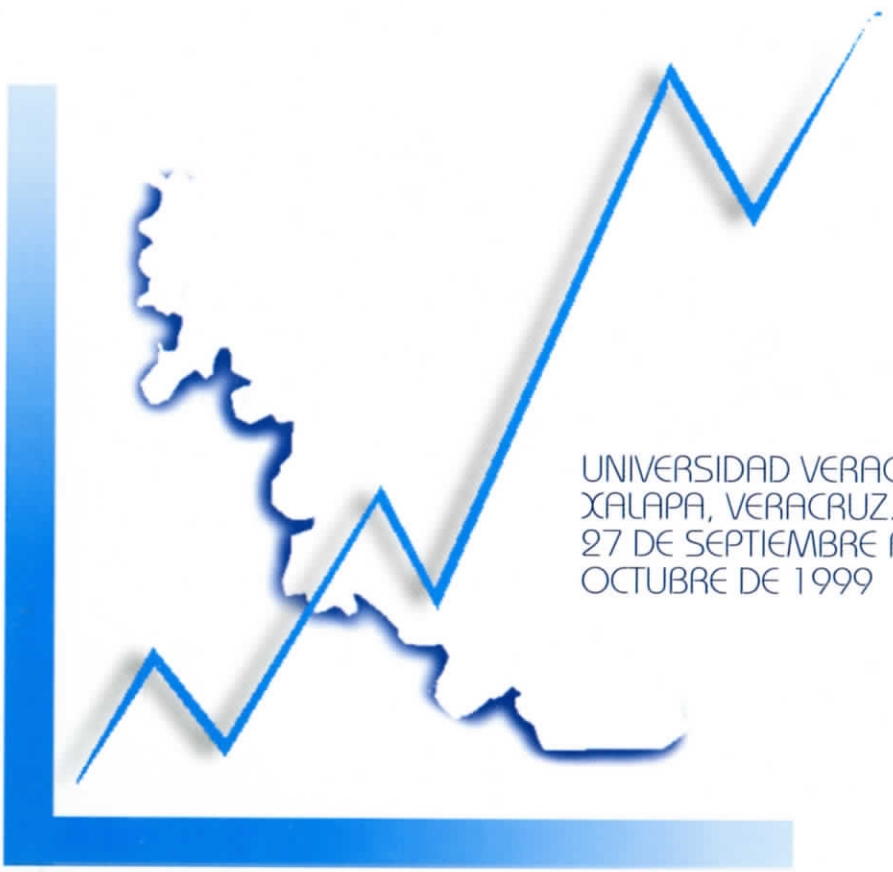
MEMORIAS
XIV FORO NACIONAL DE ESTADISTICA

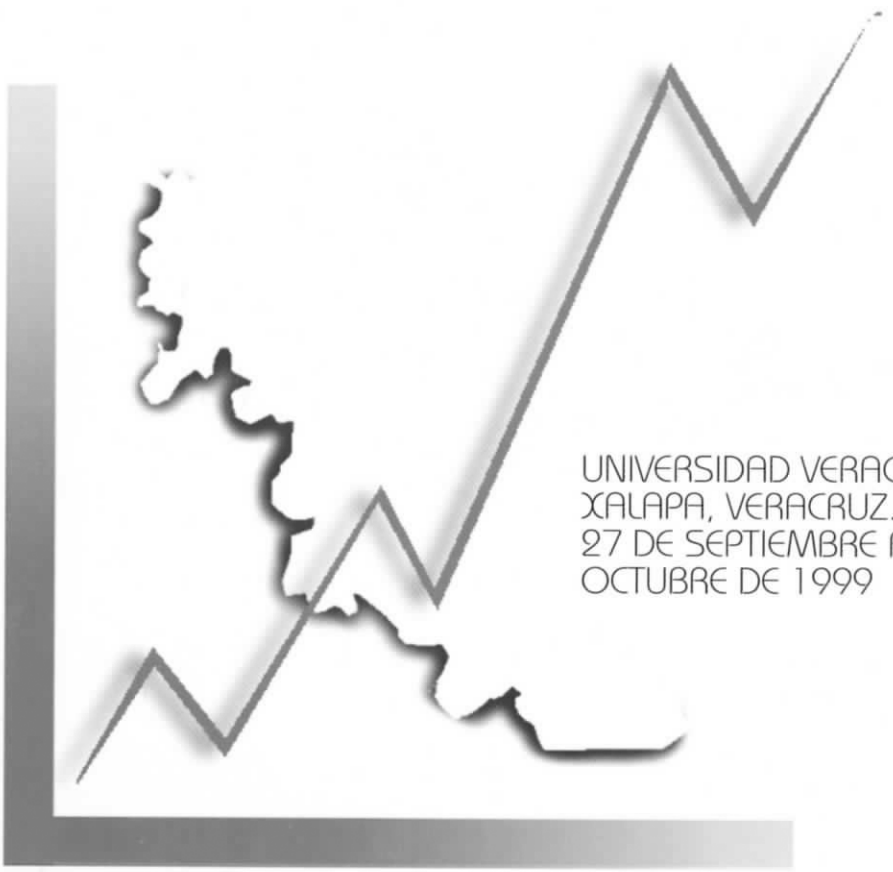


UNIVERSIDAD VERACRUZANA
XALAPA, VERACRUZ.
27 DE SEPTIEMBRE AL 1 DE
OCTUBRE DE 1999

INEGI

AME

MEMORIAS
XIV FORO NACIONAL DE ESTADISTICA



UNIVERSIDAD VERACRUZANA
XALAPA, VERACRUZ.
27 DE SEPTIEMBRE AL 1 DE
OCTUBRE DE 1999

INEGI

AME

Memorias XIV Foro Nacional de Estadística

Impreso en México
ISBN 970-13-3051-X

MEMORIA DEL XIV FORO NACIONAL DE ESTADÍSTICA

Universidad Veracruzana, Xalapa, Ver., México. 27 de Septiembre al 1 de Octubre de 1999

Resúmenes in extenso

Editado por:

José M. González-Barrios M.-*IIMAS, UNAM*

Silvia Ruiz Velasco A.-*IIMAS, UNAM y UAM Iztapalapa*

Alberto Molina Escobar-*IMUNAM, UNAM*

Presentación

El XIV Foro Nacional de Estadística se llevó a cabo del 27 de septiembre al 1o. de octubre de 1999, en la Facultad de Estadística e Informática de la Universidad Veracruzana, en Jalapa, Veracruz.

Entre otras actividades se presentaron 50 contribuciones libres y 4 conferencias magistrales. En estas memorias se presentan resúmenes de dichas contribuciones. Todos los resúmenes recibidos fueron incluidos sin un proceso de arbitraje, aunque con una detallada revisión.

La Asociación Mexicana de Estadística, agradece a la Facultad de Estadística e Informática de la Universidad Veracruzana su apoyo para la realización de este foro y al Instituto Nacional de Estadística Geografía e Informática el apoyo para la edición de estas memorias.

El Comité Editorial

José M. González B. M.

Silvia Ruiz V. A.

Alberto Molina Escobar

CONTENIDO

Modelos de Duración

Alejandro Alegría

Instituto Tecnológico Autónomo de México

1 Introducción

Los *modelos de duración* permiten analizar datos sobre la sucesión de estados que fueron ocupados por cada uno de los elementos de un conjunto, así como los tiempos en que ocurrieron estos movimientos. Los posibles estados forman un conjunto finito y deben estar bien definidos. Además, se debe contar con una regla que diga en qué estado se encuentra todo elemento en cualquier momento del tiempo.

El análisis estadístico de datos de duración también es conocido con otros nombres dependiendo del área de aplicación. Para los biólogos es *análisis de supervivencia*, porque originalmente se empleó para analizar el tiempo hasta la muerte. Los ingenieros, interesados en la vida útil de alguna maquinaria, lo llaman *análisis de confiabilidad* o de *tiempos de falla*. Los sociólogos suelen llamarlo *análisis de historia de eventos*, y los economistas *análisis de datos de duración*. El análisis econométrico de este tipo de datos es relativamente nuevo, y además plantea problemas propios de esta área de aplicación. En las siguientes secciones se presentarán algunos modelos de duración, sus características más importantes y el uso de estos modelos para interpretar datos.

2 Conceptos fundamentales y modelos

Desde el punto de vista económico es importante estudiar el tiempo que una persona o empresa permanece en un estado específico antes de abandonar dicho estado. Un ejemplo muy usado es el relacionado con la duración de períodos de desempleo. Típicamente, son dos los tipos de datos que surgen en análisis de duración. Para el primer tipo, la longitud del período de desempleo es conocida (por ejemplo, una persona encontró trabajo después de cuatro semanas). Para el segundo tipo de datos, el tiempo que se estuvo desempleado se desconoce porque al momento de realizar la observación la persona seleccionada sigue estando desempleada. Esto último da lugar a lo que se conoce como observaciones censuradas, mismas que habrá que tener en cuenta en el proceso de estimación del modelo propuesto.

El tiempo que una persona o empresa permanece en un estado específico antes de abandonar dicho estado se puede considerar una variable aleatoria. La teoría económica, mas que

interesarse por la función de densidad de esta variable, le interesa la denominada función de riesgo, así que tiene sentido elegir una especificación de la densidad de duración que produzca una función de riesgo que se comporte como uno esperaría. Esto explica porque las densidades usadas en modelos de duración ususalmente no tienen formas familiares. Ahora se formalizarán los conceptos expuestos anteriormente.

Consideremos la variable aleatoria continua T como el tiempo (duración) que transcurre desde el inicio de un evento (no tener trabajo) hasta que éste finaliza (se encontró empleo). Sean $F(t)$ y $f(t)$ las funciones de distribución y de densidad de T respectivamente. *La función de supervivencia* $S(t)$ se define como $S(t) = P(T \geq t) = 1 - F(t)$.

