_t L_{et}$ y L_{et} se obtienen a partir de A^*.

- Tomar $\hat{X}_{t+1}^t L_{p(t+1)}$ y $D_{p(t+1)}$ como entrada para la siguiente iteración.

5. CASO PARTICULAR: OBSERVACIONES ESCALARES

Considérese el sistema dado por (1) y (2) con m=1. Los valores de predicción para el instante t construidos en (t-1) según (5) y (6) toman ahora la expresión $\hat{x}_{t+1}^t = F_t \hat{x}_t^{t-1} + K_t r_{et} r_t^{-1/2} \psi_H(r_{et}^{-1} r_t^{1/2}(y_t - h_t \hat{x}_t^{t-1}))$ y $P_{t+1}^t = F_t P_t^{t-1} F_{t'} + Q_t - K_t\ r_{et} K_t'$, donde $K_t = F_t P_t^{t-1} h_t' r_{et}^{-1}$ y $r_{et} = h_t P_t^{t-1} h_t' + r_t$.

En este caso las matrices A, $(1+n)\times(1+2n)$, D,$(1+2n)\times(1+2n)$, A^*, $(1+n)\times(1+2n)$, y D^*, $(1+2n)\times(1+2n)$, toman la forma

$$A=\begin{pmatrix} 1 & \mathrm{h}_t L_{pt} & 0 \\ 0 & \mathrm{F}_t L_{pt} & \mathrm{I} \end{pmatrix}, \qquad \mathrm{D}=\begin{pmatrix} \mathrm{r}_t & 0 & 0 \\ 0 & \mathrm{D}_{\mathrm{pt}} & 0 \\ 0 & 0 & \mathrm{Q}_t \end{pmatrix},$$

$$A^*=\begin{pmatrix} 1 & 0 & 0 \\ K_t & \mathrm{L}_{\mathrm{p}\ t+1} & 0 \end{pmatrix}, \qquad \mathrm{D}^*=\begin{pmatrix} r_{et} & 0 & 0 \\ D_{pt+1} & 0 & 0 \\ 0 & 0 & \mathrm{D}_{\mathrm{a}} \end{pmatrix}.$$

Así se obtiene K_t, r_{et}, $L_{p(t+1)}$ y $D_{p(t+1)}$ a partir de A^* y D^*, y los valores de predicción para (t+1) en el instante t: $\hat{x}_{t+1}^t = F_t \hat{x}_t^{t-1} + K_t r_{et} r_t^{-1/2} \psi_H(r_{et}^{-1} r_t^{1/2}(y_t - h_t \hat{x}_t^{t-1}))$ y $P_{t+1}^t = L_{pt+1} D_{pt+1} L_{pt+1}'$.

Observación la formulación del filtro de Kalman raíz cuadrada aquí presentada (ver Romera, 1993) es susceptible de ser implementada en computación en paralelo, reduciendo además la complejidad computacional a $O(n)$.

6. RESULTADOS NUMÉRICOS

La implementación de los procedimientos de robustificación descritos anteriormente, se ha llevado a cabo mediante un programa FORTRAN. Se han utilizado datos simulados con diferentes casos de contaminación.

Se ha implementado un procedimiento de robustificación del filtro de Kalman basado en mixturas de probabilidades a posteriori de distribuciones normales (Peña y Guttman, 1988) y obtenido resultados comparativos con ambos procedimientos de robustificación. Desde el punto de vista de robustez en las estimaciones del estado obtenidas, los resultados son buenos y comparables en ambos casos. La eficiencia computacional valorada en tiempos de ejecución entre ambos métodos proporciona una reducción considerable para la robustificación basada en M-estimadores.

Se aportan a esta comunicación los valores numéricos de las simulaciones realizadas. El software desarrollado, está disponible para quien los solicite.

REFERENCIAS

Cipra, T. y Romera, R. (1991), Robust Kalman filter and its application in time series analysis, *Kybernetika*, **27**, 481-494.

Cipra, T., Romera, R. y Rubio, A., (1992). Working Paper **92-09**, *Statistics and Econometrics Series*, Universidad Carlos III de Madrid.

Meinhold, R.J.y Singpurwalla, N.D. (1989), Robustification of Kalman filter models, *J. Amer. Statist. Assoc.*, **84**, 479-486.

Peña, D. y Guttman, I. (1988), A Bayesian Approach to Robustifying Kalman Filtering, Capítulo 9 de Bayesian Analysis of Time Series and Dynamic Models, J.C. Spall, Editor, Marcel Dekker, Inc, New York.

Romera, R. y Cipra, T. (1993), A Parallel Kalman Filter via The Square Root Kalman Filtering, Working Paper 93-14, *Statistics and Econometrics Series* **12**, Universidad Carlos III de Madrid.

Romera,R. y Cipra, T. (1995), On Practical Implemetation of Robust Kalman Filter, *Commun. Statist.-Simula. and Comp.*, **24** (2), 461-488.

Servi, L.D. y Ho, Y.C. (1981): Recursive estimation in the presence of uniformly distributed measurement noise, IEEE *Trans. Automat. Control*, AC-26, 563-565.

Obtención de Un Indice de Marginación Social por Localidad en la Reserva de la Biosfera Sierra de Manantlán Utilizando Métodos Multivariados

M. P. ROSALES, J. J. ROSALES

Univ. Guadalajara, México

y

S. H. GRAF

INE-SEMARNAP, México

1. INTRODUCCIÓN

Situada en territorio de siete municipios de los Estados de Jalisco y Colima, la Reserva de la Biosfera Sierra de Manantlán alberga a poco más de 30,000 personas, distribuidas en 26 comunidades agrarias. Desde una perspectiva sociodemográfica existen grandes diferencias entre los municipios que conforman la reserva, según el Consejo Nacional de la Población (CONAPO 1990) tres de ellos son considerados entre los más marginados de la región y cuatro presentan un grado de marginación bajo. El índice de marginación obtenido por el CONAPO agrupa a los municipios según su grado de marginación, sin embargo, este índice no permite identificar las diferencias en los niveles de bienestar de la población en las diferentes localidades al interior de cada municipio en su contexto regional. Con el fin de contar con una mejor aproximación de las condiciones de bienestar social de la población que vive en la reserva, en este trabajo se obtiene un índice de marginación (IM) por localidad que permite agrupar a los diferentes asentamientos humanos según su grado de marginación. Para la construcción del IM se utilizan indicadores sociodemográficos que están relacionados con algunas características de la población como el nivel de educación y los servicios de bienestar social en cada localidad. El estudio se realizó para 74 localidades en la Reserva de la Biosfera Sierra de Manantlán. Los datos utilizados fueron tomados del XI Censo General de Población y Vivienda 1990 (INEGI 1990).

2. VARIABLES REGISTRADAS

En este estudio se consideraron 7 de los 9 indicadores utilizados por el CONAPO (1990) para la obtención del Indice de Marginación por localidad:

2.1 Analfabetismo (% de población analfabeta ≥ de 15 años de edad):

$$\text{ANALFA} = \frac{\text{PA}}{\text{PT}} * 100$$

donde PA = Población analfabeta ≥15 años de edad.
PT = Población Total ≥ 15 años de edad.

2.2 Sin Primaria (% Población sin primaria completa ≥ 15 años de edad):

$$\text{SINPRI} = \frac{\text{PSP}}{\text{PT}} * 100$$

donde PSP = Población sin primaria completa ≥ 15 años de edad.
PT = Población Total ≥ 15 años de edad.

2.3 Sin Drenaje (% ocupantes en viviendas sin disponibilidad de drenaje):

$$SINDRE = \frac{TOSD}{TO} * 100$$

donde TOSD = Total de Ocupantes sin Drenaje.
TO = Total de Ocupantes.

2.4 Sin Electricidad (% ocupantes sin disponibilidad de electricidad):

$$SINELE = \frac{TOSE}{TO} * 100$$

donde TOSE = Total de Ocupantes sin Electricidad.
TO = Total de Ocupantes.

2.5 Sin Agua (% ocupantes sin disponibilidad de agua):

$$SINAGU = \frac{TOSA}{TO} * 100$$

donde TOSA = Total de Ocupantes que no tienen agua.
TO = Total de ocupantes.

2.6 Hacinamiento (% de viviendas con algún nivel de hacinamiento):

$$HACINA = \frac{TH}{TV} * 100$$

donde HT = Total de viviendas con algún nivel de hacinamiento .
TV = Total de Viviendas.

2.7 Piso de Tierra (% de ocupantes en viviendas con piso de tierra):

$$PISOTI = \frac{TOPT}{TO} * 100$$

donde TOPT = Total de Ocupantes en viviendas con piso de tierra.
TO = Total de ocupantes en viviendas.

El total de ocupantes con piso de tierra en sus viviendas se obtuvo de la diferencia entre las viviendas totales y las viviendas sin piso de tierra multiplicado por el promedio de ocupantes por viviendas. Para la construcción del índice de marginación se procedió primero, a la aplicación de un análisis de agrupamiento K-means) sobre los 7 indicadores socioeconómicos (variables) involucrados en el estudio, la intención de este análisis exploratorio fue ver si era posible la formación natural de 5 grupos, pensando en las categorías de marginación que queremos definir sobre el total de las localidades (marginación muy alta, alta, media, baja y muy baja). Posteriormente se empleo un Análisis de Componentes Principales con objeto de construir una cl (variables) única de los siete indicadores socioeconómicos, tal que ésta explique la máxima proporción de la variación total en el conjunto de variables. Para la construcción del IM se utilizó la primera Componente Principal (CP_1), utilizando los 7 indicadores socioeconómicos; con esto a la CP_1 se le define como indicador resumen.

3. DETERMINACIÓN DE LOS GRUPOS DE MARGINACIÓN

Se utilizó el método de Estratificación Optima, el cual permite agrupar adecuadamente a las localidades, basándose en la función de densidad f(x) definida a partir del IM. Aquí es necesario encontrar los puntos x_1, x_2, x_3 y x_4 que permitan la agrupación de los valores de las localidades mediante la condición:

$$\begin{aligned} &\text{Gpo I} && x_0 \le x \le x1 \\ &\text{Gpo II} && x_1 < x \le x2 \\ &\text{Gpo III} && x_2 < x \le x3 \\ &\text{Gpo IV} && x_3 < x \le x4 \\ &\text{Gpo V} && x_4 < x \le x5 \end{aligned}$$

Se puede probar que los cortes que logran minimizar la varianza de un estimador poblacional se da con ayuda de una transformación de la función de densidad f(x). (Dalenius y Hodge, 1959). Esta es: $y(u) = \int_{-\infty}^{u} \sqrt{f(t)}\, dt$. Dados los puntos x_1, x_2, x_3 y x_4 se tiene que:

$$y(\infty) = \int_{-\infty}^{x_1} \sqrt{f(t)}\, dt + \int_{x_1}^{x_2} \sqrt{f(t)}\, dt + \ldots + \int_{x_4}^{\infty} \sqrt{f(t)}\, dt = \int_{x_0}^{x_1} \sqrt{f(t)}\, dt + \int_{x_1}^{x_2} \sqrt{f(t)}\, dt + \ldots + \int_{x_4}^{x_5} \sqrt{f(t)}\, dt$$

La $y(x_h) - y(x_{h-1})$ para h = 1,..., 5, es constante; esto es, si $y(x_h)$ y $y(x_{h-1})$ toman valores de tal manera que:

$$\int_{x_{h-1}}^{x_h} \sqrt{f(t)}\, dt = \frac{H}{5} \qquad h = 1, \ldots, 5$$

Debemos calcular el valor de H. Para ello, es preciso dividirlo entre el número de grupos deseados. Dividamos en 10 partes iguales el rango de variación del IM y contemos el número de observaciones que pertenecen a cada una de ellas. De cada frecuencia se obtiene la raíz cuadrada y se acumula a la que le antecede.

$$\int_{x_0}^{x_1} \sqrt{f(t)}\, dt + \int_{x_1}^{x_2} \sqrt{f(t)}\, dt + \ldots + \int_{x_4}^{x_5} \sqrt{f(t)}\, dt = 25.763$$

Dada esta igualdad debemos encontrar los valores x_1, x_2, x_3 y x_4 que cumplen con la igualdad:

$$\int_{x_0}^{x_1} \sqrt{f(t)}\, dt = \int_{x_1}^{x_2} \sqrt{f(t)}\, dt = \ldots = \int_{x_4}^{x_5} \sqrt{f(t)}\, dt = \frac{(25.763)}{5} = 5.153$$

De esta manera, el primer estrato o grupo está formado por las localidades cuyo índice sea menor o igual al valor de x_1. El valor x_1 es aquel en donde la función y(u) se acumula hasta 5.153. Como 5.153 es un número intermedio entre 4.0 y 6.0, las localidades que cumplen con lo anterior son las correspondientes a las primeras dos clases. Así, las primeras 8 localidades (4 de la primera clase y 4 de la segunda) pertenecen al primer grupo. De la misma manera, el valor x_2 es aquel que junto con x_1 determina que:

$$\int_{x_1}^{x_2} \sqrt{f(t)}\, dt = 5.153,$$

lo cual implica que: $\int_{x_0}^{x_2} \sqrt{f(t)}\, dt = (5.153)*2 = 10.305$

El valor 10.305 esta entre 9.464 y 12.292, por lo que el segundo grupo se compone de 10 localidades (4 de la tercera clase, 3 de la cuarta y 3 de la quinta). Siguiendo el mismo razonamiento se llega a que:

$$\int_{x_0}^{x_3} \sqrt{f(t)}\, dt = (5.153)*3 = 15.458.$$

Dado que 15.035 es mayor que 12.292 y menor que 15.898, en el tercer grupo se ubicarían las 8 localidades de las 6 clases. De manera análoga se obtiene que las 13 localidades de la séptima clase y las 14 de la octava, en total 27, pertenecen al cuarto grupo, y las 21 localidades de las clases restantes conforman el quinto grupo. El cuadro 1 muestra lo anterior.

CUADRO 1

Clase	Intervalo para el IM	Punto medio	Frec. de Clases	Raíz cuad. de Frec.	Raíz cuad. Frec. Acum.	Lím. de puntos de corte	Núm. de localidades por grupo
1	[-4.670, -3.913]	-4.292	4	2.0	2.0		
2	(-3.913, -3.155]	-3.534	4	2.0	4.0	5.153	8
3	(-3.155, -2.397]	-2.776	4	2.0	6.0		
4	(-2.397, -1.640]	-2.018	3	1.732	7.732		
5	(-1.640, -0.882]	-1.261	3	1.732	9.464	10.305	10
6	(-0.882, -0.124]	-0.503	8	2.828	12.292	15.458	8
7	(-0.124, 0.633]	-0.255	13	3.606	15.898		
8	(0.633, 1.391]	1.012	14	3.742	19.640	20.610	27
9	(1.391, 2.149]	1.770	17	4.123	23.753		
10	(2.149, 2.907]	2.528	4	2.0	25.763	25.763	21

Entonces una localidad se considera de marginación:

MUY BAJA si su IM pertenece al intervalo [-4.670, -3.155]
BAJA si su IM pertenece al intervalo (-3.155, -0.882]
MEDIA si su IM pertenece al intervalo (-0.882, -0.124]
ALTA si su IM pertenece al intervalo (-0.124, 1.391]
MUY ALTA si su IM pertenece al intervalo (1.391, 2.907]

REFERENCIAS

Dalenius, T. 1959. Minimum Variance Stratification. *Journal of American Statistical Association*. Vol. **54**. pp. 88-101.

González B.B.L. 1994. *Análisis de Componentes Principales entre Grupos de Poblaciones*. Tesis de Licenciatura en Matemáticas. Universidad de Guadalajara.

Johnson, R.A., and Wichern, D.W. 1982. *Applied Multivariate Statistical Analysis*. Prentice Hall, inc.Englewood Cliff, N.J.

SAS Institute Inc. 1988. *SAS/STAT User's Guide*, Release 6.03 Edition. Cary, NC:SAS Institute Inc. 1028 pp.

INEGI (1990) *XI Censo General de Población y Vivienda* 1990.

Análisis Post-Ajuste para Modelos Lineales Generalizados para Datos Longitudinales

SILVIA RUÍZ VELASCO A.
IIMAS-UNAM, México

Liang y Zeger (1986) propusieron una extensión de los modelos lineales generalizados para el caso de datos longitudinales. El interés fundamental es la dependencia de la variable respuesta en las variables explicativas. En este caso la dependencia temporal entre las respuestas es vista como "estorbo".

El modelo se plantea de la siguiente manera: sea $Y_i = \left(y_{i1},\ldots,y_{in_i}\right)'$ el vector de variables respuesta y $X_i = \left(x_{i1},\ldots,x_{in_i}\right)'$ la matriz de dimensión $n_i \times p$ de variables explicativas para el sujeto i-ésimo ($i = 1,\ldots,K$). Se supone que la densidad marginal de y_{it} es

$$f(y_{it}) = \exp[\{y_{it}\theta_{it} - b(\theta_{it}) + c(y_{it})\}]\phi,$$

donde $\theta_{it} = h(\eta_{it})$, $\eta_{it} = x_{it}\beta$, por lo que

$$E(y_{it}) = b'(\theta_{it}), \qquad \text{var}(y_{it}) = b''(\theta_{it})$$

Adicionalmente se supone una estructura de covarianza para las observaciones repetidas y se proponen las llamadas ecuaciones generalizadas de estimación (GEE) para obtener estimadores de los parámetros involucrados en el predictor lineal. Entonces la ecuaciones de estimación generalizadas están dadas por:

$$\sum_{i=1}^{K} D_i^T V_i^{-1} S_i = 0$$

donde $V_i = A_i^{1/2} R(\alpha) A_i^{1/2} / \phi$, $D_i = d\{b'(\theta_{it})\}/d\beta = A_i \Delta_i X_i$, $A_i = \text{diag}\{b''(\theta_{it})\}$, $\Delta_i = \text{diag}(d\theta_{it}/d\eta_{it})$, $S_i = Y_i - b'(\theta_{it})$. V_i corresponde a la cov(Y_i) si $R(\alpha)$ es la verdadera matriz de correlación de las Y_i's.

Entre las estructuras de correlación mas utilizadas se encuentran:

1. Independencia ($R(\alpha) = I$).
2. $\alpha = (\alpha_1,\ldots,\alpha_{n-1})'$, donde $a_t = \text{corr}(y_{it}, y_{it+1})$, y $R(\alpha)$ tridiagonal con elementos α_t. Casos particulares es $\alpha_t = \alpha$ para todo t (1-dep). Extensiones a m-dependencia.
3. $\text{corr}(y_{it}, y_{it'}) = \alpha$ para todo t diferente de t'. (intercambiable)
4. $\text{corr}(y_{it}, y_{it'}) = \alpha^{|t-t'|}$ autorregresivo de orden uno.
5. no especificada.