Ahora consideremos que una persona ha estado sin empleo durante un periodo de tiempo de longitud t. ¿Cuál es la probabilidad de que esta persona encuentre empleo durante el siguiente pequeño intervalo de tiempo de longitud ε? La respuesta nos la da la *función de riesgo*, que esta definida de la siguiente forma,

$$h(t) = \lim_{\epsilon \to 0} \frac{P\left(t < \mathbf{T} < t + \epsilon \mid \mathbf{T} \geq t\right)}{\epsilon} = \frac{f(t)}{F(t)}.$$

La función $h(t)$ es la tasa instantánea a la cual se termina un periodo de permanencia en un estado después de un tiempo t, dado que se estaba en dicho estado hasta el tiempo t.

En econometría, la función de riesgo surge en el problema conocido como selección de muestra, que no es otra cosa mas que una forma especial de truncamiento en los datos. En este caso $h(t)$ coincide con lo que se llama el *inverso de la razón de Mills*. (Manski, 1989) Como

$$h(t) = \frac{f(t)}{S(t)} = \frac{F'(t)}{S(t)} = -\frac{S'(t)}{S(t)} = \frac{\partial \ln S(t)}{\partial t} \Rightarrow S(t) = \exp\left\{-\int_0^t h(w)dw\right\}.$$

La función $H(t)$ definida como $H(t) = \int_0^t h(w)dw$, es conocida como el *riesgo integrado o acumulado* y resulta de utilidad en la práctica, sobre todo para verificar la especificación de modelos ($H(t)$ se puede ver como un residual generalizado. Ver Chesher et al. (1985), Lancaster (1990)). En términos económicos, la base teórica para la especificación de una función de riesgo es un modelo de elección óptima para los agentes (personas, empresas, etc.) cuyas transiciones van a ser estudiadas.

En la siguiente tabla se presentan algunas distribuciones usadas para modelar datos de duración.

Distribución	$f(t)$	$S(t)$	$h(t)$
Exponencial	$\lambda e^{-\lambda t}$	$e^{-\lambda t}$	λ
Weibull	$p\lambda(\lambda t)^{p-1}\exp\{-(\lambda t)^p\}$	$\exp\{-(\lambda t)^p\}$	$p\lambda(\lambda t)^{p-1}$
$Log-log\acute{i}stica$ $\ln(T)\sim \log istica$	$\frac{k\alpha t^{\alpha-1}}{(1+kt^\alpha)^2}$	$(1+kt^\alpha)^{-1}$	$\frac{(k\alpha t^{\alpha-1})}{(1+kt^\alpha)}$

Existen otras distribuciones que podrían ser de utilidad en casos específicos: log-normal, Gompertz, Gaussiana inversa, F y Gamma generalizadas, y también mezclas de cualquiera de éstas. Ver Kiefer (1988), Kalbfleisch and Prentice (1980).

Algo importante es que estas familias permiten considerar los siguientes aspectos:

1. Las distribuciones de duración de diferentes personas difieren porque, entre otras cosas, ellas enfrentan diferentes precios, tienen distintos niveles de riqueza e ingresos, etc.

2. Las anteriores fuentes de variabilidad pueden ser representadas por un vector de covariables $\mathbf{x}$, para cada persona, donde $\mathbf{x}$ puede tener componentes que deberían, de acuerdo a la teoría económica, haber sido observadas pero no lo fueron.

3. El vector de covariables $\mathbf{x}$, puede tener elementos que son funciones del tiempo calendario(s) y de la misma duración (t), es decir, $\mathbf{x}=\mathbf{x}(\mathbf{t},\mathbf{s})$. Por ejemplo, el estado civil, el tamaño de la familia y el nivel de educación, pueden cambiar durante el periodo de desempleo de una persona.

4. En los datos de duración es usual tener censura, pues las medidas se realizan en un momento dado y el proceso continúa su desarrollo.

5. El interés económico generalmente se centra en la función de riesgo como una función tanto de t como de $\mathbf{x}$. El valor esperado de la duración T, rara vez interesa, aunque su estudio, como una función de $\mathbf{x}$, podría ser de interés para la aplicación de algunas políticas.

3 Concentración de los modelos de duración

Es natural comparar las diferentes familias de distribuciones de duración en términos de sus funciones de riesgo. No obstante, hay otras formas de comparar distribuciones, y una en particular, que es de interés a los economistas, es estudiar su concentración. La concentración del ingreso o de la riqueza se puede estudiar por medio de la curva de Lorenz y del coeficiente de Gini. Algo semejante se puede hacer con la concentración de las duraciones de desempleo. Sea $f(t)$ la función de densidad correspondiente a la función de supervivencia $S(t)$ y a la distribución $F(t)$. Además, sea $\mu=E(T)$. La proporción del tiempo total de desempleo sufrido por aquellas personas cuyas duraciones no fueron mayores a t esta dada por

$$G(t) = \frac{\int_0^t w f(w) dw}{\mu}.$$

La curva de Lorenz resulta de graficar los puntos $(F(t), G(t))$, y el coeficiente de Gini, g, se define como dos veces el área entre esta curva y la linea $G = F$, así que, $g = 2\int_0^1 F dG - 1$ Como $\int_0^1 F dG = \int_0^\infty F(w) w f(w) w^{-1} dw$, es posible reescribir a g como

$$g = 1 - \frac{\int_0^\infty S^2(w) dw}{\mu} = 1 - \frac{\int_0^\infty S^2(w) dw}{\int_0^\infty S(w) dw}.$$

Un resultado interesante para la distribuciones de duración es el siguiente (ver Lancaster,1990): *Para cualquier distribución con función de riesgo decreciente (dependencia negativa de duración, se tiene que* $\mathbf{g} \geq 1/2$).

4 Variables exógenas

Hasta este momento no se han considerado factores externos que pudieran afectar la distribución de duraciones. Estas covariables ($\mathbf{x}$) se incorporan en los modelos especificando como influyen en la función de riesgo y buscando, al mismo tiempo, facilidad de cómputo. En una población homogénea de personas se supone que estas covariables tienen un valor constante. A continuación se presentan algunos de los modelos más importantes que incorporan el efecto de covariables que no cambian con el tiempo.

En el caso del *modelo exponencial*, se considera que $\lambda = \lambda(\mathbf{x})$, por ejemplo $\lambda = exp\{\beta'\mathbf{x}\}$, con β un vector de parámetros. Como $T \sim exp(\lambda)$, entonces resulta que $U = \ln(\lambda T) = \ln(\lambda) + \ln(T)$, con $e^u \sim exp(1)$.

Si $\mathbf{x}$ es invariante en el tiempo, se tiene un modelo de regresión para $\ln(T)$. Se puede generalizar el modelo anterior introduciendo una constante de proporcionalidad, p, en el término de error U, teniendo ahora $\ln(\lambda T) = U/p$, $p > 0$. Lo que se logra con este modelo es que $V(\ln(T))$ pueda tomar cualquier valor positivo, ya que $V(\ln(T)) = p^{-2}\ \Psi'(1)$, donde Ψ es la función digamma. La distribución de T resulta ser Weibull (ver Tabla en sección 2). Con $\lambda = exp\{\beta'\mathbf{x}\}$, la función de riesgo sería $h = \lambda p(\lambda t)^{p-1} = p \cdot exp\{p\beta'\mathbf{x}\} t^{p-1}$ de tal suerte que el efecto proporcional de las covariables es independiente del tiempo.

En el modelo anterior $\ln(\lambda T) = U/p$, con $e^u \sim exp(1)$, entonces, $T = exp\{U/p\}/\lambda$. De aquí surge otra generalización que da lugar a los *modelos de duración (vida) acelerada*. Sea ahora

$T = T_0/\lambda$,donde T_0 es una variable aleatoria que no depende de las covariables $\mathbf{x}$ ni del vector de parámetros β, y λ es función de $\beta'\mathbf{x}$. La duración de una persona con regresores $\mathbf{x}$ es acelerada o desacelerada con respecto a T_0 según sea $\lambda > 1$ ó $\lambda < 1$. Para este modelo se tiene que $\ln(T) = -\ln(\lambda(\beta'\mathbf{x})) + \ln(T_0)$, que no es otra cosa más que un modelo de regresión para $\ln(T)$ con error igual a $\ln(T_0)$ y el modelo es homocedástico. La regresión es lineal si λ es la función exponencial.

Otra familia importante de modelos es la de riesgos proporcionales. En este caso la función de riesgo $h = h(\mathbf{x}, t)$ es de la forma $h(\mathbf{x}, t) = k_1(\mathbf{x})k_2(t)$, donde k_1 y k_2 son funciones que no cambian para todos los individuos. El adjetivo proporcional se justifica al considerar dos personas con regresores $\mathbf{x}_1$ y $\mathbf{x}_2$. La razón $k_1(\mathbf{x}_1)/k_1(\mathbf{x}_2)$ no cambia para todo valor de t. La función $k_2(t)$ se denomina *riesgo base*. El modelo exponencial es de riesgos proporcionales, pues se obtiene al tomar $k_1(\mathbf{x}) = exp\{\beta'\mathbf{x}\}$ y $k_2(t) = 1$.

Este último tipo de modelo facilita el proceso de inferencia porque es posible estimar los parámetros de $k_1(\mathbf{x})$ (aún si los regresores varían con el tiempo) sin especificar la forma de la función k_2 (Cox, 1972).

La función de riesgo de una distribución Weibull es monótona. Una función de riesgo que admite un comportamiento no monótono es la que se obtiene al suponer que $\ln(T) \sim$ *logística* (ver tabla en sección 2). En este caso $h = (k(\mathbf{x})\alpha t^{\alpha-1} - 1)/(1 + k(\mathbf{x})t^{\alpha})$, y como $E(\ln(T)) = -\alpha^{-1}\ln(k(\mathbf{x}))$, se tiene un modelo de regresión para $\ln(T)$ si $k = exp\{\beta'\mathbf{x}\}$.

La función de riesgo de una distribución Weibull se puede escribir de la siguiente forma, $pk(\mathbf{x})t^{p-1} = pk(\mathbf{x})exp\{(p-1)\ln(t)\}$, donde $k(\mathbf{x})$ se pude tomar como $exp\{\beta'\mathbf{x}\}$. La transformación de Box-Cox tiene como caso particular la transformación logarítmica, por lo que se sugiere la siguiente generalización de la función de riesgo, $pk(x)exp\{(p-1)t^{\delta}\}$, donde t^{δ} es la transformación de Box-Cox. Este tipo de riesgos se conoce como *riesgos de Box-Cox*. Con $\delta = 0$ se tiene el caso de la distribución Weibull, y con $\delta = 1$ el de la distribución Gompertz. Cuando $\delta = 1$ se demuestra que cierta fracción de una cohorte que entra a algun estado, nunca pasará a otro estado.

5 Conclusión

El propósito de este trabajo ha sido el de promover el uso de los modelos de duración en las áreas socio-económicas. Se tomó como ejemplo la duración del desempleo, pero se tienen muchas posibilidades de aplicación: duración de huelgas, de matrimonios, tiempo

entre nacimientos, tiempo entre transacciones en los mercados financieros, durabilidad de productos, movilidad ocupacional o geográfica, tiempo hasta la quiebra de un negocio, tiempo desde el inicio hasta la resolución de cuestiones legales, tiempo entre la adquisición de bienes duraderos, tiempo de permanencia en cierto nivel ocupacional, tiempo de permanencia en estudios universitarios, etc.

Un modelo matemático demuestra su potencialidad en la medida que está sustentado por un marco teórico, y además, que permita explicar razonablemente bien la realidad. En la aplicación de los modelos de duración a fenómenos económicos o sociales, falta mucho por hacer en cuanto a la justificación teórica del uso de los mismos, esto tal vez debido a la complejidad natural de los procesos que se estudian en estas áreas.

Desde el punto de vista de la estadística, el uso de los modelos de duración presenta varios problemas, algunos ya resueltos y otros todavía en estudio. En un trabajo posterior se abordará el problema de inferencia. Es característico de estos modelos el contar con observaciones censuradas, tener errores de medición, usar variables proxy, tener observaciones faltantes, entre otras cosas. Posibles extensiones del modelo de duración expuesto son, el caso multivariado, covariables dependientes del tiempo, considerar más de dos posibles estados, incluir la información de la sucesión de estados que fueron ocupados, además del tiempo de permanencia en los mismos (datos de transición).

Referencias

Chesher, A., Lancaster, T. and Irish, M. (1985) On Detecting the Failure of Distributional Assumptions. *Annales de L'Insee*, **59/60**, 7-44.

Cox, D. (1972) Regression Models and Life Tables. *Journal of the Royal Statistical Society*, Series B, **34**, 187-220.

Kalbfleisch, J. and Prentice, R. (1980) *The Statistical Analysis of Failure Time Data.* New York: John Wiley and Sons.

Kiefer, N. (1988) Economic Duration Data and Hazard Functions. *Journal of Economic Literature*, **26**, 646-679.

Lancaster, T. (1990) *The Analysis of Transition Data.* New York: Cambridge University Press.

Manski, C. (1989) Anatomy of the Selection Problem. *Journal of Human Resources*, **24**, 343-360.

Fisher's Observed Information Matrix for Location and Scale Parameters under Type I Censoring

Karim Anaya Izquierdo
IIMAS, UNAM
Federico O'Reilly Togno
IIMAS, UNAM

1 Abstract

An expression for Fisher's observed information matrix is given under Type I censoring for any location-scale distribution under mild requirements. It is illustrated on a data set which has been analyzed by several authors.

2 Introduction

Situations where a location and scale distribution is used under censorship, seem frequent in practice and asymptotic results are helpful in providing approximations for estimation problems; specially interval estimation. These asymptotic results make explicit use of Fisher's expected information matrix or Fisher's observed information matrix. For inferences on one of the parameters, use of the profile likelihood is a common practice (see for example Kalbfleisch 1985, p. 65) and on the other hand, asymptotics allow for accurate approximations for this profile likelihood that use Fisher's observed information matrix rather than the expected information. In previous work, Escobar and Meeker (1994) provide Fisher's expected information matrix for some location-scale families and in the books by Nelson (1982) and Lawless (1982) one can find expressions for observed and expected information matrices in some particular cases. In this note, an expression is given to compute Fisher's observed information matrix for any location-scale family where the only explicit requirement is that the cdf F, the corresponding pdf f and its first two derivatives, may be numerically evaluated. For the families given in Escobar and Meeker (1994); the normal, the smallest and largest extreme value and the logistic, direct numerical computation of Fisher's observed information matrix, using the proposed expression, is straightforward; and for the location-scale cases discussed in Nelson (1982) and Lawless (1982), their formulae are easily reconstructed.

3 The problem

Suppose we have a random sample of size N (fixed in advance) from a general location-scale family of distributions, i.e. with pdf $(1/\sigma)f((x-\mu)/\sigma)$ and cdf $F((x-\mu)/\sigma)$ where $f(\cdot)$ and $F(\cdot)$ are known functional forms. Only those n ($n \leq N$) observations which lie on a pre-specified interval

$$(x_l, x_u), \qquad -\infty \leq x_l < x_u \leq \infty \tag{1}$$

are available so in essence, a Type I doubly censored (often called time censored) sample from $f(\cdot)$ is provided. Clearly, the number of order statistics in each part of the real line divided by x_l and x_u follows a multinomial distribution. Denote the k-th order statistic by $x_{(k)}$, the number of values which are less than x_l by r and the number of values which are greater than x_u by s, so the likelihood function can be written as

$$L(\mu, \sigma; r, s, \underline{x}) = \frac{N!}{r!\, s!}[F(z_l)]^r [1 - F(z_u)]^s \prod_{i=1}^{n} \frac{1}{\sigma} f(z_{(r+i)}), \tag{2}$$

where z denotes the standardized value of x, that is $z = (x-\mu)/\sigma$. The maximum likelihood estimator $(\hat{\mu}, \hat{\sigma})$ of μ and σ is the value that globally maximizes (2) and generally, closed forms for such an estimator do not exist. In the next section, the iterative procedure followed by us to compute the mle's, is sketched. When computing the asymptotic variances and covariance for the mle's, a possibility is to use Fisher's observed information matrix, the expression for this matrix is obtained by taking second derivatives of the natural logarithm of (2) and evaluating them at the mle's. This matrix can be written as:

$$I(\hat{\mu}, \hat{\sigma}) = I^*(\hat{\mu}, \hat{\sigma}) + L_c(\hat{\mu}, \hat{\sigma}) + R_c(\hat{\mu}, \hat{\sigma}), \tag{3}$$

where L_c and R_c stand for the matrices due to left and right censoring added to I^*, which is Fisher's observed information matrix in the uncensored case, as if the sample was of size n ($r = s = 0$ and $n = N$). In order to give explicit formulae let

$$C_{11}^t\left(z_{(r+i)}\right) = \left(\frac{1}{\sigma^2} \quad , \quad 0 \right)$$

$$C_{22}^t\left(z_{(r+i)}\right) = \left(\frac{z_{(r+i)}^2}{\sigma^2} \quad , \quad \frac{2\, z_{(r+i)}}{\sigma^2} \right)$$

$$C_{12}^t\left(z_{(r+i)}\right) = \left(\frac{z_{(r+i)}}{\sigma^2} \quad , \quad \frac{1}{\sigma^2} \right)$$

$$\Delta^t\left(z_{(r+i)}\right) = \left(\left[\left(\frac{f'(z_{(r+i)})}{f(z_{(r+i)})} \right)^2 - \frac{f''(z_{(r+i)})}{f(z_{(r+i)})} \right] \quad , \quad -\frac{f'(z_{(r+i)})}{f(z_{(r+i)})} \right)$$

for $i = 1, \ldots, n$ where $f'(x)$ and $f''(x)$ are the first and second derivatives with respect to x. In this way

$$I^*(\mu,\sigma) = \sum_{i=1}^{n} \begin{pmatrix} C_{11}^t\left(z_{(r+i)}\right)\Delta\left(z_{(r+i)}\right) & C_{12}^t\left(z_{(r+i)}\right)\Delta\left(z_{(r+i)}\right) \\ C_{12}^t\left(z_{(r+i)}\right)\Delta\left(z_{(r+i)}\right) & C_{22}^t\left(z_{(r+i)}\right)\Delta\left(z_{(r+i)}\right) - 1 \end{pmatrix}$$

$$L_c(\mu,\sigma) = \frac{-r}{F(z_l)} \begin{pmatrix} C_{11}^t(z_l)\underline{v}(z_l) & C_{12}^t(z_l)\underline{v}(z_l) \\ C_{12}^t(z_l)\underline{v}(z_l) & C_{22}^t(z_l)\underline{v}(z_l) \end{pmatrix} + r\left[\frac{f(z_l)}{\sigma F(z_l)}\right]^2 \begin{pmatrix} 1 & z_l \\ z_l & 1 \end{pmatrix}$$

and

$$R_c(\mu,\sigma) = \frac{s}{S(z_u)} \begin{pmatrix} C_{11}^t(z_u)\underline{v}(z_u) & C_{12}^t(z_u)\underline{v}(z_u) \\ C_{12}^t(z_u)\underline{v}(z_u) & C_{22}^t(z_u)\underline{v}(z_u) \end{pmatrix} + s\left[\frac{h(z_u)}{\sigma}\right]^2 \begin{pmatrix} 1 & z_u \\ z_u & 1 \end{pmatrix}$$

where

$$\underline{v}^t(\cdot) = (f'(\cdot) \ , \ f(\cdot)), \quad S(\cdot) = 1 - F(\cdot), \quad h(\cdot) = \frac{f(\cdot)}{S(\cdot)} .$$

Obtaining Fisher's observed matrix is a simple task which only needs expressions to compute F, f, f' and f''. As pointed out by a referee the above expressions might be also obtained following a suggestion on Escobar and Meeker (1992) p. 525. Inference about a single parameter (say μ) can be assessed through the profile likelihood for that parameter. The evaluation of this profile likelihood is done from the full likelihood (expression 2). If one relies on asymptotics, this profile likelihood for μ may be well approximated by a normal curve (see Barndorff-Nielsen and Cox (1994), p. 89 or Kalbfleisch (1985), p. 70) with mean equal to $\hat{\mu}$ and variance given by

$$\left[I(\hat{\mu},\hat{\sigma})_{(1,1)} - \frac{\left[I(\hat{\mu},\hat{\sigma})_{(2,1)}\right]^2}{I(\hat{\mu},\hat{\sigma})_{(2,2)}} \right]^{-1} ,$$

where $I(\hat{\mu},\hat{\sigma})_{(i,j)}$ is the (i,j)-th element of $I(\hat{\mu},\hat{\sigma})$. As for the profile likelihood for σ, it is well known that a normal approximation might be attempted for $\log\sigma$ (Box and Tiao (1979), suggest this in order to achieve a data translated likelihood). In this manner, the profile for $\log\sigma$ could be well approximated by a normal centered at $\log\hat{\sigma}$ and variance equal to

$$\left[I(\hat{\mu},\log\hat{\sigma})_{(2,2)} - \frac{\left[I(\hat{\mu},\log\hat{\sigma})_{(1,2)}\right]^2}{I(\hat{\mu},\log\hat{\sigma})_{(1,1)}} \right]^{-1},$$

where the matrix $I(\hat{\mu},\log\hat{\sigma})$ is obtained from $I(\hat{\mu},\hat{\sigma})$ by the following relation $I(\hat{\mu},\log\hat{\sigma}) = \widehat{D}'\, I(\hat{\mu},\hat{\sigma})\, \widehat{D}$, where $\widehat{D}$ =diag$(1,\hat{\sigma})$.