Estas ecuaciones pueden ser vistas como una generalización de la cuasi-verosimilitud en el caso de modelos lineales generalizados y los estimadores obtenidos gozan de las propiedades de los estimadores cuasi verosímiles en el caso de MLG. Sin embargo, la distribución de las estadísticas post-ajuste como la generalización del cociente de cuasi-verosimilitud no tiene una distribución ji cuadrada.

Una manera de ajustar estos modelos es utilizando el método Gauss-Newton, dados los estimadores (de momentos) $\tilde{\alpha}$ y $\tilde{\phi}$, se sugiere obtener el estimador de β como:

$$\hat{\beta}_{j+1} = \hat{\beta}_j - \left\{\sum_{i=1}^{K} D_i^T(\hat{\beta}_j)\tilde{V}_i^{-1}(\hat{\beta}_j)D_i(\hat{\beta}_j)\right\}^{-1}\left\{\sum_{i=1}^{K} D_i^T(\hat{\beta}_j)\tilde{V}_i^{-1}(\hat{\beta}_j)S_i(\hat{\beta}_j)\right\}$$

donde $\tilde{V}_i(\hat{\beta}_j) = V_i(\hat{\beta}_j, \tilde{\alpha}\{\hat{\beta}_j, \tilde{\phi}(\hat{\beta}_j)\})$.

Liang y Zeger (1986), muestran que la consistencia del estimador de β, así como la del estimador de la varianza de β, depende sólo de la correcta especificación de la media, pero no de la especificación de R. Asimismo, a través de un estudio de simulación, muestran que la eficiencia de los estimadores generalizados utilizando una estructura diferente con respecto a la estructura real es muy alta.

Este trabajo esta motivado en el hecho de que a pesar de obtener estimadores consistentes y hasta cierto punto eficientes con diferente especificación de la matriz de correlación, el ajuste global del modelo, así como la posible influencia de algún individuo puede ser afectada por la elección de dicha matriz.

En este trabajo nos limitaremos a las estructuras de correlación 1-3, en las que no es necesario estimar ϕ. En cuanto al ajuste general del modelo, para el caso de que la distribución marginal sea normal, se propone utilizar como estadística de prueba la T^2 de Hotelling. Para utilizar la T^2 proponemos que una vez ajustado el modelo se obtengan los residuales, marginalmente estos tienen una distribución normal univariada, y los valores para cada individuo están correlacionados. Por lo tanto si la estructura de correlación esta bien especificada, la estadística T^2 estará mas cerca de parecerse a una F, que si la estructura de correlación esta mal especificada. De hecho dependiendo del estimador utilizado para α, si la normalidad conjunta se da, la distribución sería exacta.

Con el objeto de ver qué tan lejos está de la distribución F la estadística T^2 para los casos en que no se utiliza la estructura de correlación real realizamos un estudio de simulación. Simulamos cuarenta repeticiones de 10 observaciones en 20 individuos del modelo $y=\beta_1 x_1 + \beta_2 x_2$. Donde x_1 es un entero aleatorio que puede tomar el valor 1 ó 2, y x_2 una variable aleatoria uniforme. Esto para cuatro estructuras de correlación: 1-dependencia, 2-dependencia, intercambiable e independiente. En cada uno de estos casos ajustamos las cuatro estructuras de correlación y calculamos la estadística T^2.

Los valores obtenidos fueron graficados contra las estadísticas de orden correspondientes a la distribución F. Los resultados obtenidos son: en el caso de una estructura de 2-dependencia, el uso de la estadística aparentemente es adecuado y se obtendría un mejor ajuste global que al utilizar cualquier otro. En el caso de que la estructura de correlación es 1-dependencia, vemos que realmente el ajustar un modelo de independencia o 2-dependencia, cuando el real es de 1-dependencia es muy similar. En el caso de una estructura de correlación intercambiable, todos los demás casos parecen dar una idea de sobreajuste. Finalmente en el caso de independencia (basado solo en 20 simulaciones) la estadística no se comporta bien, aunque cabe aclarar que las estimaciones de los parámetros de correlación en cualquier de las otras estructuras son muy cercanas a cero.

Calculamos la eficiencia de los estimadores en los cuatro casos, siempre es muy cercana a uno, con la sola excepción del estimador de β_2 cuando se utiliza una estructura de correlación 1-dependencia, siendo la real 2-dependencia.

En términos de influencia, para las estructuras de 2-dependencia y 1-dependencia para un individuo alteramos el valor de x_1, para medir la influencia de ese individuo, tanto en el ajuste global como en el estimador de β. Calculamos el estimador de β a un paso. Las

medidas de influencia para múltiples observaciones propuestas por Cook y Weisberg (1982) son aplicables. En términos del ajuste general del modelo el comportamiento es igual.

En cuanto al cambio en los estimadores es de la siguiente manera: cuando la estructura es de 2-dependencia, el cambio en β_1 es muy alto alrededor del 40% para cualquier estructura de correlación, sin embargo β_2 se ve afectado para las estructuras de correlación incorrectas, produciendo el cambio mas grave cuando se utiliza 1-dependencia con un 25%. Cuando la estructura correcta es de 1-dependencia el cambio en β_1 es alrededor de 15% para cualquier estructura, en este caso si se utiliza la estructura de independencia β_2 también se ve afectada cambiando alrededor de 9%.

Cambiando x_2, agregando a un individuo una variable aleatoria normal, con media 1/3 y varianza 1/3, se obtiene: cuando la estructura correcta es 2-dependencia, existe un cambio en β_2 de alrededor de 34% si esta es la estructura que se utiliza, mientras que en las otras estructura este cambio es de alrededor de 44%. Si la estructura verdadera es de 1-dependencia el cambio en β_2 es más fuerte en esta misma estructura alrededor del 30%, mientras que en las otras es del alrededor del 15%.

Conclusiones: aunque en datos reales la verdadera estructura de correlación es desconocida, es un hecho que debido a las propiedades de consistencia y eficiencia de los estimadores, muchas veces no se tiene cuidado al elegir la estructura de correlación. Del trabajo realizado se puede concluir que cuando la estructura es de *m*-dependencia, funciona mejor utilizar 2-dependencia, dado que el ajuste es muy comparable a 1-dependencia cuando esta es la verdadera estructura, y que el ajuste de 1-dependencia es muy malo cuando la verdadera estructura es 2-dependencia. Si los datos siguen una estructura intercambiable, 1-dependencia funciona muy mal.

Es necesario estudiar posibles aproximaciones a la distribución real de T^2. En el caso de influencia, se debe tratar el caso de una observación de un individuo influyente, esto se complica mucho cuando la estructura de correlación es de m-dependencia, en particular si el número de repeticiones no es grande, en el caso de independencia y correlación intercambiable, los resultados de Cook y Weisberg (1982) para modelos lineales generalizados son extendibles sin dificultad.

REFERENCIAS

Cook R. D., Weisberg S.(1982) *Residuals and Influence in Regression.* Chapman and Hall. London.

Liang y Zeger(1986) Longitudinal Data Analysis using Generalizad Linear Models. *Biometrika* **73**, 13-22.

Uso de los Modelos de Regresión Logística en Estudios de Vida de Anaquel de Exportación

MARIA I. SILVERIA GRAMONT y LORENIA LÓPEZ MAZÓN
REGINALDO BÁEZ ZAÑUDO
Universidad de Sonora, México

1. INTRODUCCIÓN

La vida de anaquel en un producto hortofrutícola se considera como el período de tiempo que transcurre desde su cosecha hasta que debe de ser retirado del anaquel por considerarse en el límite de aceptabilidad, esto es, el producto tiene fallas en sus características de calidad (Akimbolu et al; 1991; Charalambous, 1992). El tomate es uno de los productos hortofrutícolas que produce México tanto para exportación como para consumo nacional (INEGI, 1994), por lo que conocer el tiempo de duración en anaquel de dichos frutos, permitirá al productor un mejor manejo de sus ventas de mercado, así como menores pérdidas.

Existen numerosos modelos matemáticos y estadísticos para predecir la vida de anaquel de alimentos (Gacula y Singh, 1984). El de uso más generalizado es el de regresión lineal, donde la variable dependiente (Y) es alguna característica de calidad, y la independiente (X) es el tiempo. Estos modelos deben suponer distribución normal de Y, y un comportamiento lineal de la variable durante el tiempo de vida de anaquel. En la mayoría de los casos, estas dos suposiciones no se cumplen. En este trabajo se está proponiendo el uso de métodos estadísticos alternativos para estimar el comportamiento de las variables durante la vida de anaquel, así como estimar los tiempos de sobrevivencia para valores determinados de las características de calidad. Estos son: la estimación de la función de sobrevivencia (o de fallas) , y el uso de la regresión logística (simple y múltiple).

2. MATERIALES Y MÉTODOS

Se seleccionaron 200 tomates con calificación de exportación, de dos procedencias en la línea de empaque: a nivel de campo en la pisca del tomate, y al final del empaque. Los frutos se dividieron en dos sublotes, uno para colocar en refrigeración a 10°C, y el otro a 20°C. Desde los 20 días en adelante los tomates almacenados a 10°C, se almacenaron a 20°C. Cada 5 días se midió el peso de los tomates, y se tomó una muestra al azar de 5 tomates de cada lote para realizar las medidas objetivas de calidad, que fueron: color, firmeza, acidez titulable, pH, grados Brix, y pérdida de peso. Además, se observaron los frutos cada 3 días para determinar su falla en calidad de consumo (Floros, 1992; U.S.D.A, 1992).

La función de fallas permite caracterizar la vida útil de un producto alimenticio. La tasa de fallas es el porcentaje de unidades que fallan en un intervalo de tiempo t_k desde que el producto se puso en el mercado. Si al inicio del período k, hay n_i unidades en observación y fallan d_i unidades durante ese período, la probabilidad de que falle cualquier unidad en el período k será: $P(Y=1)_k = p_k$. Entonces $1-p_k$ será la probabilidad de sobrevivencia. Las funciones de falla y de sobrevivencia serán:

$$h(t_k) = P[X=1 / T>t_k]= 1-F(t_k) \text{ y } S(t_k)= 1- P[X=1 / T<t_k]= F(t_k) \quad (1)$$

donde F(t) es la función de distribución de t.

Suponiendo que no hay observaciones censuradas, la función de sobrevivencia se puede estimar usando el estimador de Kaplan-Meier, (Miller, 1981):

$S_e(t_k)= \Pi_{j\cdot k} (1- d_j/n_j)$, con una varianza asintótica de $s[S_e(t_k)] = S_e(t_k)* [S_j d_j/(n_j S(t_j))]^{0.5}$ (2)

Se pueden obtener los cuantiles, tales como $q_{.5}$, de la distribución de S(t), usando:

$$q_{.5} = \text{Mín } \{t: [1- S_e(t_j)] \geq 0.5\} \quad (3)$$

Una región de confianza asintótica de $(1-\alpha)$ para $q_{.5}$, de acuerdo a resultados publicados (Brookmeyer y Crowley, 1982), será, para el caso de observaciones no censuradas:

$$R_\alpha = \{M[S_e(t)] \mid \{S_e(q_{.5})-0.5\}^2 < c_\alpha[S_e(q_{.5})]*\Sigma_{x<M} \{d_i/[n_i(n_i+d_i)]\}\} \quad (4)$$

Los métodos de regresión permiten estimar la tasa de cambio en la calidad de un producto como una función de tiempo, y por estimación inversa se puede obtener un estimador por intervalo para el tiempo correspondiente a un valor dado de la ordenada (Yo), (Gacula y Singh, 1984; Hamilton, 1992). El modelo de regresión logístico se puede aplicar en los casos en que se consideren las pérdidas como la variable dependiente, (Gacula y Singh, 1984; Lawless, 1982; Miller, 1981). Para ello consideramos $p_k= d_k/n_k$, la tasa de fallas en el período k-ésimo; entonces la razón de momios para ese período será: $m(p_k) = p_k/(1-p_k)$. El logit de la tasa de fallas es:

$$L(p_k) = \log_e[p_k/(1-p_k)] \quad (5)$$

Para determinar la relación entre vida de anaquel y características objetivas se realizó un análisis de regresión logística múltiple (mediante el PROC LOGISTIC del SAS), (SAS, 1992):

$$\text{Logit}(p_i)= b_0 + b_1\text{Firm} + b_2\text{pH} + b_3\text{Ac} + b_4\text{Brx} + b_5\text{Ld} + b_6\text{Han} + b_7\text{Chr} + b_8\text{ PP} + \varepsilon_i \quad (6)$$

Con los valores estimados de Le(p), se pueden calcular las probabilidades de falla (o de sobrevivencia), para valores de las variables independientes, y viceversa.

Para ello se toma en cuenta que $Se(t_k) = 1- \{1/[1+e^{-Le(p)}]\}= 1- \{1/[1+e^{-(b_0+bX)}]\}$ (7)

Mediante métodos multivariados se obtuvo un índice de calidad del tomate y se analizó un modelo de regresión logística para probar el valor predictivo del índice de calidad, con respecto a las pérdidas, usando el modelo: $\text{Logit}(p_k)_i = b_0 + b_1*IC_{ki} + \varepsilon_i$ (8)

3. RESULTADOS Y DISCUSIÓN

En la figura 1 se muestran las curvas de sobrevivencia estimadas para las dos temperaturas de almacenamiento (A-10 y A-20). Una comparación de curvas por el método de Wilcoxon, (Lawless, 1982) muestra diferencias significativas ($p < 0.05$) entre las dos curvas. Cuando se analizaron las curvas de sobrevivencia para los cuatro lotes, los resultados muestran que las funciones de sobrevivencia dependen de la temperatura de almacenamiento y no de la procedencia de dichos frutos. Los límites de confianza para t_{50} (Cuadro 1), estimados de acuerdo a la ecuación (4), muestran gran semejanza con los estimados mediante la regresión logística, y con los observados directamente en los lotes.

El estimador producto-momento de la función de sobrevivencia se ajusta a datos de estudios previos hechos con tomate de exportación. Las ventajas sobre el método de regresión lineal son:

1) No requiere suposición de distribución de la variable dependiente, y 2) La línea recta no representa adecuadamente el comportamiento de las variables durante la vida de anaquel, como se ve en la figura 1.

Cuadro 1. Tiempo de 50% de Sobrevivencia en Anaquel de los Lotes de Tomates en las Diferentes Condiciones de Almacenamiento.

Procedencia	Condición de Almacenamiento	Mediana de Vida de Anaquel (Días)
Campo	20 días de 10°C	24 ± 6
Campo	20°C	21 ± 3
Empacadora	20 días de 10°C	37 ± 7.5
Empacadora	20°C	21 ± 3

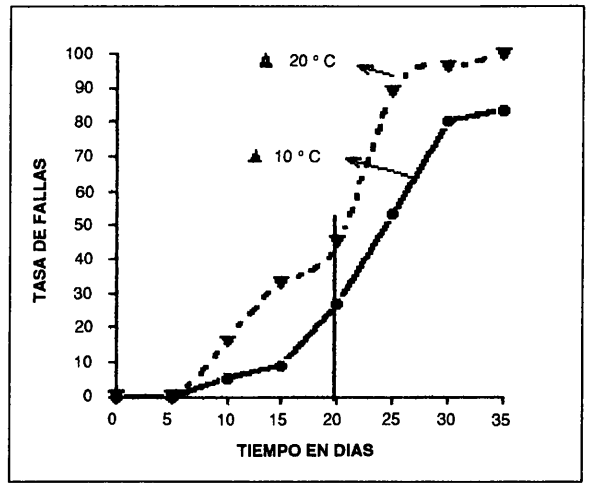


Figura 1. Funciones de Fallas para las Dos Temperaturas de Almacenamiento.

Es importante en estos estudios, verificar el comportamiento monotónico de las variables que se usan para determinar la vida de anaquel. Es aquí donde la regresión logística proporcionar una herramienta para seleccionar las variables que más se relacionen con las pérdidas (fallas) que ocurren en el tiempo. Después de verificar dicho comportamiento en algunas de las características objetivas de calidad, tales como firmeza, color, acidez, grados brix, pH y pérdida de peso, se procedió a realizar el ajuste del modelo de regresión logística propuesto en la ecuación (6). Se encontró que las pérdidas cuantitativas de tomates se relacionan con la acidez, grados Brix, la intensidad de color y la pérdida de peso, (ésta última representa aproximadamente el 75% de la relación total hallada), obteniéndose un pseudo coeficiente de determinación (R^2) de 0.80, (Aldrich y Nelson, 1984). Además, se observó en una prueba de comparación de pendientes, que los coeficientes de regresión son diferentes significativamente ($p< 0.05$) para las dos temperaturas de almacenamiento. El modelo estimado puede interpretarse en términos de probabilidades de fallas (o de sobrevivencia), y nos permite conocer cuáles determinaciones de

calidad son más indicativas de las fallas en el lote de frutos. Como ejemplo se muestra la ecuación logística obtenida para los lotes a 10° C, la cual fué:

$$L_c(p_k) = 45.54 - 1.0019*\text{IntColor} - 1.4395*\text{Perd.Peso}.$$

Para un color de 30, y una pérdida de peso del 10%, la probabilidad estimada de pérdidas del fruto será de 0.70.

Los resultados obtenidos en este estudio sugieren la posibilidad de combinar las mediciones objetivas para construir un Indice Objetivo de Calidad que pueda predecir la vida media de un lote de tomates, así como las pérdidas cuantitativas que pudiera tener el lote. Un análisis de componentes principales proporciona un índice de calidad , el cual explica el 62% de la varianza de las variables incluídas. La evaluación del modelo de la ecuación (8) nos arroja un coeficiente de correlación de 0.87, lo cual indica una buena asociación entre el índice de calidad y las pérdidas de frutos (Figura 2). En este trabajo se presentan algunos métodos estadísticos alternativos para aplicar a estudios de vida de anaquel. Se requieren estudios más amplios de las propiedades de estos métodos para dichos estudios, y una más detallada valoración estadística de sus ventajas y deficiencias.

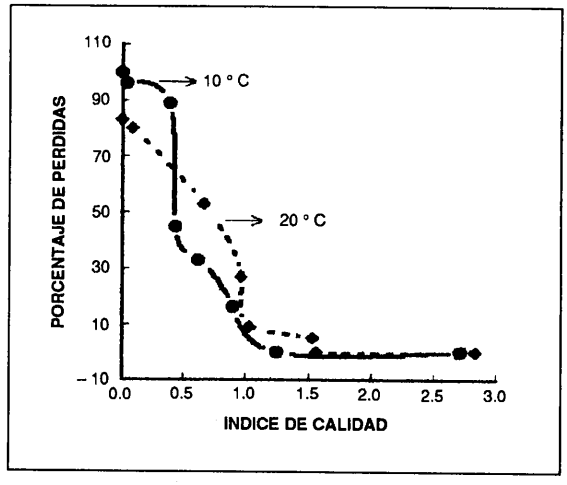


Figura 2. Relación entre el Indice de Calidad Objetiva del Lote y el Porcentaje de Pérdidas que Ocurren Durante su Vida de Anaquel.