4 Obtaining the MLE'S

Closed forms for $(\hat{\mu},\hat{\sigma})$ as a function of the sample generally do not exist, so in order to obtain them, we need to follow an iterative procedure. Taking into account that most iterative procedures proposed in literature strongly depend on the initial values we suggest to use the well known linear estimates from ordinary linear regression (like Gupta's estimates for the normal censored case, see Gupta (1952)) as a first approximation:

$$\hat{\mu}_0 = \bar{z} - \hat{\sigma}_0\bar{\eta}\,; \qquad \hat{\sigma}_0 = \frac{\sum_{i=1}^{n}(\eta_{r+i}-\bar{\eta})z_{(r+i)}}{\sum_{i=1}^{n}(\eta_{r+i}-\bar{\eta})^2},$$

where

$$\bar{z} = \frac{1}{n}\sum_{k=1}^{n} z_{(r+k)}, \qquad \bar{\eta} = \frac{1}{n}\sum_{k=1}^{n}\eta_{r+k}, \qquad \eta_{r+k} = E[Z_{(r+k)}].$$

Note that these estimates do not use all the information available in the sample. They only use the data observed in the interval (1), but they provide excellent starting values. If η_{r+k} is not available one might use the known approximation

$$\eta_{r+k} \approx F^{-1}\left(\frac{r+k}{N+1}\right).$$

With these initial estimates we could begin an iterative procedure to obtain the mle's, for

example Newton-Rhapson's method. In order to obtain $\hat{\mu}$ and $\hat{\sigma}$, we need to solve

$$\begin{aligned} S_1(\hat{\mu},\hat{\sigma}) &= \left.\frac{\partial L(\mu,\sigma)}{\partial \mu}\right|_{(\hat{\mu},\hat{\sigma})} = 0 \\ S_2(\hat{\mu},\hat{\sigma}) &= \left.\frac{\partial L(\mu,\sigma)}{\partial \sigma}\right|_{(\hat{\mu},\hat{\sigma})} = 0\,. \end{aligned}$$

Let $(\hat{\mu}_0,\hat{\sigma}_0)$ be a preliminary estimate of the solution $(\hat{\mu},\hat{\sigma})$. Newton-Rhapson's method updates this estimate to $(\hat{\mu}_1,\hat{\sigma}_1)$, then this second one is updated to $(\hat{\mu}_2,\hat{\sigma}_2)$, etc. by

$$\begin{pmatrix}\hat{\mu}_j\\ \hat{\sigma}_j\end{pmatrix} \approx \begin{pmatrix}\hat{\mu}_{j-1}\\ \hat{\sigma}_{j-1}\end{pmatrix} + [I(\hat{\mu}_{j-1},\hat{\sigma}_{j-1})]^{-1} \begin{pmatrix} S_1(\hat{\mu}_{j-1},\hat{\sigma}_{j-1}) \\ S_2(\hat{\mu}_{j-1},\hat{\sigma}_{j-1}) \end{pmatrix} \qquad j=1,2,\ldots$$

as many times as necessary.

5 An example

In Meeker and Nelson (1977) data on the number of thousands of miles (T) at which different locomotive controls failed in a life test, is analyzed. The test involved 96 controls and was terminated after 135,000 miles, at which time 37 failures had occurred. The model used for the distribution of T is a Weibull whose density is

$$g(t;\alpha,\beta) = \left(\frac{\beta}{\alpha}\right)\left(\frac{t}{\alpha}\right)^{\beta-1}\exp\left[-\left(\frac{t}{\alpha}\right)^{\beta}\right]; \qquad t>0\,,$$

where α and β are the scale and shape parameters. Analyzing the data converted into natural logarithms, the smallest extreme value distribution with location $\mu = \log\alpha$ and scale $\sigma = 1/\beta$ is used. Using the procedure sketched in section 2, we obtained $\hat{\mu}_0 = 5.199$ and $\hat{\sigma}_0 = 0.436$ as initial estimates, which are remarkably close to the mle's $\hat{\mu} = 5.212$ and $\hat{\sigma} = 0.429$. Transforming these into the original parameters one has $\hat{\alpha} = 183.399$ and $\hat{\beta} = 2.331$ which are essentially the same obtained by Meeker and Nelson. In their article, the authors provide a table to obtain approximate large sample variances and the covariance for the mle's and thus to provide confidence limits for any function of the parameters. The table provided by the authors was constructed using Fisher's expected information matrix. For the locomotive control data, Meeker and Nelson calculate approximate large sample variances and the covariance for $\hat{\alpha}$ and $\hat{\beta}$, yielding $var(\hat{\alpha}) = 272.145$, $var(\hat{\beta}) = 0.1308$ and $cov(\hat{\alpha},\hat{\beta}) = -3.7109$. We did the same exercise utilizing the formulae presented in

section 2 obtaining, $var(\hat{\alpha}) = 271.851$, $var(\hat{\beta}) = 0.1303$ and $cov(\hat{\alpha}, \hat{\beta}) = -3.6913$, which are quite close. To obtain approximate confidence limits for the $100q\text{-}th$ percentile of the Weibull distribution, Meeker and Nelson (1977) assume asymptotic normality for the natural logarithm of the estimated percentile. They report $(53.7, 85.9)$ as the 95% approximate confidence interval when $q = 0.1$. On the other hand, assuming asymptotic normality for $(\hat{\mu}, \log \hat{\sigma})$ we found confidence limits for the same percentile in the extreme value scale, yielding $(4.016, 4.476)$. Translating this confidence interval into the original scale (of the Weibull model), we obtained $(55.5, 87.9)$ which differs slightly from the one reported by Meeker and Nelson. Despite the difference, these two methods are asymptotically correct. Using the same Weibull model, Nelson (1982) p. 398, analyzed the locomotive control data providing a computer output from STATPAC software (Nelson, 1982). The output gives maximum likelihood estimates and approximate 95% confidence intervals for the location and scale parameters, for population percentiles and for the proportion failing within 80,000 miles. Their software uses Fisher's observed information (also referred there as the local estimate of Fisher's information). We found (using expression (3)) exactly the same results. Also in Nelson (1982, p. 323-333) the same data is analyzed under a normal model for the base 10 logarithms of T. The author basically repeats the former analysis using both, Fisher's observed and Fisher's expected information matrix. In order to verify that Nelson's results could be reconstructed using the formulae of section 2, we made the corresponding exercise finding again total agreement in the case of the observed information matrix. To gain some insight as to the accuracy of the approximations used in providing confidence intervals, the following graphs of the actual profile likelihoods and their corresponding normal approximations were made for both models. A similar graph for the lognormal case with a different set of data appers in Meeker and Escobar (1995).