REFERENCIAS

Akinbolu, A.M., et al. (1991). Evaluation of Post-Harvest Losses and Quality Changes in Tomatoes in Borno State, Nigeria. *Tropical Science* **31**: 235-242.

Aldrich, J.M y F. D. Nelson. (1984). *Linear Probability, Logit and Probit Models*. Ed. Sage, Beverly Hills, CA, U.S.A.

Brookmeyer, R. y J. Crowley. (1982). A confidence interval for the median survival time. *Biometrics* (**38**):29-41.

CAADES. (1993). Boletín Trimestral Informativo. Confederación de Asociaciones de Agricultores del Estado de Sinaloa, Sinaloa, México.

Charalambous, G. (1992). *The shelf life of foods and beverages. Developments in Food Science*. Elsevier, Londres.

Floros, J.D. (1993). *The Shelf life of fruits and vegetables. En: Charalambous G. Shelf life of foods and beverages*. Elsevier Sc. Publ. Londres.

Gacula, M.C. y J. Singh. (1984). *Statistical Methods in Food and Consumer Research*. Academic Press, Inc., Orlando.

Hamilton, Lawrence C. (1992). *Regression with Graphics*. Brooks/Cole Publishing Co. Pacific Grove. CA

INEGI. (1994). *El Sector Alimentario en Mexico. INEGI*, Aguascalientes, Agcs.

Lawless, J.E. (1982). *Statistical Models and Methods for Lifetime Data*, John Wiley, New York.

Miller,R. G. (1981). *Survival Analysis*. John Wiley & Sons, New York.

SAS. (1992). SAS/STAT User's Guide SAS Institute, Inc. Cary, NC.

USDA. (1992). *Import Regulations:Tomatoes*. Agricultural Marketing Service. 7CFR Chpt IX 696-698.

Caracterización de Mecanismos de Sobredispersión a Través de la Función Generatriz de Probabilidad

BELEM TREJO-VALDIVIA

IIMAS-UNAM, México

1. INTRODUCCIÓN

En muchos estudios en donde la variable de interés es un conteo, es usual suponer una distribución Poisson. Este supuesto produce que la media y la varianza de dicha variable sean iguales. En el análisis estadístico, un problema importante surgirá cuando los datos presenten sobredispersión, esto es, cuando la varianza resulte mayor que la media. Existe un buen número de trabajos reportados en la literatura, (ver, por ejemplo, Baringhaus y Henze 1992; Campbell y Oprian, 1979; Nakamura y Pérez-Abreu, 1993; Rueda, Pérez-Abreu y O'Reilly, 1991) enfocados a desarrollar pruebas para evaluar el ajuste de un modelo Poisson. Algunos de estos procedimientos están basados en el comportamiento de la función generatriz de probabilidad (fgp), o bien, en el comportamiento de la función generatriz de probabilidad empírica (fgpe), dadas respectivamente por:

$$\Phi(t) = E(t^X) \quad \text{y} \quad \Phi_n(t) = \frac{1}{n}\sum_{i=1}^{n} t^{X_i},$$

en donde X es la variable de conteo con función de distribución F sobre 0,1,2,..., y $X_1,\ldots,X_n$ es una muestra aleatoria de F. Para una distribución Poisson con media $\lambda>0$, la fgp es $\Phi(t) = e^{\lambda(t-1)}$, de donde el $\log(\Phi(t)) = \lambda(t-1)$ es una línea recta, hecho que caracteriza esta distribución.

Sin embargo, no solo es importante desarrollar métodos que permitan detectar sobredispersión en los datos, sino que además permitan identificar la fuente que la produce. En este trabajo se presentan 4 diferentes mecanismos que pueden producir datos de conteo sobredispersos con soporte sobre $\{0,1,2,\ldots\}$. En cada una de ellas, el grado de sobredispersión está modelado en forma paramétrica. El objetivo es caracterizar estas 4 situaciones mediante el análisis del comportamiento del logaritmo de la fgp y/o de alguna otra transformación de la misma.

2. ALGUNOS MECANISMOS DE SOBREDISPERSIÓN

2.1. Modelo con componentes aleatorias (Modelo tipo 'frailty'). Estos modelos resultan útiles cuando la heterogeneidad entre los individuos de la población, no se puede explicar a través de la inclusión de covariables que afectan el comportamiento de X. En este modelo, se supone que para cada individuo, la distribución condicional de X dada ξ (la llamada variable 'frailty') es Poisson con media $\lambda\xi$, en donde $\lambda>0$ y ξ es una variable aleatoria no observable que representa el grado de heterogeneidad de dicho individuo, con media finita y con función de distribución G en una familia amplia de distribuciones (que incluye tanto continuas como discretas). La función de densidad de probabilidad de X y la correspondiente fgp estarán dadas por

$$P[X=x]=\frac{\lambda^x}{x!}E_G(\xi^x e^{-\lambda\xi}) \quad \text{y} \quad \Phi_1(t)=E_G(e^{\lambda(t-1)\xi}),$$

en donde $E_G(\cdot)$ es la esperanza con respecto a G.

Una elección muy común para la función G es la distribución Gamma con parámetros $1/\theta$ y $1/\theta$ lo que produce que ξ tenga media 1 y varianza θ. Además, la distribución de X resulta ser una binomial negativa con parámetros $1/\theta$ y $1/(\lambda\theta+1)$, la fgp y su logaritmo serán entonces

$$\Phi_1(t)=(1-\lambda\theta(t-1))^{-1/\theta} \text{ y } \Psi_1(t)=\log(\Phi_1(t))=-\theta^{-1}(\log(1-\theta\lambda\,(t-1))),$$

con $\lambda>0$ y $\theta\geq 0$, de donde la media de X es $\mu(\lambda,\theta)=\lambda$ y la varianza es $\sigma^2(\lambda,\theta)=\lambda(1+\lambda\theta)$. El caso límite de $\theta=0$ corresponde a una variable ξ degenerada en el 1, lo que reproduce una población homogénea Poisson de parámetro λ.

Otras posibilidades para la elección de G son la Gaussiana Inversa, la distribución concentrada en dos puntos (lo que permite modelar una mezcla de dos distribuciones), la distribución Poisson truncada en cero, etc.

2.2. Modelo para el subregistro. En estudios epidemiológicos en donde el interés es estudiar la incidencia de determinadas enfermedades, es común que se presente un subregistro de la categoría '0'. Esto es, algunos hospitales y/o clínicas deciden no reportar la información correspondiente cuando no se registra ningún caso de la enfermedad bajo estudio. Si suponemos que el $0\leq\theta<1$ por cien por ciento de los hospitales y/o clínicas toman dicha acción, la función de densidad de probabilidad de X puede modelarse mediante

$$P[X=x]=\begin{cases}\theta+(1-\theta)e^{-\lambda} & x=0\\ (1-\theta)\dfrac{\lambda^x}{x!}e^{-\lambda} & x=1,2,\ldots\end{cases},$$

la media y la varianza de X serán $\mu(\lambda,\theta)=\lambda(1-\theta)$ y $\sigma^2(\lambda,\theta)=\lambda(1-\theta)(1+\lambda\theta)$ respectivamente. Bajo este modelo, la fgp y su logaritmo son, con $\lambda>0$ y $0\leq\theta<1$,

$$\Phi_2(t)=\theta+(1-\theta)e^{\lambda(t-1)} \quad \text{y} \quad \Psi_2(t)=\log(\Phi_2(t))=\log(\theta+(1-\theta)e^{\lambda(t-1)}).$$

Como en el caso anterior, $\theta=0$ se reduce a la distribución Poisson con media λ.

2.3. Modelo para estudios con problemas de codificación. En algunas situaciones se puede tener mecanismos defectuosos de captación de la información que produzcan que conteos contiguos tiendan a confundirse. Por simplicidad tomemos el caso en el que el '0' y el '1' son mal captados, de tal forma que una proporción θ de 'unos' son codificados como 'ceros'. La función de densidad de probabilidad de X resulta ser

$$P[X=x]=\begin{cases}e^{-\lambda}(1+\theta\lambda) & x=0\\ \lambda e^{-\lambda}(1-\theta) & x=1\\ \dfrac{\lambda^x}{x!}e^{-\lambda} & x=2,3,\ldots\end{cases}$$

con $\lambda > 0$ y $0 \leq \theta \leq 1$. De aquí se tiene que la media X está dada por $\mu(\lambda,\theta) = \lambda(1-\theta e^{-\lambda})$ y su varianza por $\sigma^2(\lambda,\theta) = \lambda(1-\theta e^{-\lambda}) + \lambda^2\theta e^{-\lambda}(2-\theta e^{-\lambda})$. La fgp y su logaritmo son entonces,

$$\Phi_3(t) = e^{\lambda(t-1)}(1-\theta\lambda(t-1)) \quad \text{y} \quad \Psi_3(t) = \log(\Phi_3(t)) = \lambda(t-1) + \log(1-\theta\lambda(t-1))$$

Aquí también se recupera el caso Poisson cuando el parámetro θ toma el valor cero.

2.4. Modelo Poisson Generalizado. Este modelo se presenta solo como una extensión del caso Poisson usual (ver Consul, 1989), está definido por la siguiente familia biparamétrica de funciones de densidad de probabilidad

$$P[X = x] = \frac{\lambda(\lambda+\theta x)^{x-1} e^{-\lambda-\theta x}}{x!}, \quad x = 0,1,2,\ldots$$

para $\lambda > 0$ y $0 \leq \theta < 1$. Aunque la familia completa acepta valores negativos del parámetro θ, aquí nos restringiremos a valores no negativos ya que en aquel caso, la variable de conteo tiene soporte finito. Para esta distribución se tiene media y varianza dadas por $\mu(\lambda,\theta) = \lambda/(1-\theta)$ y $\sigma^2(\lambda,\theta) = \lambda/(1-\theta)^3$ respectivamente. La fdp y su logaritmo en este caso estarán dadas por

$$\Phi_4(t) = e^{\lambda(u-1)} \quad \text{y} \quad \Psi_4(t) = \log(\Phi_4(t)) = \lambda(u-1)$$

con u la solución de la ecuación $u = te^{\theta(u-1)}$. De nuevo, con θ=0 se obtiene la distribución Poisson usual.

3. COMPORTAMIENTO DEL LOGARITMO DE LA FGP

Por simplicidad se tomó el parámetro de escala λ=1 para analizar el comportamiento de Ψ en cada uno de los casos anteriores y el parámetro de sobredispersión θ en [0,1). Al graficar dichas funciones se encontró que no existe una clara diferencia en el comportamiento de las curvas cuando θ es pequeño. El modelo 'frailty' produce funciones convexas, todas por encima de la línea recta, con un solo cruce en t=0. El modelo de subregistro también dá funciones convexas con cruce en cero pero por debajo de la línea para t positivo, alejándose claramente de ella a medida que θ crece. El modelo con codificación errónea muestra funciones cóncavas, de nuevo con único cruce en t=0; para θ pequeño resultan muy similares a las obtenidas en el caso anterior. Para el modelo Poisson generalizado, las funciones son claramente convexas y las curvas se cruzan en dos puntos, t=0 y t= −1.

En una situación real en la que, dada una muestra, se calculará la fgp empírica, y tomando en cuenta los resultados anteriores, no se piensa que con solo un gráfica de su logaritmo, se tenga la suficiente información para identificar la posible causa de la sobredispersión y así poder actualizar el modelo correspondiente.

4. FGP 'ESTANDARIZADA'

Por lo anterior, es deseable el contar con información adicional, quizá a partir de otra función de $\Phi(t)$, que permita resolver el problema de interés. Para estudiar una nueva función de $\Phi(t)$, se consideró entre otras cosas que debía definirse tal que fuera fácilmente

identificable el caso Poisson desde el punto de vista gráfico y que pudiera ser estimada consistentemente a partir de la muestra. Esta transformación será llamada la fgp 'estandarizada' y estará dada por

$$\Phi^*(t) = e^{-\mu(\lambda,\theta)(t-1)}\Phi(t),$$

en donde $\Phi(t)$ es como antes y $\mu(\lambda,\theta) = E(X)$ (que es estimable consistentemente a partir de los datos). Hay que notar que para el caso Poisson $\Phi^*(t)$ se reduce a la función constante igual a 1. Las correspondientes funciones 'estandarizadas' para los cuatro modelos son,

$$\Phi_1^*(t) = e^{-\lambda(t-1)}(1+\theta\lambda\,(t-1))^{-1/\theta}$$

$$\Phi_2^*(t) = e^{-(1-\theta)\lambda(t-1)}(\theta+(1-\theta)e^{\lambda(t-1)})$$

$$\Phi_3^*(t) = e^{-\lambda(1-\theta e^{-\lambda})(t-1)}(e^{\lambda(t-1)} - \theta\lambda e^{\lambda(t-1)}(t-1))$$

$$\Phi_4^*(t) = e^{\lambda(u-1)-\frac{\lambda}{1-\theta}(t-1)} \quad , \quad u = te^{\theta(u-1)}$$

De la figuras correspondientes, se observó que la fgp 'estandarizada' refleja un comportamiento más diferenciable entre los 4 modelos. Las gráficas de los modelos 'frailty' y de subregistro son similares, pero para θ grande las del logaritmo los puede diferenciar. Como el problema de diferenciar las cuatro situaciones se presenta cuando θ es cercano a cero, se hizo un análisis (basado en expansiones de Taylor) del tipo de 'alejamiento' con respecto a la línea constante que tienen estas funciones. Para el modelo 'frailty' se tiene esencialmente una función cuadrática en $(t-1)$ que es más pronunciada a medida que θ crece. Para el caso de subregistro se encontró que la funciones se comportan como exponenciales. El modelo con errores de codificación produce gráficas con término dominante lineal en $(t-1)$. Finalmente, en el caso de la Poisson generalizada se obtienen funciones de tipo cuadráticas, pero en $(u-1)$ con u solución a $u = te^{\theta(u-1)}$

Los resultados anteriores son alentadores ya que se tiene una forma, no solo gráfica sino numérica, de diferenciarlos los modelos considerados. Por lo que, el trabajo a futuro estará enfocado al estudio del comportamiento de la correspondiente versión muestral de la fdg 'estandarizada'.

REFERENCIAS

Baringhaus,L. y Henze,N. (1992). A goodness of fit test for the Poisson distribution based on the empirical generating function. *Stat. Prob. Lett.*, **13**, 269-274.

Campbell,D.B. y Oprian,C.A. (1979). On the Kolmogorov-Smirnov test for the Poisson distribution with unknown mean. *Biometrics*, **21**, 17-24.

Consul,P.C. (1989). *Generalized Poisson Distributions.* Marcel Dekker, Inc.

Nakamura,M. y Pérez-Abreu,V. (1993).Use of an empirical probability generating function for testing a Poisson model. *Cand. J. Statistics.*, **21**, 149-156.

Rueda,R., Pérez-Abreu,V. y O'Reilly,F. (1991). Goodness of fit for the Poisson distribution based on the probability generating function. *Comm. Stat. Theory Methods*, **A20**, 3093-3110.

Propiedades y Aplicaciones de Una Medida de Redundancia de la Información: el Número Equivalente

JAVIER TREJOS Z.

Univ. de Costa Rica, Costa Rica

1. INTRODUCCIÓN

El número equivalente fue estudiado por Der Mégréditchian (1979, 1988ab) en un contexto probabilístico para calcular el número de estaciones independientes en la previsión meteorológica. Para una tabla de datos X definida por p variables cuantitativas, si se introduce una métrica M en el espacio de los individuos $E=\mathbb{R}^p$, podemos adaptar la definición del número equivalente (*Neq*), al contexto euclídeo, de la manera siguiente:

Definición 1 El número equivalente asociado a la matriz X, respecto a la métrica M, es: $Neq(X,M)=\dfrac{(traza\ VM)^2}{traza(VM)^2}$ donde V es la matriz de varianzas-covarianzas de las p variables x^j.

Se tiene $Neq(X,M) = \dfrac{\left(\sum_{j=1}^{rang\ x} \lambda_j\right)^2}{\sum_{j=1}^{p} \lambda_j^2}$ donde λ_j es el j-ésimo valor propio no nulo de VM.

La siguiente propiedad, debida a Troupe (1994), precisa el sentido que damos en este contexto al *Neq* como medida de la cantidad de información no redundante aportada por un conjunto de variables cuantitativas (respecto a M).

Proposición 1
a) $Neq(X,M) \geq 1$, y $Neq(X,M) = 1$ si y sólo si hay solamente un valor propio no nulo de VM.
b) Si VM tiene al menos dos valores propios no nulos distintos, entonces $Neq(X,M) <$ rang X.
c) $Neq(M)$ = rang X si y sólo si todos los valores propios no nulos de VM son iguales.

2. CASO DE LA MÉTRICA DIAGONAL DE LAS INVERSAS DE LAS VARIANZAS

A continuación estudiamos (Trejos, 1994; 1995) el comportamiento del número equivalente en el caso en que $M = D_{1/\sigma^2}$, la diagonal de las inversas de las varianzas. Recuérdese que este es el caso usual en Análisis en Componentes Principales cuando las variables están centradas y estandarizadas. Tenemos $M = D_{1/\sigma^2}$ = diag(1/var x^j), donde var x^j es la varianza de la variable x^j. Supondremos que las variables están centradas.

Proposición 2 Si $M = D_{1/\sigma^2}$ entonces $Neq(X, D_{1/\sigma^2}) = p^2/\sum_{j=1}^{p}\sum_{k=1}^{p} \rho^2\left(x^j, x^k\right)$, donde ρ es el coeficiente de correlación lineal.

Para $M = D_{1/\sigma^2}$, la Proposición 2 permite reducir la complejidad del cálculo del *Neq*: en efecto, como el cálculo de cada correlación es en $O(n)$, la suma de los cuadrados de las p^2

correlaciones (y por consiguiente el cálculo del *Neq*) es de complejidad $O(np^2)$, mientras que con la Definición 1, la complejidad del cálculo del *Neq* es en al menos $O(np^3)$.

Corolario 3. Sea $M = D_{1/\sigma^2}$. Si se tienen m clases de variables $K_1,\ldots,K_m$ de mismo cardinal s y tales que, $\forall(x^j, x^{j'}) \in K_\ell \times K_{\ell'}$, $\rho^2(x^j, x^{j'}) = \delta_{\ell\ell'}$, entonces $Neq(M) = m$.

En presencia de grupos de variables con correlaciones intra elevadas y correlaciones inter bajas, el *Neq* tendrá un valor vecino al número de grupos: este resultado es una ilustración suplementaria del poder de medida de redundancia de la información que hemos mencionado que posee el *Neq*.

3. APLICACIONES

En sus trabajos originales, G. Der Mégréditchian estudió la aplicación del *Neq* para determinar el número de estaciones de observación meteorológica necesarias para tener toda la información pertinente, de manera tal que no se repita la información aportada por dos estaciones diferentes. Como hemos dicho, estos trabajos estaban enmarcados en un contexto probabilístico. Nosotros hemos encontrado, a partir de los desarrollos de la sección anterior, algunas aplicaciones que pueden ser interesantes en el Análisis Multivariado de Datos según la Escuela Francesa, es decir, sin asumir distribuciones de probabilidad teóricas a priori en los datos.

3.1 Análisis en Componentes Principales: Determinación del Número de Factores

El análisis en componentes principales (A.C.P.) trata de encontrar un conjunto de q variables sintéticas C^j a partir de una tabla de datos descrita por p variables cuantitativas $x^1,\ldots x^p$, tales que las C^j sean no correlacionadas y con inercia máxima, en el sentido que la proyección de la nube de puntos-individuos en $\mathbb{R}^p$ sobre el espacio generado por las C^j tenga inercia máxima. En el caso usual, las variables están centradas y se estandarizan, por lo que la métrica en $\mathbb{R}^p$ es $M = D_{1/\sigma^2}$. Es sabido que la solución de este problema se obtiene a partir de la diagonalización de la matriz VM, producto de la matriz V de varianzas-covarianzas y la métrica M sobre $\mathbb{R}^p$.

Uno de los problemas ligados a la práctica del A.C.P. es el de la determinación del número q de componentes principales (es claro que $q < p$ para que tenga sentido hacer el análisis). Diversos autores (ver, por ejemplo, Cailliez y Pagès, 1976; Escofier y Pagès, 1988; Jaumbu, 1989) han propuesto algunos criterios empíricos, tales como: tomar q tal que la inercia explicada por $C^1,\ldots,C^q$ sobrepase un umbral (porcentual, por ejemplo 70% u 80%) de la inercia total de la nube de puntos-individuos, tomar q tal que el diagrama de los valores propios de VM, ordenados en orden decreciente, muestre el punto donde el decrecimiento se aprecie como estable (este método es conocido como el método del "codo"), en el caso usual de la métrica D_{1/σ^2}, tomar q como el número de valores propios de VM mayores que 1, tomar tantas las componentes principales C^j que sean interpretables, en el sentido que haya por lo menos un individuo tal que el coseno cuadrado entre su vector en $\mathbb{R}^p$ y su proyección sobre C^j sea mayor que 0.5, o bien cuando la correlación entre al menos una variable original y C^j es 0.7. Ninguno de estos criterios es un criterio absoluto, antes bien se pregoniza la utilización conjunta de varios de ellos para decidir lo mejor posible la elección de q, y se llega incluso a afirmar que esta elección depende en mucho de la experiencia del

analista. ¿Puede entonces darse una herramienta confiable que pueda servir al usuario, lego en la materia, para la determinacion de *q*?.

Nosotros pensamos que el número equivalente puede ayudar a responder a esta cuestión. En efecto, por tratarse de una medida de la información independiente contenida en una tabla de datos, es posible que ayude a decidir cuántos factores guardar de un A.C.P.

Con el fin de estudiar esta posibilidad, calculamos el *Neq* sobre varias tablas de datos y comparamos el resultado con los criterios 1 y 3 mencionados arriba. Los resultados para varias tablas de datos se dan en la Tabla 1. Los datos de las tablas correspondientes se pueden solicitar al autor.

TABLA 1

Comparación entre el número equivalente (Neq) y el número r de valores propios mayores que 1, para varias tablas de datos de dimensiones n (número de individuos) por p (número de variables).

Tabla de datos	*n*	*p*	*Neq*	*r*	Valores propios	Inercia
Notas escolares F	9	5	2.36	2	$\lambda_1 = 2.87$	56%
					$\lambda_2 = 1.13$	80%
					$\lambda_3 = 0.98$	99%
Notas escolares CR	10	5	2.24	2	$\lambda_1 = 2.89$	58%
					$\lambda_2 = 1.62$	90%
Peces de Amiard	23	16	3.43	3	$\lambda_1 = 7.52$	46%
					$\lambda_2 = 3.69$	70%
					$\lambda_3 = 1.52$	80%
					$\lambda_4 = 0.94$	86%
Sociomatriz de Thomas	24	24	7.42	7	$\lambda_1 = 5.25$	22%
					$\lambda_2 = 4.72$	42%
					$\lambda_3 = 3.92$	58%
					$\lambda_8 = 0.84$	87%
Iris de Fisher	150	4	1.70	1	$\lambda_1 = 2.50$	62%
					$\lambda_2 = 0.91$	85%
Proteínas	25	9	3.80	3	$\lambda_1 = 4.00$	44%
					$\lambda_2 = 1.63$	63%
					$\lambda_3 = 1.12$	75%
					$\lambda_4 = 0.95$	85%
Pintores	24	4	2.52	1	$\lambda_1 = 2.27$	57%
					$\lambda_2 = 0.98$	81%

Puede verse en la tabla que el *Neq* tiende a ser superior al número de valores propios mayores que uno. Por lo tanto, es posible que el número equivalente tienda a sobreestimar el número de factores importantes de una A.C.P. Esta observación puede ser de utilidad para el usuario nuevo en el campo, que puede tener cierta aprehensión a dejar de lado información que puede ser útil para su estudio. Por ello, el número equivalente podría servirle como número de componentes principales suficientes para tomar en cuenta.

3.2 Particionamiento: Determinación del Número de Clases

En clasificación automática, los métodos de particionamiento tratan de obtener una partición de un conjunto de objetos sobre los que se han observado una serie de variables, de manera tal que los elementos de una misma clase sean lo más parecidos posible, y los elementos de clases distintas sean bastante diferentes (ver Cailliez y Pagès, 1976; Diday, Lemaire, Pauget y Testv, 1982; Jambu, 1989). Usualmente, se aplican métodos que fijan *a priori* el número de clases, tales como los métodos de nubes dinámicas, de las *k*-medias, de transferencias, etc., al contrario de métodos como Isodata que estiman el número de clases pero con base en un gran número de parámetros difíciles de controlar para un usuario poco experimentado. Sería por lo tanto útil contar con un método que estime el número de clases antes de implementar la metodología de particionamiento.

Para abordar esta cuestión, hemos pensado en usar una adaptación del número equivalente que presentamos anteriormente. En efecto, los métodos de particionamiento buscan tipologías de los *individuos*, mientras que los métodos factoriales hacen tipologías de las *variables*. Las medidas del "parecido" entre individuos generalmente están basadas en criterios de *disimilitud* o *distancia*: entre menor sea el índice más parecidos son los objetos, mientras que las medidas del "parecido" entre variables están basadas en criterios de asociación estadística, tales como la correlación lineal: entre mayor sea el índice de asociación más parecido es el comportamiento de las variables.

Sea Ω un conjunto de n individuos, sobre los que se dispone de una medida de disimilitud d: $\Omega \times \Omega \rightarrow I\!R^+$ (d puede ser una distancia). Sea d^* el máximo valor que alcanza d, entonces se define la similitud s:

$$s(i,j) = \frac{(d^*)^2 - d^2(i,j)}{(d^*)^2}.$$

Obsérvese que así el valor máximo de $s(i,j)$ es 1, y corresponde al caso en que $i = j$. Se denota S la matriz de similitudes calculadas sobre los elementos de Ω.

Definición 2. Dado un conjunto Ω con una medida de similitud s: $\Omega \times \Omega \rightarrow [0,1]$, se define el número equivalente $Neq(\Omega,S)$ por:

$$Neq(\Omega, S) = \frac{(traza\, S)^2}{traza(S^2)} = n^2 / \sum_{i=1}^{n} \sum_{j=1}^{n} s^2(i,j)$$

Adaptando la Proposición 3 a la definición anterior, se tiene el resultado enunciado en la Proposición 4.

Proposición 4. Si existe una partición $C_1,\ldots, C_k$ de Ω en k clases de mismo cardinal π, tales que $\forall (i,j) \in C_\ell \times C_{\ell'}$, $s(i,j) = \delta_{\ell\ell'}$, entonces $Neq(\Omega, s) = k$.

El resultado anterior sugiere que, si se tienen k clases bastante homogéneas y de cardinal similar, el número equivalente puede dar una aproximación de ese número de clases. En caso que las clases no tengan mismo cardinal, entonces $Neq(\Omega, s) = \left(\sum_{\ell=1}^{k} \pi_\ell\right)^2 / \sum_{\ell=1}^{k} \pi_\ell^2$, donde π_ℓ es el cardinal de la clase C_ℓ.

Hemos medido el número equivalente definido sobre una matriz de similitudes, para algunas de las tablas de datos estudiadas anteriormente. Estos resultados se dan en la Tabla

2. Para algunas de las tablas mostradas, el número equivalente da una idea del número de clases que podrían tomarse en una clasificación. Recuérdese que los árboles de clasificación jerárquica construidos ascendentemente, normalmente dan buenas agrupaciones en las partes bajas del árbol pero la calidad de la clasificación disminuye conforme se asciende en la construcción. Contrariamente, los árboles construidos descendentemente dan una mejor calidad en las partes superiores la calidad disminuye en las partes inferiores. Estas comparaciones deben ser ampliadas, con diversos métodos de clasificación, así como con diversos criterios de estimación del número de clases. Murillo (1996) propone un índice para "cortar" un árbol de clasificación jerárquica, basado en conjuntos difusos. Además se hacen comparaciones entre 8 índices para estimar el número de clases, entre ellos el que aquí proponemos basado en el número equivalente. Una próxima publicación dará cuenta de estas comparaciones.

TABLA 2

Comparación entre el número equivalente (Neq) y el número de clases sugeridas por el árbol de clasificación jerárquica, para varias tablas de datos de dimensiones n (número de individuos) por p (número de variables)

Tabla de Datos	*n*	*p*	*Neq*
Notas escolares F	9	5	3.00
Notas escolares CR	10	5	3.21
Peces de Amiard	23	16	2.24
Sociomatriz de Thomas	24	24	4.68
Iris de Fisher	150	4	2.14
Proteinas	25	9	2.76
Pintores	24	4	2.92

4. CONCLUSIONES Y PERSPECTIVAS

El número equivalente tiene propiedades interesantes que pueden explotarse en análisis de datos. Las aplicaciones mostradas han ayudado a abordar problemas abiertos que tiene el análisis de datos, pudiéndose aún profundizar en algunas propiedades teóricas que podrían ayudar a esclarecer mejor los problemas planteados. En otras publicaciones hemos estudiado el uso del número equivalente en la determinación del número de componentes de conjunciones para la generación de reglas de producción (Schektman, Trejos y Troupé, 1994; Trejos, 1994, 1995; Troupé, 1994).

Sin embargo, las investigaciones deben continuarse para hacer comparaciones con métodos y criterios existentes para la determinación del número de factores en un análisis factorial o el número de clases en clasificación automática.

También debe tratar de generalizarse al caso en que se tenga una tabla con variables cualitativas, o cuando se tiene una tabla de contingencia. Este último caso sería particularmente útil para estimar el número de componentes en un Análisis de Correspondencias.

Por otro lado, es posible que el número equivalente encuentre aplicaciones en otros campos del análisis de datos, como en regresión y en discriminación. En efecto, uno podría pensar en abordar el problema del número de variables explicativas necesarias para un problema de regresión (no necesariamente lineal, y sin suponer ninguna distribución de probabilidad, ni en las variables activas ni en los residuos); así mismo, se podría pensar en

que el número equivalente puede ser útil para la determinación del número de variables explicativas significativas en discriminación (de nuevo sin hacer hipótesis de probabilidad). Por otra parte, el conocido problema de la determinación del número de neuronas en una red neuronal con una capa escondida (para la aplicación del método de retropropagación del gradiente), podría encontrar alguna luz desde el punto de vista del número equivalente, adaptando su definición al uso de los pesos sinápticos entre las neuronas. Estas cuestiones serán estudiadas en futuras investigaciones dentro del Programa de Investigación en Modelos y Análisis de Datos de la Universidad de Costa Rica.

REFERENCIAS

Cailliez, F.; Pages J.P. (1976) *Introduction à l'Analyse des Données*. Société de Mathématiques Appliquées et de Sciences Humaines, Paris.

Der Mégreditchian, G. (1979) L'optimisation des réseaux d'observation des champs météorologiques, *La Météorologie*, **6(17):** 51-66.

Der Mégreditchian, G. (1988a) Análisis espacial de los campos meteorológicos y aplicación a la optimización de redes de medida. En: Memorias IV Simposio Métodos Matemáticos Aplicados a las Ciencias, B. Montero & J. Poltronieri (eds.), 1984, *Editorial de la Universidad de Costa Rica*, pp. 1-34.

Der Mégreditchian, G. (1988b) Condensación óptima de la información meteorológica por medio del análisis en componentes principales. En: Memorias IV Simposio Métodos Matemáticos Aplicados a las Ciencias, B. Montero & J. Poltronieri (eds.), 1984, *Editorial de la Universidad de Costa Rica*, pp. 35-61.

Diday, E.; Lemaire, J.; Pouget, J.; Testu, F. (1982) *Eléments d'Analyse de Données*. Dunod, Paris.

Escofier, B.; Pagès, J. (1988) *Analyses Factorielles Simples et Multiples*: Objectifs, Méthodes et Interprétation. Dunod, Paris.

Jambu, M. (1989) *Exploration Informatique et Statistique des Donneés*. Dunod, Paris.

Murillo, A. (1996) *Proposición de un índice para la interpretación de árboles de clasificación basado en conjuntos difusos*. Tesis para optar al grado de Magister Scientiæ en Computación, Instituto Tecnológico de Costa Rica, Cartago.

Schektman, Y.; Trejos, J.; Troupé, M. (1994) Generación de reglas estadísticas a partir de grandes bases de datos, *Revista de Matemática: Teoría y Aplicaciones*, **1(1):** 87-100.

Trejos, J. (1994) *Contribution à l'Acquisition Automatique de Connaissances à Partir de Donneés Qualitatives*. Thèse de doctorat, Université Paul Sabatier, Toulouse.

Trejos, J. (1995) El número equivalente como medida de la información en análisis de datos, *Revista de Matemática: Teoría y Aplicaciones*, **2(2):** 75-86.

Troupé, M. (1994) *Contribution à la Régression Multiple Multidimensionnelle et à la Génération de Règles Incertaines*. Thèse de doctorat, Université Paul Sabatier, Toulouse.

Modelos Antedependientes de Primer Orden: Estimación Máximo Verosímil y Aspectos Computacionales

DALE L. ZIMMERMAN y VICENTE NÚÑEZ A.

The University of Iowa, U.S.A. *Univ. del Pais Vasco, España*

1. INTRODUCCIÓN

Los modelos antedependientes de primer orden para estructuras de covarianza en datos longitudinales, pueden ser útiles cuando se tiene correlación en serie, pero las asunciones de un modelo estacionario autorregresivo no son válidas. En este trabajo describimos la flexibilidad de estos modelos e indicamos como la estructura antedependiente de una matriz de covarianzas puede utilizarse para reducir los costes computacionales necesarios para la estimación máximo verosímil de los parámetros en el modelo.

Consideremos una situación en la que se tienen medidas univariantes continuas a lo largo del tiempo para cada uno de los n sujetos. Estas mediciones no deben ser necesariamente efectuadas en lapsos de tiempo que se encuentren igualmente espaciados, ni deben ser las mismas para cada sujeto. Datos de este tipo se denominan datos longitudinales o medidas repetidas. Varios autores han analizado casos particulares de este tipo de datos basándose en el modelo lineal general:

$$Y_k = X_k\beta + e_k, \quad k = 1,2,\ldots,n, \tag{1}$$

donde Y_k es el vector de respuestas de p_k x 1 para el sujeto k; X_k es una matriz de diseño de p_k x q, de rango q para el sujeto k; los e_k's son vectores aleatorios independientes cuya distribución es normal multivariante con media 0 y matriz de covarianzas $\Sigma_k = \Sigma_k(\theta)$; y tanto β como θ son vectores de parámetros desconocidos.

Los análisis de datos longitudinales que se han basado en el modelo (1) difieren principalmente en lo que respecta a la estructura de covarianzas, debido a la dependencia de Σ_k en θ. Tres tipos de modelos de covarianza han recibido la mayor atención: los modelos multivariantes, los modelos con variación intra-grupos (o con efectos aleatorios), y los modelos autorregresivos estacionarios. Ware (1985) revisa estos tipos de modelos y analiza sus ventajas y desventajas. Estos modelos o variaciones menores de ellos también han sido considerados por Jennrich y Schluchter (1986), Lee (1988), Diggle (1988), Schluchter (1988), Jones (1990), Jones y Boadi-Boateng (1991), y Muñoz, Carey, Schouten, Segal y Rosner (1992).

Un cuarto modelo para la estructura de covarianza de (1) que ha recibido una menor atención es el modelo antedependiente, introducido inicialmente por Gabriel (1962), también considerado por Byrne y Arnold (1987), y comentado brevemente por Jones (1993). Una sucesión de variables W_1, W_2,..., W_m, cuya distribución conjunta es normal multivariante, se dice que es antedependiente de orden s si W_i y W_{i+k+1} condicionados en W_{i+1}, W_{i+2},..., W_{i+k} son independientes, para toda i y $k \geq s$. Los modelos antedependientes, al igual que los modelos estacionarios autorregresivos, son modelos para correlación en serie pero son más generales en que estos modelos no requieren que todas las varianzas sean iguales o que las

correlaciones entre todos los pares equidistantes en la escala temporal sean iguales. Nos centraremos en el caso de los modelos antedependientes de primer orden, a los que llamamos AD(1) y que generalizan el conocido modelo autorregresivo estacionario de primer orden.

En general, el coste de usar el modelo AD(1) más general, en lugar de usar el modelo AR(1) consiste en un incremento de las operaciones computacionales a realizar para la estimación de los parámetros del modelo. Como describiremos en este trabajo, sin embargo, la cantidad de estas operaciones puede ser reducida de manera considerable, especialmente en el caso de modelos AD(1) altamente estructurados o parametrizados. Describiremos varios de estos modelos antedependientes, y analizaremos diversos aspectos computacionales de su estimación usando la teoría de verosimilitud. Finalmente llevaremos a cabo un estudio de Monte Carlo para estudiar las ventajas computacionales de los metodos propuestos.