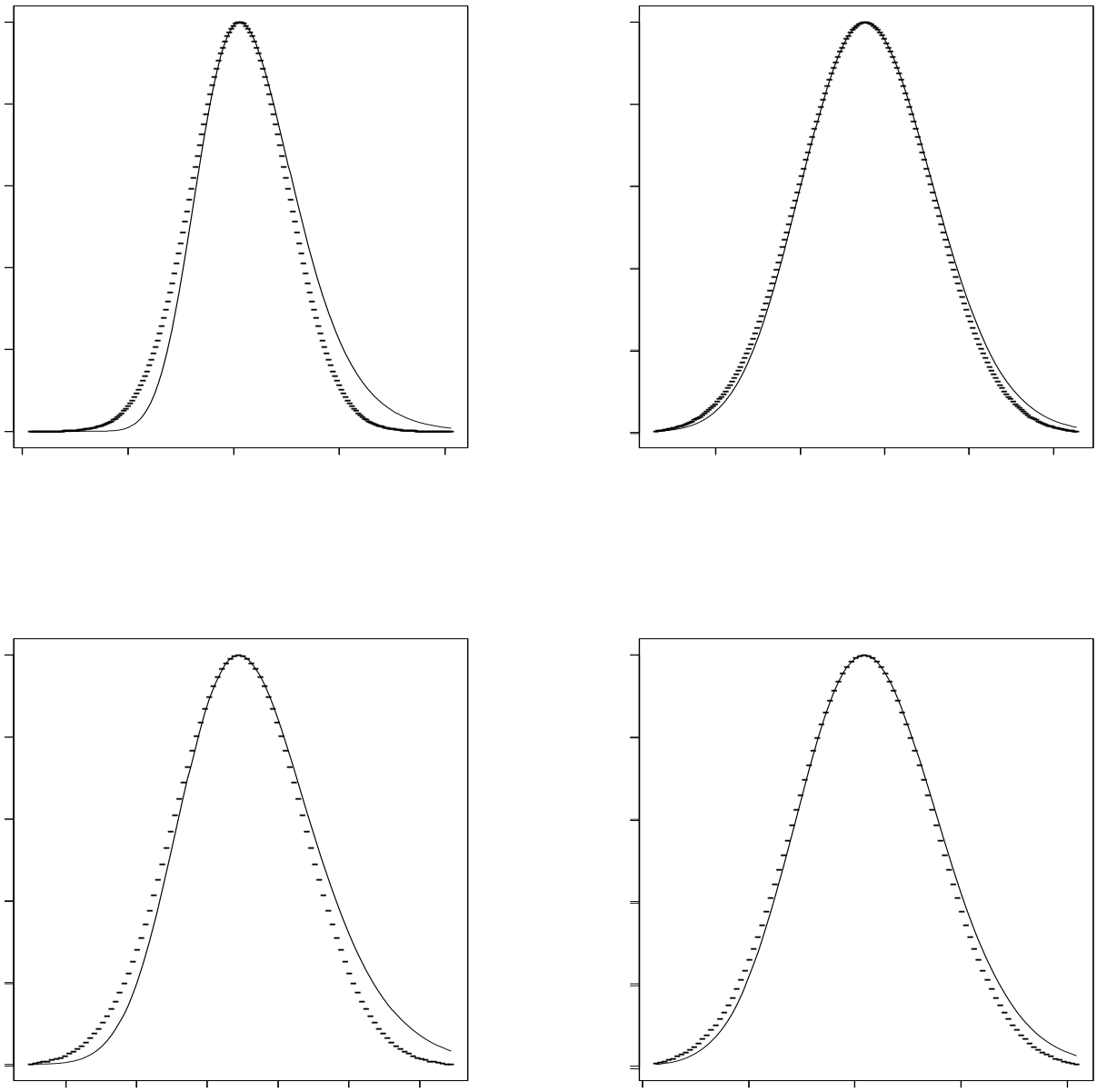


Figure 1

Approximate and exact profiles for μ and $\log \sigma$ under both models

(solid=exact; dashed=normal approximation)

6 Comments

Given that the data set on section 4 was analyzed using two different distributions, it seemed reasonable to verify in each case if there was a good fit of the assumed distribution. In goodness of fit literature there are not many tests for these distributions under censoring. Of course one might use the well known probability plots. In Meeker and Escobar (1998) p. 177, an interesting picture shows the adequate fit of both distributions in the range of the observed data and the difference in the upper values of the distribution by assuming the Weibull versus the lognormal model. If using Anderson Darling's A^2 for the lognormal test of fit (as discussed in D'Agostino and Stephens (1986), section 4.8.4) a value of A^2 =0.063 was obtained, that according to their tables is far from rejection. When testing the fit for the Weibull model, the version of Anderson Darling's A^2 based on normalized spacings gave the value A^2 =0.502 (see D'Agostino and Stephens, (1996) section 4.20.2) which according to the corresponding tables is also far from rejection, thus in accordance to the appraisal in the probability plots. In spite of the results in which both distributions are "very much accepted", inferences on the upper percentiles strongly depend on the choice of the distribution as already pointed by the probability plots. In the following table (from Nelson (1982), p. 325 and p. 398) a comparison is made of the resulting intervals with either of the two models.

Table 1

Approximate 95% confidence intervals for percentiles under both models using Fisher's observed information matrix

Percentage	Lognormal	Weibull
10	(55, 82)	(55, 88)
20	(78, 109)	(82, 113)
50	(136, 205)	(135, 182)
80	(219, 416)	(180, 280)
90	(279, 609)	(203, 339)
95	(339, 837)	(220, 391)

In this case, extreme care should be taken in trying to make inferences on characteristics of the distribution that relate to upper percentiles, since the choice of the distribution has a

huge impact on them. On the other hand, it is not surprising that the tests of fit accept both distributions since these tests use the available information which consists precisely of the lower 40% of an independent sample, where both distributions are quite similar. Trying to make inferences on the upper percentiles is essentially a questionable case of extrapolation in the sense that one is "estimating" far away from where the data lies.

References

Barndorff-Nielsen, O.E. and Cox, D.R. (1994). *Inference and Asymptotics.* Chapman and Hall.

Box, G.E. and Tiao, G.C. (1979). *Bayesian Inference in Statistical Analysis.* John Wiley and Sons, Inc.

D'Agostino, R. and Stephens, M. (1986). *Goodness of Fit Techniques* . Marcel Dekker, Inc.

Escobar, L. and Meeker, W. (1992). Assesing Local Influence in Regression Analysis with Censored Data. *Biometrics.*, **48**, 507-528.

Escobar, L. and Meeker, W. (1994). Fisher Information Matrix for the Extreme Value, Normal and Logistic Distributions and Censored Data. *Appl. Statist.*, **43**, 533-540.

Gupta, A. (1952). Estimation of the Mean and Standard Deviation of Normal Population from a Censored Sample. *Biometrika.*, **39**, 260-273.

Kalbfleisch, J. (1985). *Probability and Statistical Inference (2nd ed.).* Volume 2. Springer Verlag, New York Inc.

Lawless, J.F. (1982). *Statistical Models and Methods for Lifetime Data.* John Wiley and Sons, Inc.

Meeker, W. and Escobar, L. (1995). Teaching about Approximate Confidence Regions based on Maximum Likelihood Estimation. *The American Statistician*, **49**, 48-53.

Meeker, W. and Escobar, L. (1998). *Statistical Methods for Reliability Data.* John Wiley and Sons, Inc.

Meeker, W. and Nelson, W. (1977). Weibull Variances and Confidence Limits by Maximum Likelihood for Singly Censored Data. *Technometrics*, **19**, 473-476.

Nelson, W. (1982). *Applied Life Data Analysis.* John Wiley and Sons, Inc.

Estimación de la Densidad Espectral para Algunos Modelos de Series de Tiempo Estacionarias Usando Wavelet Packets

Alberto Contreras Cristán
IIMAS, UNAM
Andrew Walden
Imperial College of Science Technology and Medicine

1 Estimación de la densidad espectral de una serie de tiempo estacionaria de segundo orden

Sea $\{X_t\}$ un proceso estacionario de segundo orden con densidad espectral $S(f)$. Algunos métodos para estimación de $S(f)$ haciendo wavelet thresholding de un primer estimador no paramétrico basado en una muestra finita $X_1, X_2, \ldots, X_N$, han sido sugeridos primeramente por Moulin (1994) y Gao (1997). La idea básica en estos artículos es calcular el logaritmo del periodograma para esta muestra finita, aplicar la *transformada wavelet discreta* (DWT) a las ordenadas del periodograma, hacer thresholding de los coeficientes resultantes y por último invertir la transformación. El uso de la técnica de *wavelet thresholding* (Donoho and Johnstone (1995)) se justifica ya que el logaritmo del periodograma se puede escribir como una relación funcional (el logaritmo de $S(f)$ más un término de ruido) . El método descrito tiene algunos problemas como el hecho de la distribución del término de ruido no es Gaussiana además de que el periodograma puede ser un mal estimador espectral cuando $S(f)$ tiene un rango dinámico alto. Este trabajo trata de generalizar un método presentado en Walden et al. (1998) donde el periodograma se substituye por el *estimador espectral multitaper* en el algoritmo descrito arriba. La generalización consiste en usar wavelet packets para definir una transformación alternativa a la (DWT), que permita usar la teoría de Donoho y Johnstone para thresholding.

2 Estimadores espectrales multitaper

Sean U un entero positivo, $N = 2^q$, con q un entero positivo y $\widehat{S}^{(\mathrm{mt})}(f)$ el estimador multi-

taper para $S(f)$ basado en U tapers (Thomson (1982))

$$\widehat{S}^{(\mathrm{mt})}(f) \equiv \frac{1}{U}\sum_{u=0}^{U-1}\widehat{S}_u^{(\mathrm{mt})}(f), \tag{1}$$

donde

$$\widehat{S}_u^{(\mathrm{mt})}(f) = \left|\sum_{t=0}^{N-1} a_{t,u}X_t e^{-i2\pi ft}\right|^2.$$

Los tapers $\{a_{t,u} : t = 0, 1, 2, \ldots, N-1\}$ se construyen ortonormales: $\sum_t a_{t,j}a_{t,k} = 0$, si $j \neq k$ y $\sum_t a_{t,u}^2 = 1$, para cada $u = 0, 1, \ldots, U-1$.

Sean $\varphi(\cdot)$ and $\varphi^{'}(\cdot)$ las funciones digama y trigama respectivamente y sea

$$Y(f) \equiv \log \widehat{S}^{(\mathrm{mt})}(f) - \varphi(U) + \log(U),$$

Walden et al. (1998) observan que en tal caso

$$Y(f) = \log S(f) + \eta(f), \tag{2}$$

donde $\eta(f)$ es una variable aleatoria con distribución aproximadamente normal con media cero y varianza $\sigma_\eta^2 = \varphi^{'}(U)$. En el mismo artículo se establece que pese a que este modelo sería muy cercano al que se asume para la metodología de Donoho y Johnstone para wavelet thresholding, al evaluar (2) en las frecuencias de Fourier $\{f_i\}$ tendríamos

$$\mathbf{Y} = \begin{bmatrix} Y(f_0) \\ \vdots \\ Y(f_{M-1}) \end{bmatrix} = \mathbf{S} + \mathbf{N} = \begin{bmatrix} \log S(f_0) \\ \vdots \\ \log S(f_{M-1}) \end{bmatrix} + \begin{bmatrix} \eta(f_0) \\ \vdots \\ \eta(f_{M-1}) \end{bmatrix}, \tag{3}$$

donde $\mathbf{N}$ es un vector normal con matriz de covarianzas

$$\Sigma_{\mathbf{N}} \equiv \begin{bmatrix} s_\eta(f_0) & s_\eta(f_1) & \cdots & s_\eta(f_{\frac{M}{2}-1}) & s_\eta(f_{\frac{M}{2}}) & s_\eta(f_{\frac{M}{2}-1}) & \cdots & s_\eta(f_1) \\ s_\eta(f_1) & s_\eta(f_0) & \cdots & s_\eta(f_{\frac{M}{2}-2}) & s_\eta(f_{\frac{M}{2}-1}) & s_\eta(f_{\frac{M}{2}}) & \cdots & s_\eta(f_2) \\ s_\eta(f_2) & s_\eta(f_1) & \cdots & s_\eta(f_{\frac{M}{2}-3}) & s_\eta(f_{\frac{M}{2}-2}) & s_\eta(f_{\frac{M}{2}-1}) & \cdots & s_\eta(f_3) \\ \vdots & \vdots & & \vdots & \vdots & \vdots & & \vdots \\ s_\eta(f_1) & s_\eta(f_2) & \cdots & s_\eta(f_{\frac{M}{2}}) & s_\eta(f_{\frac{M}{2}-1}) & s_\eta(f_{\frac{M}{2}-2}) & \cdots & s_\eta(f_0) \end{bmatrix} \tag{4}$$

y $s_\eta(f_l)$ puede aproximarse por

$$\check{s_\eta}(\nu) = \begin{cases} \sigma_\eta^2 \left(1 - \frac{|\nu| N}{U+1}\right), & \text{if } |\nu| \leq (U+1)/N; \\ 0 & \text{otherwise.} \end{cases} \tag{5}$$

Entonces las componentes en $\mathbf{N}$ no necesariamente son no correlacionadas.

3 Transformada Wavelet Packet

Los filtros wavelet packet definen todo un sistema de transformaciones ortonormales. Para cada $j = 1, 2, \ldots, \log_2(N)$ y $n = 0, 1, \ldots, 2^j - 1$, denotemos por $\mathcal{I}_{j,n}$ al intervalo diádico $[n/2^{j+1}, (n+1)/2^{j+1})$. Coifman, Meyer y Wickerhauser (1992), establecen que toda partición $\rho = \{\mathcal{I}_{j_1,n_1}, \mathcal{I}_{j_2,n_2}, \ldots, \mathcal{I}_{j_m,n_m}\}$ del intervalo de frecuencias $[0, \frac{1}{2}]$ tiene asociada (de forma biunívoca) una transformación wavelet packet. La forma de construir los renglones o filtros pasa-banda que definen la matríz ortonormal $\mathcal{W}_\mathcal{P}$ asociada a esta transformación puede consultarse, por ejemplo, en Walden y Contreras-Cristán (1998). Aquí solamente describimos el algoritmo para seleccionar una base o transformación ortonormal, del conjunto de transformaciones conocido como wavelet packets.

Sea $\mathbf{y}$ el vector resultante de aplicar la matríz $\mathcal{W}_\mathcal{P}$ al vector aleatorio $\mathbf{Y}$ en la ecuación (3). De acuerdo con la discusión en la seccion 2, $\mathbf{y}$ tiene distribución normal multivariada con matriz de covarianzas $\Sigma_\mathbf{y} \equiv \mathcal{W}_\mathcal{P} \Sigma_\mathbf{N} \mathcal{W}_\mathcal{P}^T$. Más específicamente, sean $M_j = N/2^j$ y $\mathbf{y}_{j,n}$ el sub-vector de $\mathbf{y}$ con coeficientes wavelet packet asociados a la banda de frecuencias $\mathcal{I}_{j,n}$, entonces la distribución de $\mathbf{y}_{j,n}$ es normal multivariada con matríz de covarianzas $M_j \times M_j$ dimensional dada por

$$\Sigma_{\mathbf{y}_{j,n}} \equiv \mathcal{P}_{j,n} \Sigma_\mathbf{N} \mathcal{P}_{j,n}^T,$$

donde $\mathcal{P}_{j,n}$ es la submatriz de $\mathcal{W}_\mathcal{P}$ cuyos renglones son filtros con banda de frecuencias asociada $\mathcal{I}_{j,n}$.

Asumiendo que $\Sigma_{\mathbf{y}_{j,n}} = \sigma_{j,n}^2 \mathbf{I}_{M_j}$, donde $\mathbf{I}_{M_j}$, es la matriz identidad M_j dimensional, entonces tendríamos los ingredientes necesarios para poder usar los resultados de Donoho y Johnstone (1994) y (1995) , en particular el valor de umbral $T_{M,j,n} = \sigma_{j,n} \sqrt{(2 \log M)}$.

Lo anterior motiva el siguiente algoritmo para seleccionar una transformación.

- A nivel $j = 1$ de la transformación calculamos los coeficientes wavelet packet $\mathbf{y}_{1,n} = \mathcal{P}_{1,n}\mathbf{Y}$ $n = 0, 1$, donde $\mathbf{Y}$ está dado por la ecuación (3).

- Aplicamos una prueba de hipótesis para verificar si cada $\mathbf{y}_{1,n}, n = 0, 1$ es ruido blanco. Si no rechazamos que $\mathbf{y}_{1,n}$ es ruido blanco, entonces la banda de frecuencias $\mathcal{I}_{1,n}$ es un elemento de ρ. En caso de rechazar la hipótesis de que $\mathbf{y}_{1,n}$ sea ruido blanco, calculamos los coeficientes wavelet packet correspondientes a $\mathcal{I}_{2,2n}$ y $\mathcal{I}_{2,2n+1}$. Una vez que una banda ha sido seleccionada como elemeto de ρ, el árbol wavelet packet no se extiende más y la banda constituye el final de una rama.

- Para cada $j \geq 2$, repetimos iterativamente los pasos en el inciso anterior de esta descripción para aquellos coeficientes $\mathbf{y}_{j,n}$ que no provengan de un ancestro (banda) seleccionada como elemento de ρ.

4 Estimando la densidad espectral vía wavelet packet thresholding

En esta sección describimos un ejercicio de simulación para producir estimadores de $S(f)$ así como para estudiar propiedades de estos estimadores.
Usando filtros tipo Daubechies *mínima asimetría* con 8 coeficientes (Daubechies (1992)) repetimos los siguientes pasos 2000 veces:

a) Generamos una muestra $X_0, \ldots, X_{N-1}$ de un proceso AR(2) con N = 2048 y parámetros autoregresivos $\phi_{1,2} = 0.97\sqrt{2}$ y $\phi_{2,2} = -(0.97)^2$.

b) Se calcula el estimador multitaper $\{\widehat{S}^{(\mathrm{mt})}(f_l) : l = 0, 1, \ldots, N/2\}$ usando $U = 10$ tapers tipo sinusoidal (ver Riedel y Sidorenko (1995) para la definición de estos tapers).

c) Usando el algoritmo descrito en la sección 3, se selecciona una base ortonormal del árbol de wavelet packets para el logaritmo del estimador multitaper estandarizado. Usando un estudio de simulación anterior se determinó un nivel de significancia $\alpha = 0.01$ para la prueba de hipótesis así como un nivel máximo $j = 8$ como valores óptimos.

d) Se calcula la transformación wavelet packet (DWPT) de $\mathbf{Y}$ correspondiente a la base encontrada. Un procedimiento de thresholding se aplica a los coeficientes resultantes $\mathbf{y}$. Para este proceso se usa el valor de umbral $T_{M,j,n}$ definido en la sección 3 con

$\sigma_{j,n}$ calculada como en Walden et al. (1998), la forma de estos valores aparece en Contreras-Cristán y Walden (2000).

e) Se calcula la DWPT inversa de los coeficientes $\mathbf{y}^*$ resultantes del thresholding. El resultado de esta operación es nuestro estimador $\widehat{S}^{(wptmt)}(f_l)$ de $S(f)$

La figura 1 muestra una gráfica que describe una base ortonormal "más popular" en el sentido de que fué la más escogida a lo largo de las 2000 simulaciones. Para su elaboración se guardaron registros de las frecuencias con las que las bandas $\mathcal{I}_{j,n}$ fueron seleccionadas.
La figura 2 muestra en linea delgada la verdadera densidad $S(f)$ correspondiente al proceso AR(2) y en linea gruesa la estimación $\widehat{S}^{(wptmt)}(f)$ correspondiente a un error cuadrático medio (ECM) más cercano al ECM promedio durante las 2000 simulaciones.

5 Conclusiones

Este método de estimacion espectral puede rendir buenos resultados, no obstante queda mucho por estudiar para entender porque las estimaciones que hemos obtenido hasta el momento no superan substancialmente los resultados obtenidos por Walden et al. (1998). En particular el método se puede usar como una forma de selección de base para los wavelet packets

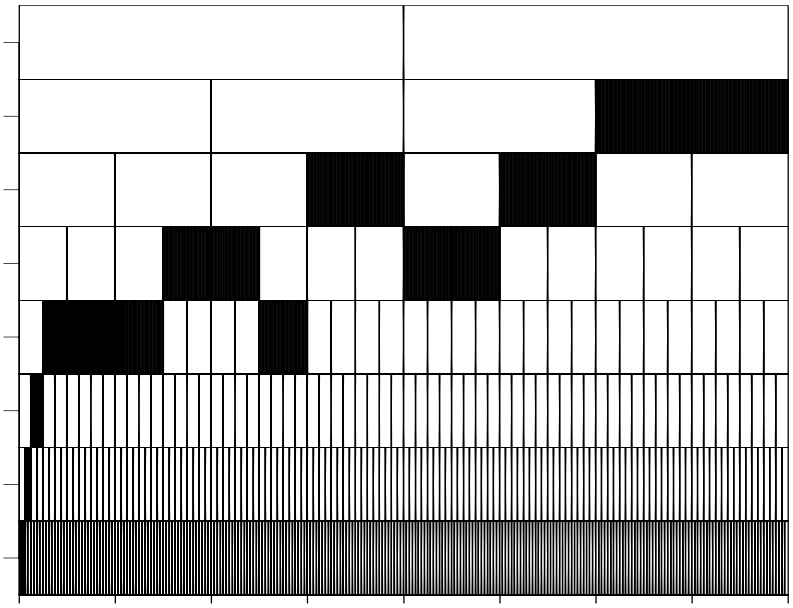


Figura 1. Base ortonormal modal para el proceso $AR(2)$ en el árbol wavelet packet.