2. MODELO ANTEDEPENDIENTE DE PRIMER ORDEN

Supongamos que tenemos un modelo antedependiente de primer orden, al que llamaremos en adelante modelo AD(1), para los elementos de cada vector de errores $\varepsilon_1,...,\varepsilon_n$. Para un sujeto arbitrario k, consideremos la matriz de covarianzas $\Sigma_k(\theta)$ de ε_k, y por simplicidad omitamos la dependencia con respecto a k y a θ. Sean $0 < t_1 < t_2 <...< t_p$ los tiempos a los cuales se efectuaron las observaciones en este sujeto, y sea Y_i la respuesta en el tiempo t_i. Sean además, $\sigma_i^2 = \mathrm{Var}(Y_i)$ y $\sigma_i\sigma_j\rho_{ij} = \mathrm{Cov}(Y_i, Y_j)$. La definición del modelo AD(1) es equivalente a (Feller, 1966; Secc. III.8)

$$\rho_{ij} = \rho_{im}\rho_{mj}, \quad i < m < j,$$

expresión equivalente a la condición

$$\rho_{ij} = \prod_{l=i}^{j-1} \rho_{l,l+1}, \quad i+1 < j.$$

Entonces, si $\rho_i = \rho_{i,i+1}$, para un modelo AD(1), podremos escribir Σ como

$$\begin{bmatrix} \sigma_1^2 & \sigma_1\sigma_2\rho_1 & \sigma_1\sigma_3\rho_1\rho_2 & \sigma_1\sigma_4\rho_1\rho_2\rho_3 & \cdots & \sigma_1\sigma_p\prod_{i=1}^{p-1}\rho_i \\ & \sigma_2^2 & \sigma_2\sigma_3\rho_2 & \sigma_2\sigma_4\rho_2\rho_3 & \cdots & \sigma_2\sigma_p\prod_{i=2}^{p-1}\rho_i \\ & & \sigma_3^2 & \sigma_3\sigma_4\rho_3 & \cdots & \cdot \\ & & & \cdot & \cdot & \cdot \\ & \text{simet.} & & & \sigma_{p-1}^2 & \sigma_{p-1}\sigma_p\rho_{p-1} \\ & & & & & \sigma_p^2 \end{bmatrix} \tag{2}$$

Las restricciones que un modelo AD(1) impone en las varianzas es que sean positivas, y en las correlaciones que $-1 < \rho_i < 1$, para $i = 1,..., p-1$. Además, el modelo especifica que para cada tiempo t_i, la correlación entre la medición efectuada en el tiempo t_i y las mediciones efectuadas en tiempos posteriores sea una función monótona decreciente del espacio temporal, pero esta función no tiene que ser la misma para todo i. La matriz de

covarianzas de un modelo AD(1) está completamente determinada por los $(2p - 1)$ elementos sobre la diagonal principal y la primera superdiagonal, o equivalentemente por los parámetros $\sigma_1^2,\ldots,\sigma_p^2,\rho_1,\ldots,\rho_{p-1}$. Es decir, hay menos parámetros en un modelo AD(1) que en el modelo multivariante que tiene $p(p + 1)/2$ parámetros, pero hay más parámetros que en un modelo autorregresivo estacionario de primer orden, que solamente tiene dos parámetros.

Una de las especificaciones flexibles de los modelos AD(1) está dada por

$$\rho_i = \rho^{f(t_{i+1},\lambda)-f(t_i,\lambda)}, \tag{3}$$

$$\sigma_i^2 = \sigma^2 g(t_i;\psi),$$

donde $0 < \rho < 1$, $\sigma^2 > 0$, y λ y ψ son vectores de parámetros. En esta familia de modelos AD(1) las correlaciones están restringidas a ser positivas, una característica típica de datos longitudinales reales. Varias posibilidades para elegir f y g incluyen modelos lineales simples (lineales en los elementos de λ y ψ) y modelos de potencias. El caso especial en que λ es un escalar,

$$f(t;\lambda) = \begin{cases} (t^\lambda - 1)/\lambda & \lambda \neq 0 \\ \log(t) & \lambda = 0 \end{cases}, \tag{4}$$

y $g(t; \psi) \equiv 1$, con espacio de parámetros $\{(\rho, \lambda, \sigma^2): 0 < \rho < 1, -\infty < \lambda < \infty, \sigma^2 >\}$, fue utilizado por Núñez Antón y Woodworth (1994) en un estudio sobre la eficacia de implantes auditivos. Este modelo, en el cual la familia de transformaciones de Box-Cox fue aplicada a la escala temporal, se asumió para explicar el hecho de que los sujetos "aprenden" en el tiempo, con el resultado de que respuestas equidistantes en tiempo estaban más altamente correlacionadas a medida que el estudio avanzaba ($\lambda < 1$) (menos correlacionadas si $\lambda > 1$, y el análogo del proceso estacionario autorregresivo de primer orden correspondería a $\lambda = 1$). También puede especificarse que la función g sea un modelo de potencias, útil cuando se espera que la varianza de las respuestas sea una función monótona creciente (o decreciente) del tiempo, como por ejemplo en los casos típicos de datos de crecimiento.

3. ASPECTOS COMPUTACIONALES DE LA ESTIMACIÓN MÁXIMO VEROSÍMIL

Bajo el modelo (1), tenemos que el logaritmo de la función de verosimilitud, aparte de una constante aditiva, es

$$L(\beta,\theta;Y_1,\ldots,Y_n) = -\frac{1}{2}\sum_{i=1}^{n} \log|\Sigma_k(\theta)| - \frac{1}{2}\sum_{i=1}^{n} (Y_k - X_k\beta)'\Sigma_k^{-1}(\theta)(Y_k - X_k\beta),$$

o equivalentemente, $\hat{\theta}$ es cualquier valor de θ que maximice

$$L^*(\theta;Y_1,\ldots,Y_n) = -\frac{1}{2}\sum_{k=1}^{n} \log|\Sigma_k(\theta)| - \frac{1}{2}\sum_{k=1}^{n} Y_k'\Sigma_k^{-1}(\theta)Y_k + \frac{1}{2}\hat{\beta}'(\theta)\left[\sum_{k=1}^{n} X_k'\Sigma_k^{-1}(\theta)Y_k\right],$$

donde $\hat{\beta}(\theta) = \left[\Sigma_{k=1}^{n} X_k' \Sigma_k^{-1}(\theta) X_k\right]^{-1} \left[\Sigma_{k=1}^{n} X_k' \Sigma_k^{-1}(\theta) Y_k\right]$. Un estimador de máxima verosimilitud restringido (REML) de θ (Diggle, 1988; Muñoz et al., 1992; Núñez Antón y Woodworth, 1994) es cualquier valor $\tilde{\theta}$ que maximice

$$L^{**}(\theta; Y_1, \ldots, Y_n) = L^{*}(\theta; Y_1, \ldots, Y_n) - \frac{1}{2} \log \left| \sum_{k=1}^{n} X_k' \Sigma_k^{-1}(\theta) X_k \right|.$$

Consideremos la estimación de los parámetros en $\Sigma_k(\theta)$ cuando es AD(1) no estructurada, para toda $k = 1,\ldots,n$. En el caso de un modelo AD(1) completamente no estructurado puede haber problemas de identificación, ya que habrá más parámetros que estimar que observaciones disponibles. Si los tiempos a los cuales se efectuan las mediciones son idénticos para todos los sujetos, aunque no necesariamente igualmente espaciados, entonces $\Sigma_1 = \Sigma_2 = \ldots = \Sigma_k \equiv \Sigma$ y si $n \geq q+2$, existe una expresión explícita para los estimadores máximo verosímiles de Σ (Byrne y Arnold, 1983). Por contra, si dichos tiempos no son los mismos para todos los sujetos, o se desea ajustar un modelo AD(1) más estructurado, tal y como los descritos anteriormente, expresiones explícitas para los estimadores máximo verosímiles (o REML) generalmente no existen y los estimadores de θ deben obtenerse maximizando numericamente L* o L**. Este proceso de maximización es complejo, desde el punto de vista computacional, particularmente cuando el número de mediciones por sujeto es elevado, ya que será necesario obtener los determinantes e inversas de matrices de gran dimensión.

Estas operaciones pueden abreviarse enormemente calculando dicho determinante e inversa sólo una vez para cada patrón de tiempos de medición en el estudio. Sin embargo, el ahorro será aún mayor utilizando las expresiones explícitas existentes para el cálculo del determinante e inversa de una matriz AD(1) definida positiva (casos especiales de Barret, 1979; o Byrne y Arnold, 1983). Además es un hecho conocido que la inversa de una matriz de covarianzas AD(1) no singular es tridiagonal (Gabriel, 1962). Las fórmulas de Barret para $|\Sigma|$ y $\Sigma^{-1} = (\sigma^{ij})$, para una matriz Σ definida positiva, antedependiente no estructurada de $p \times p$ son:

$$|\Sigma| = \left(\prod_{i=1}^{p} \sigma_i^2\right)\left(\prod_{i=1}^{p-1}\left(1-\rho_i^2\right)\right), \tag{5}$$

$$\sigma^{ij} = \begin{cases} \left\{\sigma_1^2\left(1-\rho_1^2\right)\right\}^{-1} & i = j = 1 \\ \left\{\sigma_p^2\left(1-\rho_{p-1}^2\right)\right\}^{-1} & i = j = p \\ \left(1-\rho_{i-1}^2\rho_i^2\right)\left\{\sigma_i^2\left(1-\rho_{i-1}^2\right)\left(1-\rho_i^2\right)\right\}^{-1} & i = j \neq 1, p \\ -\rho_i\left\{\sigma_i\sigma_j\left(1-\rho_i^2\right)\right\}^{-1} & |i-j| = 1 \\ 0 & |i-j| > 1 \end{cases} \tag{6}$$

4. SIMULACIÓN

Para investigar cuánto tiempo se puede ahorrar mediante el uso de las fórmulas anteriores, hemos realizado un pequeño ejemplo computacional, consistente en calcular tanto la inversa como el determinante de matrices de covarianza AD(1) de N x N utilizando cada uno de los métodos:

Método 1: usando las fórmulas anteriores

Método 2: usando las subrutinas de IMSL: DLINDS (para la inversa), y DLFTDS y DLFDDS (para el determinante)

Las subrutinas de IMSL solamente utilizan el hecho de que la matriz sea real, simétrica y definida positiva. El estudio se llevo a cabo en un supercomputador Convex 3840, utilizando programas escritos en FORTRAN además de opciones de optimización y vectorización. Todos los cálculos se realizaron en doble precisión. Los resultados se muestran en la siguiente tabla, que indica los tiempos de CPU en segundos utilizado por cada método para un rango de valores de N

N	Método 1	Método 2
10	0.00176	0.00411
50	0.00763	0.02837
100	0.02574	0.10930
500	0.60614	4.90808
1000	2.81839	31.17996
5000	114.86062	4050.86823

Estos resultados indican que a medida que N aumenta de 10 (correspondiendo a un número pequeño de mediciones por sujeto) a 5000 (que podría verse como un número poco real; sin embargo, veáse Wegman, 1994), la reducción relativa del tiempo de CPU utilizado por el Método 1 se incrementa de un 57% a un 97%.

5. CONCLUSIONES

Hemos mostrado cómo se puede reducir significativamente los costes computacionales de estimación máximo verosímil en modelos antedependientes. Esperamos que estos resultados sean útiles para los investigadores que deseen ajustar modelos AD(1) (particularmente los altamente estructurados) a sus datos longitudinales, pero que no los utilizan debido a las dificultades computacionales de los mismos. Como punto concluyente mencionaremos que los modelos AD(1) tienen una característica que algunas veces puede no ser muy real, especialmente cuando la obtención de las mediciones implica un submuestreo, y es que las correlaciones entre observaciones en el mismo sujeto tienden a uno a medida que su espaciamiento se hace menor. Esto se soluciona con la inclusión de errores independientes de medición, cada uno con varianza τ_k^2, que serán añadidos al modelo AD(1) inicial, y que permiten aprovechar las mismas ecuaciones para el cálculo de inversas y determinantes.

REFERENCIAS

Barret, W.W. (1979). A Theorem on Inverses of Tridiagonal Inverses. *Linear Algebra Applications*, **27**, 211-217.

Byrne, P.J. y Arnold, S.F. (1983). Inference about Multivariate Means for a Nonstationary Autoregressive Model. *Journal of the American Statistical Association*, **78**, 850-855.

Diggle, P.J. (1988). An Approach to the Analysis of Repeated Measurements. *Biometrics*, **44**, 959-971.

Feller, W. (1966). *An Introduction to Probability Theory and its Applications* Volume II. Chichester: Wiley.

Gabriel, K.R. (1962). Ante-dependence Analysis of an Ordered Set of Variables. *Annals of Mathematical Statistics*, **33**, 201-212.

Jennrich, R.L. y Schluchter, M.D. (1986). Unbalanced Repeated-Measures Models with Structured Covariance Matrices. *Biometrics*, **42**, 805-820.

Jones, R.H. (1990). Serial Correlation or Random Subject Effects? Communications in Statistics B, *Simulation and Computation*, **19(3)**, 1105-1123.

Jones, R.H. (1993). *Longitudinal Data with Serial Correlation: A State Space Approach*, London: Chapman and Hall.

Jones, R.H. y Boadi-Boateng, F. (1991). Unequally Spaced Longitudinal Data with AR(1) Serial Correlation. *Biometrics*, **47**, 161-175.

Kenward, M.G. (1987). A Method for Comparing Profiles of Repeated Measurements. *Applied Statistics*, **36**, 296-308.

Lee, J.C. (1988). Prediction and Estimation of Growth Curves with Special Covariance Structures. *Journal of the American Statistical Association*, **83**, 432-440.

Macchiavelli, R.E. y Arnold, S.F. (1994). Variable Order Ante-dependence Models. *Communications in Statistics-Theory and Methods*, **23**, 2683-2699.

Munoz, A., Carey, V., Schouten, J.P., Segal, M. y Rosner, B. (1992). A Parametric Family of Correlation Structures for the Analysis of Longitudinal Data. *Biometrics*, **48**, 733-742.

Nunz Anton, V. y Woodworth, G.G. (1994). Analysis of Longitudinal Data with Unequally Spaced Observations and Time Dependent Correlated Errors. *Biometrics*, **50**, 445-456.

Schluchter, M.D. (1988). Analysis of Incomplete Multivariate Data using Structured Covariance Matrices. *Statistics in Medicine*, **7**, 317-324.

Ware, J.H. (1985). Linear Models for the Analysis of Longitudinal Studies. *The American Statistician*, **39**, 95-101.

Wegman, E.J. (1994). *Huge Data Sets and the Frontiers of Computation Feasibility*. Technical Report No. 110, Center for Computational Statistics, George Mason University.

Checking Normality in Possibly Non-Linear Simultaneous Equations Models

VÍCTOR M. AGUIRRE-TORRES and MARIO CORTINA-BORJA

ITAM, México *Univ. de Oxford, UK*

1. INTRODUCTION

There are several reasons why it is important to check normality of residuals from a multivariate system of non linear equations. They may be separated into the phases of estimation, hypothesis testing, checking model specification, and using the model for simulation. From the point of view of model estimation it is a well known fact that in univariate regression models, the presence of outliers can severely affect the results. See for example Gallant (1987, p. 307). Least squares estimations are also maximum likelihood in the univariate set up, therefore one may expect that multivariate outliers may impact the estimation of a system of equations by means of NL3SLS or FIML. Also, if there is no strong evidence against normality, then this would be a reassurance that by using any of the above methods of estimation an asymptotically optimal estimation procedure has been employed. Considering hypothesis testing, it is shown in Gallant (1975) that likelihood ratio tests (LRT) are better behaved than Wald type tests with respect to their large sample approximations. The usual assumption of most statistical computer packages for the likelihood is multivariate normal, therefore it would be advisable to make a previous check on this assumption. With respect to model specification, just to know that an observation is an outlier could be an indication of the existence of other sources of variation that affect the vector of endogenous variables and that are not considered explicitly in the model. Also, this test could be used as a general test for misspecification, because in the null hypothesis it is also being assumed that the model's functional form is correct. Finally the model may be used for simulations including random shocks, where the distribution of choice is usually mutivariate normal. Most of the work on testing multivariate normality has been done under the i.i.d. case, see for example Wagle (1968), Mardia (1970), and Rincón-Gallardo, Quesenberry and O'Reilly (1979). Checking normality in the univariate regression set up is a standard feature in econometric computer packages like TSP, where a histogram and the Jarque-Bera test are available. However, getting a significant value of a test statistic is not good enough, and a histogram may not be all that informative either. It would be much better to see, for example, a normal plot of the residuals and a box-plot to check for outliers and possible abnormalities. See for example Nelson (1979) for a broader discussion on the use of normal plots, and Gallant (1987) for an application of this idea to nonlinear regression residuals. In the multivariate regression case, Jarque and McKenzie (1982) propose without justification the use of Mardia's sample measures of multivariate skewness and kurtosis evaluated on regression residuals as a test for the disturbance normality assumption. Again, rejecting the hypothesis would be evidence of the lack of normality, however it would not be very informative. One may want to check the normal plots, or box-plots, or hanging rootograms or the time sequence plot for each entry, but these again would be incomplete in the multivariate situation, since due to correlation there could be multivariate outliers that do

not appear to be univariate outliers. For this reason, in this paper we propose the use of relplots and quelplots of the residuals in conjunction with normality tests. The relplots and quelplots are two extensions of the univariate box-plots, and are proposed in Goldberg and Iglewicz (1992).

The presentation of the material is divided into four sections: Section 2 presents the statistical model from a mathematical point of view, the joint normality tests based on Mardia's (1970) tests and an outline of the proof of the asymptotic distribution of the tests. Section 3 presents the construction of the quelplot, relplot, some examples and its relation to sample multivariate skewness and kurtosis. Section 4 gives an application to an econometric example.

2. THE STATISTICAL MODEL

The model is assumed to be in an implicit form consisting of M (possibly) nonlinear equations given by

$$q(y, x, \lambda^*) = e \tag{1}$$

where y is an $M \times 1$ vector of endogenous variables, x is a $K \times 1$ vector of exogenous variables, λ^* is the unknown parameter of the model with dimension L, and e is an $M \times 1$ vector of unobserved disturbances due to slight errors in the specification and/or errors of observation. The data $\{y_t, x_t\}_{t=1}^{n}$ available for estimation of the structural parameters are assumed to follow (1), the errors are independent and identically normally distributed each having zero mean and unknown positive definite variance-covariance matrix Σ.