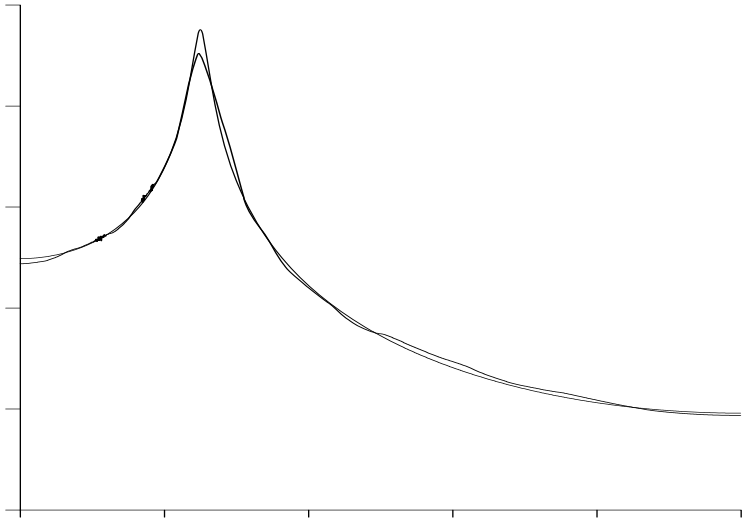


Figura 2. Estimación correspondiente a MSE más cercano al MSE promedio durante las 2000 simulaciones.

Referencias

Coifman, R., Meyer, Y. and Wickerhauser, V. (1992) Size properties of wavelet packets. In *Wavelets and their Applications,* Eds. M.B. Ruskai et al., pp. 453-470. Boston: Jones and Bartlett.

Contreras Cristán, A., Walden A.T. (2000) *Wavelet packet thresholding of multitaper spectral estimators.* En preparación.

Daubechies, I. (1992) *Ten Lectures on Wavelets.* CBMS-NSF Regional Conference Series in Applied Mathematics, SIAM, Philadelphia.

Donoho, D.L., Johnstone, I.M. (1994) Ideal spatial adaptation by wavelet shrinkage. *Biometrika,* **81**, 425-455.

Donoho, D.L., Johnstone, I.M. (1995) Adapting to unknown smoothness via wavelet shrinkage. *Journal of the American Statistical Association,* (to appear).

Gao, H. Y. (1997) Choice of thresholds for wavelet shrinkage estimate of the spectrum. *J. Time Series Analysis,* 18, (to appear).

Moulin, P. (1994) Wavelet thresholding techniques for power spectrum estimation. *IEEE Trans. Signal Processing,* **42,** 3126-3136.

Riedel, K.S. and Sidorenko, A. (1995) Minimum biased multitaper spectral estimation, *IEEE Trans. Signal Processing,* vol. **43**, pp. 188-95.

Thomson, D.J. (1982) Spectrum estimation and harmonic analysis. *Proc. IEEE,* vol. **70**, pp. 1055-96, 1982.

Walden, A.T., Contreras Cristán, A. (1998) The phase-corrected undecimated discrete wavelet packet transform and its application to determining the timing of events. *Proc. Roy. Soc. Lond. Ser.* A, **454** August, to appear.

Walden, A. T., Percival, D. B. and McCoy, E. J. (1998) Spectrum estimation by wavelet thresholding of multitaper estimators, *IEEE Transactions on Signal Processing.* To appear.

Máquinas de Vector Soporte para Clasificación

Sergio De los Cobos Silva
Universidad Autónoma Metropolitana-Iztapalapa
John Goddard
Universidad Autónoma Metropolitana-Iztapalapa
Blanca Rosa Pérez Salvador
Universidad Autónoma Metropolitana-Iztapalapa
Miguel Angel Gutiérrez Andrade
Universidad Autónoma Metropolitana-Azcapotzalco

1 Resumen

En los últimos años ha surgido una amplia variedad de trabajos en áreas muy diversas entre las que se encuentran entre otras: clasificación, reconocimiento de patrones, regresión y estimación de funciones, que utilizan la técnica llamada *Máquinas de Vector Soporte*, la cual, en general, permite encontrar máquinas de aprendizaje que generalizan bien sobre datos no previamente vistos.

El propósito de este trabajo es doble, por una parte sirve como una introducción a las máquinas de vector soporte (SVM del inglés Support Vector Machines), y por otra parte, intenta proporcionar una revisión de los recientes desarrollos en el campo.

2 Introducción

Las máquinas de vector soporte (SVM), en su forma actual fueron desarrolladas en los laboratorios AT&T-Bell por Vapnik y colaboradores (Boser et al. (1992), Guyon et al. (1993), Cortes y Vapnik (1995), Scholkopf et al. (1995), Vapnik et al. (1997)). Debido a su contexto industrial, la investigación sobre las SVM se ha orientado a aplicaciones del mundo real. Se utilizaron inicialmente en reconocimiento de caracteres ópticos con gran éxito (Scholkopf et al. (1996), (1999)). También en aplicaciones de regresión y de pronóstico con series de tiempo donde se obtuvieron excelentes resultados (Muller et al. (1997), Drucker et al. (1997), Stitson el al. (1999), Mattera y Haykin (1999)), así como en estimación de densidades y descomposición de análisis de varianza (Weston et al., 1997, Stitson et al. (1996), Scholkopf et al. (1999)). Cabe mencionar que existen excelentes tutoriales al respecto

(Stitson et al. (1996), Smola y Scholkopf (1998) y Burges (1998)).

El problema con que se enfrentaron en el desarrollo inicial de las SVM, era el de la negociación que se debería alcanzar entre el sesgo y el control de capacidad, es decir, para un trabajo de aprendizaje dado, con cierta cantidad de datos de entrenamiento, la mejor realización se obtendrá si existe un balance correcto entre los valores alcanzados a través del conjunto de entrenamiento y la "capacidad" de la máquina, es decir, la habilidad de que la máquina aprenda cualquier conjunto de entrenamiento sin error.

Las máquinas de vector soporte implementan reglas de decisión mediante una función no lineal (que mapea los puntos de entrenamiento a un *espacio característico* de mayor dimensión, donde los puntos de entrenamiento son linealmente separables).

Las máquinas de vector soporte forman una extensión del método del hiperplano de separación optimal, para tal efecto, suponga que se tiene un conjunto de entrenamiento

$\{(x_i, y_i), i = 1, 2, \ldots, l\} \subset \phi \times \mathbb{R}$, donde por ejemplo $\phi = \mathbb{R}^n$, $y_i \in \{-1, +1\}$.

Se define una máquina como un conjunto de posibles mapeos $x \to f(x, \alpha)$, donde las funciones $f(x, \alpha)$ están etiquetadas por el parámetro ajustable α. Dado un valor particular de α se genera lo que se conoce como una *máquina de entrenamiento*.

En la regresión ϵ-SV (Vapnik (1995)), el objetivo es encontrar una función $f(x)$ que esté a lo más ϵ desviaciones de los objetivos y_i para todos los datos de entrenamiento, y que a la vez sea tan "plana" como sea posible.

3 Caso lineal

Para el caso en que f sea función lineal, es decir, $f(x) =< w, x > +b$, donde $w \in \phi$, $b \in \mathbb{R}$ y $< \cdot, \cdot >$ denota el producto punto en ϕ. El hiperplano $< w, x > +b = 0$ satisface las condiciones:

$$y_i[< w, x > +b] \geq 1, \quad i = 1, 2, \ldots, l.$$

Este no es un buen hiperplano de separación para algunos de los puntos que estén muy cerca de él, puesto que pueden ser clasificados de forma diferente aunque pertenezcan a la misma clase. Por tanto, lo que se desea es tener un hiperplano que tenga siempre un margen de separación lo más grande posible. Este hiperplano de margen máximo es conocido como el hiperplano de separación óptimo.

El hiperplano de separación óptimo será entonces aquél que maximice el margen, es decir,

las distancias mínimas entre él y cualquier conjunto de datos de entrenamiento. Se puede observar que:

$$\rho(w,b) = \min_{\{x_i;y_i=1\}} \frac{< w, x_i > + b}{\|w\|} - \max_{\{x_i;y_i=-1\}} \frac{< w, x_i > + b}{\|w\|}$$

es la suma de las distancias a los dos puntos más cercanos al hiperplano de separación, la cual se desea maximizar, se obtiene:

$$\rho(w,b) = \min_{\{x_i;yi=1\}} \frac{1}{\|w\|} - \max_{\{x_i;y_i=-1\}} \frac{1}{\|w\|} = \frac{2}{\|w\|}$$

por lo que el hiperplano de separación optimal por tanto se puede expresar como un problema de optimización convexa:

$$\text{Minimizar} \quad 1/2\|w\|^2$$

$$\text{sujeto a:} \quad y_i[< w, x > + b] \geq 1, \quad i = 1, \ldots, l.$$

Mediante la formulación Lagrangeana podemos obtener el margen maximal:

$$\text{Minimizar} \quad L(w,b,\alpha_i) = 1/2\|w\|^2 - \sum_{i=1}^{l} \alpha_i \{y_i[< w, x > + b] - 1\},$$

donde α_i son multiplicadores de Lagrange positivos. Utilizando la formulación dual de Wolfe y utilizando las condiciones de Karush-Khun-Tucker se llega al problema:

$$\text{Maximizar} \quad W(\alpha) = \sum_{i=1}^{l} \alpha_i - 1/2 \sum_{i,j=1}^{l} \alpha_i \alpha_j y_i y_j < x_i, x_j >,$$

$$\text{sujeto a:} \quad \alpha_i \geq 0, i = 1, \ldots, l, \sum_{i=1}^{l} \alpha_i y_i = 0,$$

por lo que la solución es:

$$w = \sum_{i=1}^{l} \alpha_i y_i x_i.$$

A los x_i con coeficientes α_i diferentes de cero se les denominan *vectores soporte*.

Obsérvese que:

$$f(x) =< w, x >+b = \sum_{i=1}^{s} \alpha_i y_i < x_i, x >+b,$$

donde s es el número de vectores soporte, a esta expresión se le denomina *expansión de vector soporte*.