We assume that the parameters of the model are estimated as follows:

$$\hat{\lambda} \text{ minimizes } S(\lambda) = (1/n) \sum_{t=1}^{n} S(y_t, x_t, \lambda, \hat{\tau}) \tag{2}$$

where in the above formula $\hat{\tau}$ is an estimate of the nuisance parameters of the model, in particular τ may consist of Σ. Depending on the choice of the function $S(*)$, formula (2) contains estimation methods such as: maximum likelihood and the so called "Zellner-type", or "seemingly unrelated regression method", and two or three stage linear or nonlinear least squares. If the later method is used then testing normality would be particularly useful to check whether it is worth pursuing the use of an optimal estimation method like FIML. Regularity conditions under which $\hat{\lambda}$ is a strongly consistent estimator of λ^* and $\sqrt{n}(\hat{\lambda} - \lambda^*)$ is asymptotically normal are given in detail in Gallant (1987), or Gallant (1977). Once the model has been fitted, the residuals from regression (1) are computed as follows:

$$\hat{e}_t = y_t - f(x_t, \hat{\lambda}) \quad t = 1,2,\ldots,n. \tag{3}$$

Let

$$\bar{\hat{e}} = (1/n) \sum_{t=1}^{n} \hat{e}_t \qquad \text{and} \qquad S = (1/n) \sum_{t=1}^{n} (\hat{e}_t - \bar{\hat{e}})(\hat{e}_t - \bar{\hat{e}})^T \tag{4}$$

Form the generalized Mahalanobis distances given by

$$\hat{D}_{tj} = (\hat{e}_t - \bar{\hat{e}})^T S^{-1} (\hat{e}_j - \bar{\hat{e}}), \qquad t, j = 1,2,\ldots,n.$$

Then the measures of multivariate skewness and Kurtosis are given by:

$$b_{1M} = \frac{1}{n^2} \sum_t \sum_j D_{tj}^3 \qquad \text{and} \qquad b_{2M} = \frac{1}{n} \sum_t D_{tt}^2 \tag{5}$$

Mardia (1970) discusses the motivation of both measures in detail. Using the results of that paper, it may be concluded that in the i.i.d case, and if b_{1M}^* and b_{2M}^* are the corresponding measures of skewness and kurtosis but computed on the unobserved discrepancies of the model given (1) and under the assumption of multivariate normality, then

$$nb_{1M}^* / 6 \xrightarrow{L} \chi^2(M(M+1)(M+2)/6) \tag{6}$$

and

$$\sqrt{n}\left[b_{2M}^* - M(M+2)\right] / \sqrt{8M(M+2)} \xrightarrow{L} N(0,1) \tag{7}$$

It is shown in Aguirre (1986) that (7) holds also true when b_{2M} is evaluated on the regression residuals like (3) and under the normality assumption. A similar procedure should work for b_{2M} and (6). The idea of the proof is to show that b_{2M} and b_{2M}^* are asymptotically equivalent. That is, it is shown that

$$\sqrt{n}\,(b_{2M} - b_{2M}^*) \xrightarrow{P} 0$$

and hence from Stutsky's theorem b_{2M} has the same large sample distribution as b_{2M}^*. The expressions on the left hand side of (6) and (7) will be used as the test statistics.

3. RELPLOTS AND QUELPLOTS

As we have pointed out, we are interested in using exploratory procedures which are helpful for identifying outliers and assessing the hypothesis of normality. To do so we take advantage of the bivariate extensions of the boxplot developed by Goldberg and Iglewicz (1992). The Boxplot was introduced by John W. Tukey for univariate data. There have been a great number of graphical methods for extending this technique for bivariate data. A good review of these attempts appears in the paper by Goldberg and Iglewicz (1992). Goldberg and Iglewicz (1992) start by noting that for a bivariate Normal distribution optimal confidence limits are ellipses; thus they propose to use elliptical regions as a natural generalization of the univariate boxplot. In order to have a better performance in the presence of outliers they use robust estimators for location, scale and correlation parameters. These authors propose elliptical regions as an exploratory data analysis tool. One, called *relplot* is a robust symmetric elliptical plot. Since fully elliptical plots assume symmetrical data, they can be considered as somehow restrictive for an initial exploratory data analysis. To overcome this objection, Goldberg and Iglewicz (1992) introduced the *quelplot*. This graph is formed by building four separate quarter ellipses which are matched on their major and minor axes in order to obtain a smooth non symmetric elliptical plot. Both plots consist of two concentric elliptical graphs. As in the univariate boxplot, the inner one (the *hinge*) contains 50% of the data while the outer one (the *fence*) is useful for detecting outliers, defined as points which lie beyond it. Note that the hinge contains 50% of the data rather than a 50% probability region, which makes the method more exploratory in nature, with a minimum of distributional assumptions. The fence is built as an approximate 99% confidence bound calculated assuming a bivariate Normal model. The algorithms for drawing relplots and quelplots are discussed by Goldberg and Iglewicz (1992). We follow very closely these

algorithms. It should be stressed, however, that the estimates that we use to obtain robust estimates of location, scale and correlation do differ from those used by Goldberg and Iglewicz. We take advantage of the Robust Statistics routines implemented in S-plus by Venables and Ripley (1994). For the location estimate we chose an M-estimate. For the location and correlation estimates we use the minimum volume ellipsoid estimate for the variance-covariance matrix. This robust estimate is the covariance matrix that is defined by the ellipsoid with minimum volume of those ellipsoids that contain (n+M+1)/2 of the data points. It is too computationally intense to find the actual estimate, so an approximation is found using the genetic algorithm used by S-plus. The minimum volume ellipsoid covariance estimator has a breakdown point that is almost one-half. That is, the estimate can not be made arbitrarily bad without changing about half of the data. A covariance matrix is considered to be arbitrarily bad if either a component goes to infinity (just as in the breakdown of a location or regression estimate), or if the matrix becomes deficient in rank. This is analogous to a scale estimate breaking down if the estimate is going either to infinity or to zero. Location, scale and correlation are shown in the plot as follows: the intersection of two line segments, which may be either the minor and major axis or the robust regression lines of Y on X and X on Y marks the location of the data; the latter segments are plot within the range of points lying within the fence, so they can be considered as a robust indicator of scale; finally, the acute angle between the regression lines is small for a large absolute value of correlation and large for a small absolute value of correlation; the ratio of the ellipse's axes is also an indicator of correlation. These features make the elliptical plots a very good exploratory tool for bivariate data.

4. AN EXAMPLE

Consider the following model of supply and demand of a certain product harvested in the US. The model has two equations

$$\left.\begin{array}{rl} \text{demand} & y_1 = \alpha_o + \alpha_1 y_2 + \alpha_2 x + e_1 \\ \text{supply} & y_1 = \beta_o + \beta_1 y_2 + e_2 \end{array}\right\} \tag{8}$$

where: y_1 = Production index of the harvest (1977=100), y_2= Price index of the harvest (1977=100), x = Per capita personal expenditure, the source of the data is the "Economic Report of the President, 1986", Table B-26 for x, B-94 for y_1 and B-96 for y_2. The data are shown in Table 1. The system of equations in formula (8) is in implicit form where

$$q(y, x, \lambda) = (y_1 - \alpha_0 - \alpha_1 y_1 - \alpha_2 x, y_1 - \beta_0 - \beta_1 y_2)$$

and the parameters of interest are stacked in λ. That, is in terms of notation (1), we have:

$$\lambda^T = (\alpha_0, \alpha_1, \alpha_2, \beta_0, \beta_1) \qquad y = \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} \qquad e = \begin{bmatrix} e_1 \\ e_2 \end{bmatrix}.$$

Since an endogenous variable (y_2) is present on the right hand side of both equations, we used two stage least squares (TSLS) on each equation to estimate λ. The results presented were obtained using EViews, the Windows version of TSP. From those tables

$$\hat{\lambda}^T = (6.87,\ 0.20,\ 0.008,\ 49.35,\ 0.47).$$

TABLE 1

Production, price of a certain harvest in the US, per capita personal expenditure, (1977=100), 1970-1985. Residuals from model.

Year	Production Index	Price Index	Personal Expenditure	Residuals Equation 1	Residuals Equation 2
1970	77	52	7275	-1.114672	2.917458
1971	86	56	7409	5.964077	10.05041
1972	87	60	7726	3.512569	9.183367
1973	92	91	7972	0.249812	-0.286238
1974	84	117	7826	-11.73410	-20.42204
1975	93	105	7926	-1.168103	-5.820898
1976	92	102	8272	-4.460827	-5.420614
1977	100	100	8551	1.606525	3.512909
1978	102	105	8808	0.456557	3.179102
1979	113	116	8902	8.451780	9.044726
1980	101	125	8784	-4.34625	-7.156127
1981	116	134	8798	8.734855	3.643020
1982	118	121	8825	13.11146	11.71092
1983	88	127	9148	-20.79058	-21.08965
1984	110	138	9462	-3.618290	-4.224026
1985	117	120	9682	5.145360	11.17768

Table 1 also gives the residuals from the fit, and Figure 1 is the corresponding relplot. Notice the strong contemporaneous correlation, and an outlying observation which corresponds to the year of 1983. Table 2 gives the values of the multivariate skewness and kurtosis as well as the corresponding P-values for the tests of joint normality. From there we may see that there is not evidence of departure from normality of the observations.

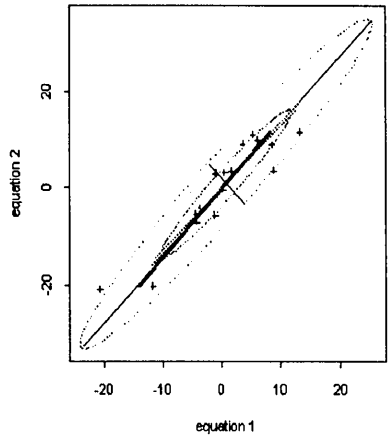


Fig. 1. Relplot residuals model for harvest data.

As a result of this check we get the following: *i*) Since there is no evidence of lack of normality and there is evidence of contemporaneous correlation, a better estimation method should be used (*e.g.* three stage least squares or full information maximum likelihood); *ii*) It would be interesting to investigate what special condition could produce the outlying observation of year 1983. Since the residuals are negative one should look for a factor that could have lowered the harvest much more than expected.

TABLE 4.2
Lack of Normality Tests, Model for Harvest Data

	Coefficient	**Test Statistic**	**P-value**
Skewness (b_{1M})	0.754	2.012	.27
Kurtosis (b_{2M})	7.197	-0.401	.69

REFERENCES

Aguirre, V. (1986). A test for joint normality in multivariate nonlinear regression models. *Com. Int. IIMAS-UNAM*, Serie Naranja, No. 427, 19 pages.

Gallant, A. R. (1975). Nonlinear regression. *The American Statistician* **29**, 73-81.

Gallant, A. R. (1977). Three-stage least squares estimation for a system of simultaneous nonlinear implicit equations. *Journal of Econometrics* **5**, 71-88.

Gallant, A. R. (1987). *Nonlinear statistical models*. John Wiley and Sons, New York.

Goldberg, K. M. and B. Iglewicz (1992). bivariate extensions of the boxplot. *Technometrics* **34**, 307-320.

Jarque, C., and McKenzie, C. (1982). Testing for multivariate normality in simultaneous equations models. Working paper in Economics and Econometrics, No. **82**. The Australian National University.

Mardia, K. V. (1970). Measures of multivariate skewness and kurtosis with applications. *Biometrika* **57**, 519-530.

Nelson, W. (1979). *How to analyze data with simple plots*. How to series, ASQC Quality Press.

Rincón-Gallardo, S., Quesenberry, C. P., and O'Reilly, F. (1979). Conditional probability integral transformations and goodness-of-fit test for multivariate normal distributions. *The Annals of Statistics* **7**, 1052-1057.

Venables, W. N. and B.D. Ripley (1994). *Modern applied statistics with S-plus*, Springer-Verlag, New York

An Algebraic Approach to the Yule-Walker Equations in Time Series Analysis

ROLANDO CAVAZOS-CADENA

Univ. Autónoma Agraria Antonio Narro, Saltillo Coahuila, México

1. INTRODUCTION

Let $\{X_t\}$ be an autoregressive process of order p ($AR(p)$), i.e., for some real numbers $\varphi_1,\ldots, \varphi_p$,

$$X_t + \varphi_1 X_{t-1} + \ldots + \varphi_p X_{t-p} = Z_t, \tag{1}$$

where the Z_t's are zero-mean uncorrelated random variables with common variance $\sigma^2 > 0$; see, for instance, Anderson (1971, pp. 166-176), or Box and Jenkins (1976, pp. 53-65). As noted in Remarks 3 and 5 of Brocwell and Davis (1987, pp. 86-88), it can be assumed that the polynomial φ is causal, i.e., $\varphi(z) \neq 0$ for all z with $|z| \leq 1$, a condition that is supposed to hold true in the following discussion. Now consider the problem of determining the autocovariance function $\gamma(\cdot)$ of $\{X_t\}$, wich is given by $\gamma(h)$: = Cov $[X_{t+h}, X_t]$ for every integer h. The following two step method, which is described in page 97 of Brockwell and Davis (1987), is a computationally convenient tool to determine $\gamma(\cdot)$.

Step 1. Find $\gamma(0)$, $\gamma(1)$, ..., $\gamma(p)$ by solving the Yule - Walker (Y-W) system of equations associated to φ:

$$\sum_{k=0}^{p} \gamma(|i-k|)\, \varphi_k = \begin{cases} \sigma^2, & i = 0 \\ 0, & i = 1,2,\ldots,p. \end{cases} \tag{2}$$

Step 2. Use that $\gamma(i) = -\sum_{k=1}^{p} \gamma(i-k)\, \varphi_k$, $i > p$, to determine $\gamma(p+1)$, $\gamma(p+2)$, ..., in a recursive way.

To establish this method on firm grounds, it is necessary to show that (2) has a unique solution, a fact that can be easily verified when the degree of φ is small, say p = 1 or p = 2. The main objective of this work is to give a necessary and sufficient criterion so that (2) has a unique solution for arbitrary p. The result in this direction, stated below in Theorem 3.1, can be summarized as follows: The Y-W equations (2) have a unique solution if and only if

$$r_i r_j \neq 1, \qquad i,j = 1,2,\ldots,p, \tag{3}$$

where r_1, ..., r_p are the roots of $\varphi(z) = 1 + \varphi_1 z + \cdots + \varphi_p z^p$. When φ is a causal polynomial all of its roots lie outside the unit disk, and then (3) is clearly satisfied. The proof of Theorem 3.1 is by induction and, although elementary, is rather technical and relies on specially builded vector space ideas.

2. NOTATION AND TERMINOLOGY

Throughout the remainder **Z** stands for the set of all intefers, **IN** := {0,1,2, ···} and **C** denotes the set of all complex numbers. The complex vector space **L** consists of all functions

(sequences) $v: \mathbf{IN} \to \mathbf{C}$ with the property that $v(k) = 0$ for all k large enough, and is endowed with the usual addition and scalar multiplication. The (right) shift operator $s : \mathbf{L} \to \mathbf{L}$ is defined as follows: For $v \in \mathbf{L}$,

$$s(v)(0) := 0, \qquad \text{and} \qquad s(v)(k) := v(k-1), \quad k = 1,2, \ldots ; \tag{4}$$

in addition, for $v \in \mathbf{L}$,

$$s^0(v) = v \qquad \text{and} \qquad s^n(v) = s^{n-1}\big(s(v)\big). \tag{5}$$

On the other hand, rows and columns of a squared matrix M are numbered starting from zero and Det M denotes the determinat of M. Define square matrices of order $n + 1$ as follows:

$$\text{For } v_0, v_1, \ldots, v_n \in \mathbf{L}, \qquad M_{n+1}(v_0, v_1, \ldots, v_n)_{ij} := v_i(j), \qquad i,j = 0,1, \ldots , n \tag{6}$$

Hereafter, Span$\{v_0, v_1, \ldots, v_n\}$ stands for the vector space generated by $v_0, v_1, \ldots, v_n$, and Dim(Span$\{v_0, \ldots, v_n\}$) denotes the corresponding dimension. To conclude, with a given polynomial $\varphi(z) = \varphi_0 + \varphi_1 z + \cdots + \varphi_p z^p$ of degree p we associate two vectors $\vec{\varphi}$ and $\overleftarrow{\varphi}$ in L defined as follows:

$$\vec{\varphi}(k) := \varphi_k \quad \text{and} \quad \overleftarrow{\varphi}(k) := \varphi_{p-k}, \quad k = 0,1,\ldots, p; \quad \vec{\varphi}(k) := 0, \quad \text{and} \quad \overleftarrow{\varphi}(k) := \quad k > p \tag{7}$$

Finally, it is convenient to introduce the next convention concerning the coeficients of $\varphi(z)$:

$$\varphi_k := 0 \qquad \text{for} \quad k < 0 \qquad \text{or} \quad k > p. \tag{8}$$

3. THE RESULT

Let φ be a polynomial of degree p. In this section we state a formula for the determinant of the matrix corresponding to the Y-W system (2) associated to φ. To begin with, notice that for every positive integer i, $\sum_{k=0}^{p} \gamma(|i-k|)\varphi_k = \sum_{k=0}^{i} \gamma(i-k)\varphi_k + \sum_{k=i+1}^{p} \gamma(k-i)\varphi_k = \sum_{j=0}^{i} \gamma(j)\varphi_{i-j} + \sum_{k=1}^{p-i} \gamma(j)\varphi_{i+j}$, and using convention (8) this yields that $\sum_{k=0}^{p} \gamma(|i-k|)\varphi_k = \gamma(0)\varphi_i + \sum_{j=1}^{p} \gamma(j)\big[\varphi_{i-j} + \varphi_{i+j}\big]$. Thus, the Y-W system (2) can be equivalently written as

$$\sum_{k=0}^{p} \gamma(k)\varphi_k = \sigma^2 \qquad \text{and} \qquad \gamma(0)\varphi_i + \sum_{j=1}^{p} \gamma(j)\big[\varphi_{i-j} + \varphi_{i+j}\big] = 0, \qquad i = 1,2,\ldots, p \tag{9}$$

The (squared) matrix of this system will be denoted by $M(\varphi)$. Clearly, $M(\varphi)$ is of order $(p+1)$ and its entries are determined as follows: For $i = 0,1,2,\ldots, p$,

$$M(\varphi)_{i0} := \varphi_i, \qquad \text{and} \qquad M(\varphi)_{ij} := \varphi_{i-j} + \varphi_{i+j} \quad j = 1,2,\ldots, p; \tag{10}$$

for instance, $M(\varphi)_{00} = \varphi_0$, and for $j > 0$, $M(\varphi)_{0j} = \varphi_{0-j} + \varphi_{0+j} = \varphi_j$, in accordance with the first equation in (9). The next theorem contains a formula for the determinant of M(φ) and, as a by-product, a criterion for the nonsingularity of M(φ) is obtained.

Theorem 3.1 Let $\varphi(z) = 1 + \varphi_1 z + \ldots + \varphi_p z^p$ be a complex polynomial of degree p. If the roots of φ are $r_1, \ldots, r_p$ and $M(\varphi)$ is as in (10), then (*i*) and (*ii*) below occur.

$$(i)\ \operatorname{Det} M(\varphi) = \prod_{1 \le i < j \le p} \left[1 - (r_i r_j)^{-1}\right] \prod_{i=1}^{p} \left(1 - r_i^{-2}\right) \tag{11}$$

by (the usual) convention, for $p = 1$ the first product in (11) is 1.

(*ii*) $M(\varphi)$ is non-singular if and only if $r_i r_j \neq 1$ for all $i, j = 1, \ldots, p$.

A proof of Theorem 3.1 will be presented in Section 5. Presently, we just note that (*ii*) follows immediately from part (*i*). Also, (11) is easily verified for small values of p. For instance, for $p = 1$, $\varphi(z) = 1 + \varphi_1 z$ and

$$M(\varphi) = \begin{bmatrix} 1 & \varphi_1 \\ \varphi_1 & 1 \end{bmatrix};$$

see (10). In this case $\operatorname{Det} M(\varphi) = 1 - \varphi_1^2$, and this yields (11) for $p = 1$, since φ has the unique root $r_1 = -1/\varphi_1$. For $p = 2$ factorize $\varphi(z)$ as $\varphi(z) = (1 + a_1 z)(1 + a_2 z)$, where the roots of φ are $r_i = -1/a_i$, $i = 1,2$. Then $\varphi(z) = 1 + (a_1 + a_2) z + a_1 a_2 z^2$, and M ($\varphi$) is given by

$$M(\varphi) = \begin{bmatrix} 1 & a_1 + a_2 & a_1 a_2 \\ a_1 + a_2 & a_1 a_2 + 1 & 0 \\ a_1 a_2 & a_1 + a_2 & 1 \end{bmatrix}.$$

Then, expanding Det $M(\varphi)$ by the third column we get Det $M(\varphi) = (1 - a_1 a_2)(1 - a_1^2)(1 - a_2^2)$ and replacing a_i by $-1/r_i$ we get (11) with $p = 2$. The proof of (11) for arbitrary p is contained in Section 5.

4. PRELIMINARIES

This section contains the preliminary results that will be used to establish Theorem 3.1. The starting point is the following.

Definition 4.1 Let $\varphi(z) = 1 + \varphi_1 z + \ldots + \varphi_p z^p$ be a polynomial of degree $p > 0$. The sequence $V^\varphi = \{V_t^\varphi \mid t \in \mathbf{Z}\} \subset \mathbf{L}$ is defined by

(*i*) For $1 \le n < p$, $V_n^\varphi(0) := \varphi_n$, and $V_n^\varphi(k) := \varphi_{n+k} + \varphi_{n-k}$, $k = 1, 2, \ldots$ (see (4)-(8) for notation);

(*ii*) For $n \in \mathbf{IN}$, $V_{-n}^\varphi := s^n(\tilde{\varphi})$ and $V_{n+p}^\varphi := s_n(\tilde{\varphi})$.

A glance to Definition 4.1, (6)-(8) and (10) shows that the sequence V^φ is related to $M(\varphi)$ through the following equality:

$$M(\varphi) = M_{p+1}\left(V_0^\varphi, V_1^\varphi, \ldots, V_p^\varphi\right). \tag{12}$$

The key technical result of this section follows.

Theorem 4.1 Suppose that $\varphi(z) = 1 + \varphi_1 z + \ldots + \varphi_p z^p$ has degree p and satisfies $\varphi(b) = \varphi(1/b) = 0$ for some $b \in \mathbf{C} \setminus \{0\}$. Then (*i*)-(*iii*) occur.

(*i*) $\mathrm{Dim}(\mathrm{Span}\{V_1^{\varphi}, V_0^{\varphi}, \ldots, V_{p+1}^{\varphi}\}) \le p+1$.

(ii) For each $a \in \mathbf{C}$, $\mathrm{Det}\, M_{p+2}(V_0^{\varphi} + aV_{-1}^{\varphi}, V_1^{\varphi} + aV_0^{\varphi}, \ldots, V_{p+1}^{\varphi} + aV_p^{\varphi}) = 0$

(iii) For all $a \in \mathbf{C}$, $\mathrm{Det}\, M[(1+az)\varphi] = 0$.

The proof of Theorem 4.1 has been divided into several pieces given in the form of Lemmas 4.1-4.4; these preliminaries involve the notions introduced in the next definition.

Definition 4.2 Let $V = \{V_t \mid t \in \mathbf{Z}\}$ be a sequence in $\mathbf{L}$ and $k \in \mathbf{IN}$.

(*i*) The sequence V has property $D(k)$ if, for all $n \in \mathbf{IN}$,

$$\mathrm{Dim}\left(\mathrm{Span}\left\{V_t \mid -n \le t \le k+n\right\}\right) \le k+n.$$

(ii) Given $a \in \mathbf{C}$, the sequence $T_aV = \{T_aV_t \mid t \in \mathbf{Z}\}$ is defined by $T_aV_t := V_t + aV_{t-1}$, $t \in \mathbf{Z}$.

The next lemma is the starting point in the walk to the proof of Theorem 4.1.

Lemma 4.1 Let $\varphi(z) = 1 + \varphi_1 z + \ldots + \varphi_p z^p$ be a polynomial of degree p and $a \in \mathbf{C} \setminus \{0\}$. If $\theta(z) = (1+az)\varphi(z)$, then $V^{\theta} = T_aV^{\varphi}$.

We now study how property $D(k)$ is related to a transformation T_a.

Lemma 4.2 Let $a \in \mathbf{C}$ be arbitrary and suppose that $V = \{V_t \mid t \in \mathbf{IN}\} \subset \mathbf{L}$ has property $D(k)$. Then T_aV has property $D(k+1)$.

The next two lemmas relate property $D(k)$ with sequences V^{φ}.

Lemma 4.3 Let $\varphi(z)$ be a polynomial of degree $p \ge 1$ and suppose that $\varphi(1)=0$ or $\varphi(-1) = 0$. Then, V^{φ} has property $D(p)$.

Lemma 4.4 Let $\varphi(z)$ be a polynomial of degree $p \ge 2$ that $\varphi(b) = \varphi(1/b) = 0$ for some $b \in \mathbf{C}\setminus\{0,1,-1\}$. Then, V^{φ} has property $D(p)$.

We now put all the pieces together.

Proof of Theorem 4.1. Let $\varphi(z) = 1 + \varphi_1 z + \ldots + \varphi_p z^p$ be a polynomial of degree p with $\varphi(b) = \varphi(1/b) = 0$ for some $b \in \mathbf{C}\setminus\{0\}$.

(*i*) When $b = 1$ or $b = -1$ Lemma 4.3 yields that V^{φ} has property $D(p)$ and, by Lemma 4.4, the same holds when $b \ne 1, -1$. Then, by Definition 4.2 (*i*), we conclude that $\mathrm{Dim}(\mathrm{Span}\{V_{-1}^{\varphi}, V_o^{\varphi}, \ldots, V_p^{\varphi}, V_{p+1}^{\varphi}\}) \le p+1$.

(*ii*) First notice that $\mathrm{Span}\{V_r^{\varphi} + a \cdot V_{r-1}^{\varphi} \mid r = 0,1,\ldots,p+1\} \subset \mathrm{Span}\{V_t^{\varphi} \mid -1 \le t \le p+1\}$, and then by part (*i*), $\mathrm{Dim}(\mathrm{Span}\{V_r^{\varphi} + a \cdot V_{r-1}^{\varphi} \mid r = 0,1,\ldots,p+1\} \le p+1$. It follows that the $p+2$ vectors $V_r^{\varphi} + a \cdot V_{r-1}^{\varphi}$, $r = 0,1,2,\ldots,(p+1)$ are linearly dependent in $\mathbf{L}$, and it is clear that this fact implies the linear dependence of the rows of $M_{p+2}\left(V_0^{\varphi} + a \cdot V_{-1}^{\varphi}, \ldots, V_{p+1}^{\varphi} + a \cdot V_p^{\varphi}\right)$; consequently, $\mathrm{Det}\, M_{p+2}\left(V_0^{p} + a \cdot V_{-1}^{\varphi}, \ldots, V_{p+1}^{\varphi} + a \cdot V_p^{\varphi}\right) = 0$; see, for instance, Chapter 5 in Hoffman and Kunze (1971).

(*iii*) Set $\psi(z) := (1 + az) \cdot \varphi(z)$. Then ψ has degree $p+1$, and using (12) with $p+1$ and ψ instead of p and φ respectively, we see that $M(\psi) = M_{p+2}\left(V_0^{\psi}, V_1^{\psi}, \ldots, V_{p+1}^{\psi}\right) = M_{p+2}\left(V_0^{\varphi} + a \cdot V_{-1}^{\varphi}, \ldots, V_{p+1}^{\varphi} + a \cdot V_p^{\varphi}\right)$, where Lemma 4.1 was used to obtain the second equality. Finally, part (*ii*) yields $\mathrm{Det}\, M(\psi) = 0$.

To conclude this section we state a simple fact that will be useful in the proof of Theorem 3.1.

Lemma 4.5 Let $\varphi(z)$ be a polynomial of degree p with $\varphi(0) = 1$. Then,

$$\text{Det } M(\varphi) = \text{Det } M_{p+2}\left(V_0^{\varphi}, V_1^{\varphi}, \ldots, V_{p+1}^{\varphi}\right).$$

5. PROOF OF THEOREM 3.1

We are finally ready to establish Therem 3.1.

Proof of Theorem 3.1 As already noted it is sufficient to prove part (*i*). Let $\varphi(z) = 1 + \varphi_1 z + \ldots + \varphi_p z^p$ be a polynomial of degree p and factorize φ as $\varphi(z) = \prod_{i=1}^{p} (1 + a_i z)$, where the roots of φ are $-1/a_i$, , $i = 1,2, \ldots, p$. With this notation (11) is equivalent to

$$\text{Det } M\left[\prod_{i=1}^{p} (1 + a_i z)\right] = \prod_{1 \le i < j \le p} (1 - a_i a_j) \prod_{i=1}^{p} (1 - a_i^2), \tag{13}$$

an equality that was verified in Section 3 for $p = 1$ and $p = 2$. We complete the proof of (13) by an induction argument. Suppose that (13) holds for $p = n \ge 2$, and let $a_1, a_2, \ldots, a_{n+1}$ be no-null complex numbers. Define

$$\psi(z) := \prod_{i=1}^{n} (1 + a_i z), \tag{14}$$

and for $c \in \mathbf{C}$, set

$$F(c) := \text{Det } M_{n+2}\left(V_0^{\psi} + c \cdot V_{-1}^{\psi}, \ldots, V_{n+1}^{\psi} + c \cdot V_n^{\psi}\right) \tag{15}$$

Combining Lemma 4.1 and (12) wet get

(a) $F(c) = \text{Det } M[(1+cz)\psi(z)]$; in particular,

$$F(a_{n+1}) = \text{Det } M\left[\prod_{i=1}^{n+1} (1 + a_i z)\right]. \tag{16}$$

Using the multilinearity of the determinant function, (15) yields

(b) $F(c)$ is a polynomial in c with degree $\le n + 2$; see, for instance, Hoffmann and Kunze (1971) Chapter 5. Next, we find the roots of $F(c)$. First observe that

(c) $F(1) = F(-1) = 0$.

To see this set $\psi^*(\mathbf{z}) := (1+z) \cdot \prod_{i=2}^{n} (1 + a_i z)$ and notice that $\psi^*(-1) = 0$, and $(1+z)\psi(z) = (1+a_1 z)\psi^*(z)$; see (14). Then (a) above yields $F(1) = \text{Det } M[(1+z)\cdot\psi(z)] = \text{Det } M[1+a_1 z)\psi^*(z)] = 0$, where the last equality is due to Theorem 4.1(*iii*) with ψ^* and -1 instead of φ and b, respectively. Similary, it can be shown that $F(-1) = 0$.

(d) $F(1/a_i) = 0$, $i = 1,2, \ldots, n$.

To show this, pick k, $i \in \{1,2, \ldots, n\}$ with $k \ne i$, and define

$$\tilde{\psi}(z) := (1 + z/a_i) \prod^{(k)} (1 + a_j z), \tag{17}$$

where $\prod^{(k)}$ indicates product over all $j \ne k$, $1 \le j \le n$; notice that

$$(1 + z/a_i) \cdot \psi(z) = (1 + a_k z) \cdot \tilde{\psi}(z). \tag{18}$$

From (17) it is clear that $\tilde{\psi}(-a_i) = 0$ and since $k \neq i$, $\tilde{\psi}(z)$ contains the factor $(1+a_i z)$, and then $\tilde{\psi}(-1/a_i) = 0$. Therefore, from (a) and (18) we obtain $F(1/a_i) = \text{Det } M[(1+z/a_i)\cdot\psi(z)] = \text{Det } M[(1+a_k z)\cdot\tilde{\psi}(z)]$, and using Theorem 4.1(*iii*) with $\tilde{\psi}$ and $-a_i$ instead of φ and b, respectively, we conclude that $F(1/a_i) = 0$. To continue, suppose for the moment that $a_1, a_2, \ldots, a_n$ are different numbers in $\mathbf{C}\setminus\{0,1,-1\}$. In this case, (c) and (d) show that the polynomial $F(c)$ has $n+2$ roots, namely, $1, -1$, and $1/a_i$, $i = 1,2, \ldots, n$. Combining this fact with (b) we see that $F(\cdot)$ has degree $n+2$ and it can be factorized as $F(c) = F(0)\cdot(1-c)(1+c)\cdot\prod_{i=1}^{n}(1-a_i c)$. Setting $c = a_{n+1}$ and using (a) it follows that

$$\text{Det } M\left[\prod_{i=1}^{n+1}(1+a_i z)\right] = \left(1-a_{n+1}^2\right)\cdot\prod_{i=1}^{n}\left(1-a_i a_{n+1}\right)\cdot F(0). \tag{19}$$

Next, using (15) we see that $F(0) = \text{Det } M_{n+2}\left(V_0^{\psi}, V_1^{\psi}, \ldots, V_{n+1}^{\psi}\right)$, and then $F(0) = \text{Det } M(\psi)$, by Lemma 4.5 applied to ψ, which has degree n. Now, the induction hypothesis yields $F(0) = \prod_{i=1}^{n}\left(1-a_i^2\right)\cdot\prod_{1\le i,\, j\le n}\left(1-a_i a_j\right)$, and combining this equality with (19) we get

$$\text{Det } M\left[\prod_{i=1}^{n+1}(1+a_i z)\right] = \prod_{i=1}^{n+1}\left(1-a_i^2\right)\cdot\prod_{1\le i,\, j\le n+1}\left(1-a_i a_j\right), \tag{20}$$

which is (13) with $p = n+1$. Although (20) has been established under the assumption that $a_1, \ldots a_n$ are different numbers in $\mathbf{C}\setminus\{0,1,-1\}$, the equality holds for arbitrary $a_1, \ldots, a_n \in \mathbf{C}\setminus\{0\}$, since both sides of (20) are continuous functions of the a_i's. In short, assuming that (13) holds for $p = n$ we have seen that it is also valid for $p = n+1$. This completes the proof of Theorem 3.1.

ACKNOWLEDGEMENTS

This work was partially supported by the PSF Organization under Grant 200-300/3-95, and by the MAXTOR Foundation for Applied Probability and Statistics (MAXFAPS) under Grant 01-01-56/5-93

REFERENCES

Anderson, T. W. (1971), *The Statistical Analysis of Time Series, Wiley*, New York.

Box, G. E. P. and G. M. Jenkins (1976), *Time Series Analysis, Forecasting and Control, Holden-Day*, San Francisco.

Brockwell P. J. and R. A. Davis (1987), *Time Series: Theory and Applications, Springer-Verlag*, New York.

Hoffman K. and r. Kunze (1971), *Linear Algebra, Prentice-Hall*, Englewood Cliffs, Massachusetts.

Combining Information in Time Series Analysis

VICTOR M. GUERRERO and DANIEL PEÑA

ITAM, México *Univ. Carlos III de Madrid*

1. INTRODUCTION

Combining information has such a common place in the practice of statistics that the practicing statistician many times overlooks it. Hedges and Olkin (1985) presented many statistical problems that can be analyzed from this point of view. Draper et al. (1992) provided a thorough review of this field with many examples and ideas for future research. Similarly, Peña (1994) considered combining information with emphasis on understanding the structure and properties of the estimators involved in the combination.

This work presents a basic (least squares) rule that has been frequently used by time series analysts for combining information. We assume here that some information, additional to that employed by a time series model, is available in the form of linear restrictions that have to be fulfilled exactly by an optimal estimator. Our basic concern is to obtain (conditionally) unbiased Minimum Mean Square Error Linear Estimators (MMSELE) of random vectors. Hence no distributional assumption will be required for obtaining the optimal estimators, although when normality is a reasonable assumption, the linear qualification can be dropped from MMSELE.

2. BASIC COMBINING RULE

In this section we establish an optimal combining rule that can be employed when two basic sources of information are available. (i) A statistical model that produces the MMSELE, **W**, of the random vector **Z**, based on an observed set of explanatory variables X, and (ii) some extra-model information **Y** given in the form of linear restrictions imposed on **Z**. The model implied by (i) may be written as $\mathbf{Z} = \mathrm{X}\boldsymbol{\beta} + \mathbf{u}$, with $\boldsymbol{\beta}$ a fixed parameter vector and **u** a random vector, but this model form will not be used explicity in what follows. In fact, we shall assume that the model is known, as well as its parameters. We now establish the rule and illustrate its use in different situations afterwards.

Basic Combining Rule: Let us suppose that **Z**, **W** and **Y** are related by

$$\mathbf{W} = \mathbf{Z} + \mathbf{e} \tag{1}$$

and

$$\mathbf{Y} = \mathbf{CZ}, \tag{2}$$

where **e** is a random vector such that $\mathrm{E}(\mathbf{e}|\mathrm{X}) = 0$, $\mathrm{E}(\mathbf{Ze}'|\mathrm{X}) = 0$, $\mathrm{Cov}(\mathbf{e}|\mathrm{X}) = \Sigma_e$ and $\mathbf{W} = E(\mathbf{Z}|X)$ is the MMSELE of **Z**. If C is a known full-rank matrix and **W**, **Y** and Σ_e are also known, with Σ_e nonsingular, then the MMSELE of **Z** based on **W** and **Y** is given by

$$\hat{\mathbf{Z}} = \mathbf{W} + \Sigma_e C'(C\Sigma_e C')^{-1}(\mathbf{Y} - C\mathbf{W}), \tag{3}$$

with MSE matrix

$$Cov(\hat{\mathbf{Z}} - \mathbf{Z}) = \Sigma_e - \Sigma_e C'(C\Sigma_e C')^{-1} C\Sigma_e. \tag{4}$$

This Basic Rule allows us to combine **W** and **Y** in an optimal manner. However, it does not necessarily follow that **W** and **Y** should always be combined. In particular, it will not be sensible to combine them when they contradict each other. Then, it makes sense to test for compatibility between **W** and **Y** to see whether or not the combined predictor is reasonable. To that end, a compatibility test derived on the assumption of normality for **e** was proposed by Guerrero (1989). That is, let us consider as null hypothesis $H_o: \mathbf{Y} = CE(\mathbf{Z}|X)$. On this hypothesis $\mathbf{Y} - C\mathbf{W}$ is normally distributed with mean vector zero and covariance matrix $C\Sigma_e C'$, therefore a statistic for testing compatibility between **Y** and **W** is given by

$$K = (\mathbf{Y} - C\mathbf{W})'(C\Sigma_e C')^{-1}(\mathbf{Y} - C\mathbf{W}) \sim \chi^2_m \tag{5}$$

where m is the dimension of **Y**.

It is important to realize that the Basic Rule as well as its companion compatibility test, can be obtained within a more general setting in which (2) is replaced by $\mathbf{Y} = C\mathbf{Z} + \mathbf{u}$ with **u** a random vector such that E(**u**)=**0** and Cov (**u**)=U. However, in all cases considered here U = 0, so there is no need of considering **u** explicity. In another work we shall consider situations in which $U \neq 0$, so that a more general combining rule will be established together with its associated relevant analysis.

Now, when talking about modelling a univariate time series $\{Z_t\}$, we shall assume that it admits an ARIMA representation. We also let $X = (Z_1, \ldots, Z_N)'$ be the observed data and $\mathbf{Z} = (Z_{N+1}, \ldots, Z_{N+H})'$ be the H>1 future values to be forecasted with origin at time N. Then we know that

$$Z_{N+h} - E(Z_{N+h}|X) = \sum_{j=0}^{h-1} \Psi_j a_{N+h-j}, \text{ for } h = 1, \ldots, H \tag{6}$$

where the $\Psi_j's$ are the pure moving average (MA) weights of the model and $\{a_t\}$ is a zero-mean white noise process with variance σ^2. Expression (6) can be rewritten in matrix notation as

$$\mathbf{Z} - E(\mathbf{Z}|X) = \Psi\mathbf{a}, \tag{7}$$

with $\mathbf{a}' = (a_{N-1}, \ldots, a_{N+H})'$ and Ψ a lower triangular matrix with $\Psi_o = 1$ in the main diagonal, Ψ_1 in the second diagonal, and so on

Notice in particular that (7) holds true for both stationary and nonstationary time series. For a stationary series with $E(\mathbf{Z}|X) = \mathbf{0}$ we have $\mathbf{e} = \Psi\mathbf{a}$ and $\Sigma_Z = \sigma^2\Psi\Psi'$. Also $\Pi\mathbf{Z} = \mathbf{a}$, where $\Pi = \Psi^{-1}$ is a lower triangular matrix with ones in the main diagonal, $-\pi_1$ in the second diagonal and so on. The $\pi_i's$ are the pure autoregressive (AR) coefficients of the ARMA process. Then $\Sigma_Z^{-1} = \sigma^{-2}\Pi'\Pi$ is the inverse autocorrelation function of the process. In general we shall call $\mathbf{W} = E(\mathbf{Z}|X)$, $\mathbf{e} = \Psi\mathbf{a}$ and $\Sigma_e = \sigma^2\Psi\Psi'$, and for a stationary process $\Sigma_Z = \Sigma_e$. It should also be stressed that even though most of the problems considered in this paper make explicit reference to univariate time series, the same ideas can be employed with multiple time series. The basic change required in that situation, from a theoretical viewpoint, will be notational.

3. APPLICATIONS OF THE BASIC RULE

The situations considered in this section are used as illustrative examples of application of the Rule previously established.

3.1 Forecast Updating

We consider first the problem of updating a vector of ARIMA forecasts, initially obtained for lead times h=1, ..., H, with origin at time N. As soon as we have access to a new observation Z_{N+1}, its forecast $\hat{Z}_N(1)$ becomes useless and $\hat{Z}_N(2),\ldots,\hat{Z}_N(H)$ are suboptimal by not taking into account all the available information. The solution to this problem appears in Box and Jenkins (1976, Ch. 5), but it can be solved alternatively by applying the Basic Rule.

3.2 Estimation of Missing Data

The basic need for completing a time series which has some missing values is that most time series analysis require a data set without gaps. Besides, in some cases the estimation of missing observations is the main objective of the analysis. The basic references for this problem, from the present standpoint, are Peña (1987), Peña and Maravall (1991) and Guerrero (1994).

3.3 Restricted Forecasting Without Uncertainty in the Restrictions

This case occurs when some restrictions to be imposed on the time series forecast are known to be true in advance. For instance we may consider imposing budget constraints, or else we may view this kind of application as a scenario (or what if) analysis. For instance, Guerrero (1989) mentions the problem of forecasting the monthly Financing Granted by the Mexican Bank System when **Y**, the total annual financing, was known in advance.

3.4 Change Foreseen in the Deterministic or Stochastic Structure of the Model

Let us now suppose that a structural change in the structure of the time series model is foreseen to occur during the forecast horizon of interest. This idea may come from subject matter considerations, for example when an intervention is anticipated. This case may be considered as an ex-ante intervention analysis, in which the whole effect of the intervention is presumably accounted for by way of some linear restrictions on the future values of the series. Guerrero (1991) considered this situation and provided the solution.

3.5 Change in Parameter Values Due to an Intervention in the Forecast Horizon

The problem now is combining ARIMA forecasts with some linear restrictions when an intervention is anticipated and its effects are feared to change the original values of the AR and MA parameters. A solution to this problem was given by Guerrero (1990a).

3.6 Temporal Disaggregation

The problem of temporal disaggregating a time series is that of estimating an unobserved random vector $\mathbf{Z} = (Z_1,\ldots,Z_{mn})'$ on the basis of knowing some linear aggregates

$Y_i = \sum_{j=1}^{n} c_j Z_{n(i-1)+j}$, with i=1,...,m. Here n denotes the intraperiod frecuency of observation (i.e. if $\{Y_i\}$ is observed annualy and $\{Z_t\}$ is a monthly series, n=12) m is the number of whole-period observations and $\mathbf{c} = (c_1,\ldots,c_n)' \neq \mathbf{0}$. Some usual forms of $\mathbf{c}$ are: $\mathbf{c} = (0,0,\ldots,0,1)'$ for interpolating a stock series, $\mathbf{c} = (1,1,\ldots,1)'$ for distributing a flow series and $\mathbf{c} = (1/n, 1/n, \ldots, 1/n)'$ for distributing an index series. The basic references are now Cohen, Müller and Padberg (1971), Harvey and Pierce (1984), and Wei and Stram (1990), when no auxiliary information is available, whereas Chow and Lin (1971), Denton (1971) and Guerrero (1990b) make use of auxiliary information.

3.7 Detecting and Measuring the Effect of Influential Outliers

We consider here the single outlier case with known time of occurrence T. So, we employ the two basic mechanisms that may generate an Additive Outlier (AO) or an Innovational Outlier (IO). These were considered for instance, by Tsay (1986) and Peña (1990).

3.8 Reallocation Outliers

A situation considered and illustrated by Wu, Hosking and Ravishanker (1993), basically consists in restricting a block of consecutive observations affected by outliers to produce the same sum as if no outliers were present.

4. CONCLUSIONS

The Basic Rule for combining information from two different sources is a very useful tool for solving time series problems. Such a rule produces a weighted average of two different predictors coming from each source of information. Its optimality is easily revealed by the corresponding Mean Square Error matrix which is not only minimum for the class of linear and unbiased estimators considered, but because it shows that using the extra-model information reduces the original variability in the model estimator.

Realizing that many statistical procedures are derived by combining information is important from a unifying point of view. Besides, we advocate the use of compatibility tests in order to appreciate whether the combination makes sense or not. Some of these tests have already appeared in the time series literature, associated mainly with likelihood-based inferences.

REFERENCES

Box, G.E.P. and Jenkins, G.M. (1976). *Time Series Analysis, Forecasting and Control.* Rev. ed. San Francisco: Holden-Day.

Cohen, K.J., Müller, W. and Padberg, M.W. (1971). Autoregressive Approaches to Disaggregation of Time Series Data. *Appl. Statist.* **20**, 119-129

Chow, G.C. and Lin, A. (1971) Best Linear Interpolation, Distribution and Extrapolation of Time Series by Related Series. *Rev. Econ. Statist.* **53**, 372-375.

Denton, F.T. (1971). Adjustment of Monthly or Quarterly Series to Annual Totals: An Approach Based on Quadratic Minimization. *J. Am. Statist. Assoc.* **66**, 99-102.

Draper, D., Gaver, D.P. Jr., Goel, P.K., Greenhouse, J.B., Hedges, L.V., Morris, C.N., Tucker, J.R. and Waternaux, C.M. (1992) *Combining Information. Statistical Issues and Opportunities for Research.* Washington: National Academy Press.

Guerrero, V.M. (1989). Optimal Conditional ARIMA Forecasts. *J. Forecasting* **8**, 215-229.

Guerrero, V.M. (1990a). Restricted ARIMA Forecasts which Account for Parameter Changes. *ESTADISTICA* **42**, 17-31.

Guerrero, V.M. (1990b). Temporal Disaggregation of Time Series. An ARIMA-based Approach. *Int. Statist. Rev.* **58**, 29-46.

Guerrero, V.M. (1991). ARIMA Forecasts with Restrictions Derived from a Structural Change. *Int. J. Forecasting* **7**, 339-347.

Guerrero, V.M. (1994). Restricted Forecasts of Missing Observations in Univariate Time Series Models. ESTADISTICA 46, 1-23.

Harvey, A.C. and Pierse, R.G. (1984). Estimating Missing Observations in Economic Time Series. *J. Am. Statist. Assoc.* **79**, 125-131.

Hedges, L.V. and Olkin, I. (1985). *Statistical Methods for Meta-Analysis.* New York: Academic Press.

Peña, D. (1987) Measuring the Importance of Outliers in ARIMA Models. *In New Perspectives in Theoretical and Applied Statistics*, M. Puri et al (eds) New York: Wiley, 109-112

Peña, D. (1990) Influential Observations in Time Series. *J. Bus Econ. Statist.* **8**, 235-241.

Peña, D. (1994). Combining Information in Statistical Modelling. Working Paper. Universidad Carlos III de Madrid.

Peña, D. and Maravall, A. (1991). Missing Observations, Additive Outliers and Inverse Autocorrelation Function. *Comm. Statist.* (Theory and Methods) **A-20**, 3175-3186.

Tsay, R.S. (1986). Time Series Model Specification in the Presence of Outliers. *J. Amer. Statist. Assoc.* **81**, 132-141.

Wei, W.W.S. and Stram, D.O. (1990). Disaggregation of Time Series Models. *J. Roy Statist. Soc.* **B-52**, 453-467

Wu, L.S.Y., Hosking, J.R.M. and Ravishanker, N. (1993). Reallocation Outliers in Time Series. *Appl. Statist.* **42**, 301-313.

Blockmodels: A Complement to Log-Linear Models

J. VAN HOREBEEK and J.L. TEUGELS

CIMAT, Guanajuato, México *K.U. Leuven, Belgium*

1. BASIC PROBLEM

In this paper we present a group of models to analyze categorical data. We first show how the classical approach by means of log-linear models fails in some particular cases. Therefore we introduce a complementary family where the basic parameters are the (higher) moments of the underlying distribution and derive the corresponding transformation rules.

Consider the following data set taken from Hagenaars (1990). It concerns a study about the changes in political preferences during the post election period February 1977 and March 1977; people were asked for which party (X_1) and for which prime minister (X_3) they voted in February and for which party (X_2) and prime minister (X_4) they would vote if one organized new elections at that moment (March 1977).

TABLE 1

	X_3: March		
	Christ. Dem.	Left Wing	Others
X_1: February			
Christian Democratic	242	17	25
Left Wing	9	350	26
Others	294	418	388
	X_4: March		
	Van Agt	Den Uyl	Others
X_2: February			
Van Agt	111	19	42
Den Uyl	16	410	49
Others	180	495	425

Questions of interest are:

1. Did the party preference change? i.e. $P(X_1) = P(X_3)$ (uni-marginal homogeneity)?
2. Is the change in party preference equal to the change in the preference for the prime minister i.e. $P(X_1,X_2) = P(X_3,X_4)$? (bi-marginal homogeneity)?
3. Since it is clear that the marginals differ, one sometimes tests partial marginal homogeneity: 'this is bi-marginal homogeneity but we don't require uni-marginal homogeneity" (Hagenaars, 1990).

In a typical analysis of categorical data, one would resort to a log-linear parametrization and translate the above hypotheses in terms of restrictions on those parameters. Hòwever, since they are defined by means of conditional characteristics (log-odds ratios), this is not a straightforward task and leads to tricky calculations; e.g. the hypothesis of marginal homogeneity is often indirectly checked by tests for symmetry and quasi-symmetry.

In the next section we avoid those problems by constructing a new class of models parametrized by means of the (higher) moments.

2. BLOCKMODELS

Definition 2.1. Define $M(r_1,...,r_n) = R^{\{0,...,r_1-1\} \times ... \times \{0,...,r_n-1\}}$; an element belonging to $M(r_1,..., r_n)$ is called a block.

Definition 2.2 A flat $(A^1|...|A^n)$ is an ordered sequence of matrices where $A^i \in M(r_i, s_i)$, $1 \leq i \leq n$.

In a straightforward way, one defines the addition and multiplication of flats as the componentwise addition resp. multiplication of the matrices of the flats.

Definition 2.3 If A is a flat $(A^1|...|A^n)$ with $A^i \in M(r_i, s_i)$ and $B \in M(s_1,..., s_n)$, define the flatproduct $A \triangleright B \in M(r_1,..., r_n)$ as

$$(A \triangleright B)_{i_1,...,i_n} = \sum_{k_1=0}^{s_1} A^1_{i_1,k_1} \sum_{k_2=0}^{s_2-1} A^2_{i_2,k_2} \cdots \sum_{k_n=0}^{s_n-1} A^n_{i_n,k_n} B_{k_1,...k_n}$$

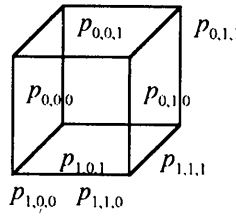


Fig. 1. Block representation of a three dimensional binary variable

For some nice properties of those operators, we refer to Teugels and Van Horebeek (1995).

Definition 2.4 Given a sequence of random variables $Z_{i,k}$, $1 \leq k \leq n$ and $0 \leq i \leq r_k - 1$, a block is associated with in as follows

$$B_{i_1,...,i_n}(Z) = E(Z_{i_1,1}...Z_{i_n,n}). \quad (1)$$

Suppose one has given the multivariate discrete variable $(X_1,...,X_n)$. Other sequences $Z_{i,k}$ with their corresponding blocks can be constructed in a variety of ways.

Example 2.1

- Choose in (1), $Z_{i,k} = I(X_k = i)$. We obtain a block built up with the cell probabilities since $E(I(X_1 = i_1)\cdots I(X_n = i_n)) = p_{i_1,...,i_n}$. We denote it by $\mathcal{B}^p(X)$. An example is shown in Figure 1.
- Choose in (1), $Z_{i,k} = X^i_k$. We get the block built up with the moments. This block is denoted by $\mathcal{B}^\mu(X)$;
- Choose $Z_{0,k} = 1$ and $Z_{i,k} = X^i_k - EX^i_k$. We get a block built up with the central moments. We denote this block by $\mathcal{B}^\sigma(X)$.

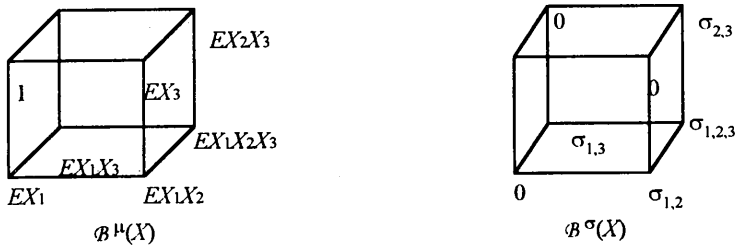


Fig. 2: Moment and central moment representation of a there dimensional binary variable

$$\sigma_{1,2,3} = E(X_1 - EX_1)(X_2 - EX_2)(X_3 - EX_3)$$

The above operators can be used to obtain formulas to write the original probabilities in terms of the new representation and viceversa.

Property 2.1 The operator to transform $\mathcal{B}^{p}(X)$ into $\mathcal{B}^{\mu}(X)$ or $\mathcal{B}^{p}(X)$ and vice versa is the flatproduct, where the flats are defined as in the following scheme:

transformation	*flat*
(*a*) $\mathcal{B}^{p}(X) \rightarrow \mathcal{B}^{\mu}(X)$	$(A^1 \vert \ldots \vert A^n)$ with $A^k \in M(r_k, r_k)$: $A^k_{i,j} = j^i$
(*b*) $\mathcal{B}^{p}(X) \rightarrow \mathcal{B}^{\sigma}(X)$	$(B^1 \vert \ldots \vert B^n)$ with $B^k \in M(r_k, r_k)$: $B^k_{i,j} = \begin{cases} 1 & i = 0 \\ j^i - EX^i_k & i \neq 0 \end{cases}$
(*c*) $\mathcal{B}^{\mu}(X) \rightarrow \mathcal{B}^{p}(X)$	$(C^1 \vert \ldots \vert C^n)$ with $C^k \in M(r_k, r_k)$: $C^k = \begin{bmatrix} 1 & -e^T Z^{r_k - 1} \\ 0 & Z^{r_k - 1} \end{bmatrix}$ and $e \in M(r_k - 1, 1)$: $e = [1, \ldots, 1]^T$, $Z^t \in M(t,t)$: $Z^t_{i,j} = \frac{(-1)^{i+t}}{(i+1)!(t-i-1)!} \sum_{k=0}^{j+1} (i+1)^{k-j-2} \begin{bmatrix} t+1 \\ k \end{bmatrix}$
$(d) B^{\sigma}(X) \rightarrow B^{p}(X)$	$(D^1 \vert \ldots \vert D^n)$ with $D^k \in M(r_k, r_k)$: $D^k_{i,j} = \begin{cases} \sum_{s=0}^{r_k - 1} EX^s_k C^k_{i,s} & j = 0 \\ C^k_{i,j} & j \neq 0 \end{cases}$

where $\begin{bmatrix} t+1 \\ k \end{bmatrix}$ represents a Stirling number of the first kind defined by the relation

$$x(x-1)\ldots(x-t)=\sum_{k=1}^{t}\begin{bmatrix} t+1 \\ k \end{bmatrix}x^{k}. \qquad (2)$$

A proof can be found in Teugels and Van Horebeek (1995).

3. BLOCKMODELS IN PRACTICE

Instead of the λ parameters of a log-linear model, we use now the (central) moments defined by $\mathcal{B}^{\mu}(X)$ resp. $\mathcal{B}^{\sigma}(X)$. For the above example, we get:

$$EX_1EX_2,\ldots,EX_1^2,\ldots,E(X_1-EX_1)(X_2-EX_2),\ldots.$$

The hypotheses can be formulated immediately in terms of those parameters, and by means of property 2.1 the corresponding likelihood equations can be written down. More examples can be found in Teugels and Van Horebeek (1995).

REFERENCES

Hagenaars, J. (1990). *Categorical Longitudinal Data.* Sage Publications.

Teugels, J.L., and Van Horebeek, J. (1995). Algebraic Descriptions of Nominal Multivariate Discrete Data. *Paper submitted for publications.*

Esta publicación consta de 410 ejemplares y se terminó de
imprimir en el mes de octubre de 1996 en los talleres gráficos del
Instituto Nacional de Estadística, Geografía e Informática
Av. Héroe de Nacozari Núm. 2301 Sur, Acceso 11, P.B.
Fracc. Jardines del Parque, CP 20270
Aguascalientes, Ags.
México