En el caso de datos no separables, se puede tener un problema no factible, por lo que, se pueden introducir variables de holgura para hacer frente a las restricciones que de otra manera son infactibles, obteniéndose:

$$y_i[< w, x >+b] \geq 1 - \xi_i$$

Para que ocurra un error, el correspondiente ξ_i debe exceder a la unidad por lo que $\sum_{i=1}^{l} \xi_i$ es una cota superior en el número de errores. De manera natural se puede asignar una penalización, por lo que se tendría (Burges (1998)) una formulación dual como:

$$\text{Maximizar} \quad W(\alpha, \beta) = \sum_{i=1}^{l} \alpha_i - 1/2 \sum_{i,j=1}^{l} \alpha_i \alpha_j y_i y_j < x_i, x_j >,$$

$$\text{sujeto a:} \quad 0 \leq \alpha_i \leq C, \quad \sum_{i=1}^{l} \alpha_i y_i = 0,$$

obteniendo la solución:

$$w = \sum_{i=1}^{s} \alpha_i y_i x_i,$$

donde s es el número de vectores soporte. La constante C determina la negociación entre lo plano de f y los errores tolerados. Observe que $f(x) =< w, x >+b = \sum_{i=1}^{s} \alpha_i y_i < x_i, x >+b$.

4 Caso no lineal

El siguiente paso es hacer el algoritmo SV (support vector, vector soporte) no lineal. Un viejo truco (Aizerman et al. (1964)) puede utilizarse. Preprocesando los patrones de entrenamiento x_i mediante un mapeo $\Phi : \phi \rightarrow \tau$, donde τ es un espacio característico de mayor dimensión que el espacio de entrada lo cual permite hacer separable el conjunto de entrenamiento (en Φ) mediante hiperplanos.

Se tiene que $w = \sum_{i=1}^{s} \alpha_i y_i x_i$ y además que $f(x) = \sum_{i=1}^{s} \alpha_i y_i < x_i, x > + b$, por lo que al preprocesar se obtendrá $w = \sum_{i=1}^{l} \alpha_i y_i \Phi(x_i)$, $f(x) = \sum_{i=1}^{s} \alpha_i y_i < \Phi(x_i), \Phi(x) > + b$, donde s es el número de vectores soporte. Este acercamiento no es factible para todo mapeo (puesto que la dimensión del espacio característico puede ser infactible computacionalmente. Lo que se utiliza para evitar esto, son funciones núcleos (kernel, en inglés) en lugar del producto punto, es decir, $k(x_i, x) = < \Phi(x_i), \Phi(x) >$ (por ejemplo, funciones núcleo que son funciones del producto punto de los x_i en ϕ) tales que el algoritmo de entrenamiento y la solución encontrada son independientes de las dimensiones tanto de ϕ como de Φ.

De manera análoga al caso lineal se obtiene un planteamiento dual (Smola y Scholkopf (1998)):

$$\text{Maximizar} \quad -1/2 \sum_{i,j=1}^{l} (\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*) < x_i, x_j > -\epsilon \sum_{i=1}^{l} (\alpha_i + \alpha_i^*) + 2 \sum_{i=1}^{l} (\alpha_i - \alpha_i^*) y_i$$

$$\text{sujeto a:} \quad \sum_{i=1}^{l} (\alpha_i - \alpha_i^*) = 0, \quad 0 \leq \alpha_i,\, \alpha_i^* \leq C, \quad i = 1, \ldots, l.$$

5 Conclusiones y nuevas líneas de investigación

Las máquinas de vector soporte, entre otras aplicaciones, proporcionan una nueva alternativa para el reconocimiento de patrones, clasificación y para estimaciones de funciones y regresión, que difiere significativamente de otras aproximaciones como por ejemplo, en el caso de redes neuronales, puesto que las VSM siempre encuentran una solucin teóricamente optimal, además de que tienen una interpretación geométrica simple, aunque la implementación no necesariamente es sencilla. Cabe mencionar que la elección y descubrimiento de nuevas funciones núcleo, así como la implementación de algoritmos para VSM y de nuevas aplicaciones, son líneas de investigación abiertas actualmente.

Referencias

Aizerman, M.A., Braverman, E. and Rozoner, L. (1964) Theoretical Foundations of the Potential Function Method in Pattern Recognition Learning. *Automation and Remote Control*, **25**, 821-837.

Boser E., Guyon I. and Vapnik V. (1992) A Trainning Algorithm for Optimal Margin Classifiers. In Hausler D., Ed, 5th *Anual ACM Workshop on COLT*, Pittsburg, 142-152.

Burges C. (1998) A Tutorial on Support Vector Machines for Patter Recognition. *Data Mining and Knowledge Discovery*, Volume 2, 1-43. Kluwer Academic Publishers, Boston.

Cortes and Vapnik V. (1995) *Support Vector Networks*, M. Learning 20:273-297.

Drucker H., Burges C., Kaufman L., Smola A. and Vapnik V. (1997) Support Vector Regression Machines, Mozer, Jordan and Petsche eds., *Advances in Neural Information Processing Systems*, Cambridge, 155-161.

Guyon, Boser and Vapnik V. (1993) Automatic Capacity Tuning of Very Large VC-dimension Classifiers, Hanson, Cowan and Giles eds., *Advances in Neural Information Processing Systems*, Morgan Kaufmann, 147-155.

Mattera D. and Haykin S. (1999) Support Vector Machines for Dynamic Reconstruction of a Chaotic System, Scholkopf, Burges and Smola eds., *Advances in Kernel Methods-Support Vector Learning*, Cambridge, 211-242.

Muller K., Smola A., Ratsh G., Scholkopf B., Kohlmorgen J. and Vapnik V. (1997) Predicting Time Series with Support Vector Machines, Gerstner, Germond, Hasler and Nicoud eds., *ICANN-97*, Berlin, 999-1004.

Scholkopf B., Bartlett P., Smola A. and Williamson R. (1995) Support Vector Regression with Automatic Acurrancy Control, Fayyad and Uthurusamy eds., *First International Conference on Knowledge Discovery & Data Minning*, Menlo Park.

Scholkopf B., Burges C. and Vapnik V. (1996) Incorporating Invariances in Support Vector Learning Machines, Von der Malsburg, Von Seelen, Vorburggen and Sendhoff eds., *ICANN-96*, Berlin, 47-52.

Scholkopf B., Mika S., Burges C., Knirsch P., Muller K., Ratsh G. and Smola A. (1999) Input Space vs. Feature Space in Kernel-Based Methods, *IEEE Transactions on Neural Networks*, (in press).

Smola A. and Scholkopf B. (1998) A Tutorial on Support Vector Regression, *Neuro COLT2 Technical Report Series.*

Stitson M., Weston J., Gammerman A., Vork V. and Vapnik V. (1996) Theory of Support Vector Machines, *Tech. Report CSD-TR-96-17, Royal Holloway College.*

Stitson M., Gammerman A., Vapnik V., Vovk V., Watkins C. and Weston J. (1999) Support Vector Regression with ANOVA Descomposition Kernels. Scholkopf, Burges and Smola eds., Advances in Kernel Methods-Support Vector Learning, Cambridge, 285-292.

Vapnik V. (1995) *The Nature of Stastistical Learning Theory*, Springer-Verlag, New York.

Vapnik V., Golowich S. and Smola A. (1997) Support Vector Method for Function Approximation, Regression Estimation, and Signal Processing, in Mozer, Jordan & Petche eds., *Advances in Neural Information Processing Systems* **9**, MIT Press, 281-287.

Westorn J., Gammerman A., Stitson M., Vapnik V. and Watkins C. (1997) Support Vector Density Estimation. Scholkopf, Burges and Smola eds., *Advances in Kernel Methods-Support Vector Learning*, Cambridge, 293-306.

Specification Testing of Conditional Quantile Functions

Miguel A. Delgado
Universidad Carlos III de Madrid
Manuel A. Domínguez
Instituto Tecnolgico Autnomo de México

1 Introduction

Let $(Y, X')'$ be a $\mathbb{R} \times \mathbb{R}^d$ -valued random vector and $F(\cdot \mid x)$ the conditional distribution of Y given $X = x$. For any $\theta \in (0,1)$, we are interested in testing if he θth conditional quantile of Y given X can be represented by the a parametric specification $m(x, \beta_0)$, for some $\beta_0 \in B \subset \mathbb{R}^b$. Regression quantiles were introduced by Koenker and Bassett (1978). A quantile regression model is usually written as

$$Y = m(X, \beta_0) + \varepsilon(\beta_0),$$

where $\varepsilon(\beta_0)$ is the quantile regression error. When $F(\cdot \mid x)$ is continuous along $y = m(x, \beta_0)$, the null hypothesis imposes that $\Pr(\varepsilon(\beta_0) \leq 0 \mid X) = \theta$ $a.s.$. Hence,

$$H_0 : E[1(\varepsilon(\beta_0) \leq 0) - \theta \mid X] = 0 \ a.s. \text{ some } \beta_0 \in B, \tag{1}$$

where $1(A)$ is the indicator function of the event A. The alternative hypothesis is the negation of the null.

Zheng (1998) proposed a consistent test statistic converging to a standard normal under H_0, but it requires to use smoothers. Our test statistic is inspired by the traditional goodness-of-fit tests. These tests were adapted to testing conditional mean specification by Brunk (1970) and extended by Stute (1997) and references therein. This type of tests statistics are not distribution free and are implemented using bootstrap. The main advantage is that they do not require any type of smoothing and are able to detect Pitman's local alternatives converging to H_0 at the rate $n^{-1/2}$.

Finally, note that under H_0, $T(x) = E([1(\varepsilon(\beta_0) \leq 0) - \theta] 1(X \leq x)) = 0$ for every x.

2 Testing procedure

Let $\left\{(Y_i, X_i')', i = 1, .., n\right\}$ be independent observations of $(Y, X')'$ and $\hat{\beta}_n$ some reasonable estimator of β_0. Define $\varepsilon_i(\beta) = Y_i - m(X_i; \beta)$. Then, $T(x)$ can be estimated by

$$T_n(x) = \frac{1}{n}\sum_{i=1}^{n}\left[1\left(\varepsilon_i\left(\hat{\beta}_n\right) \leq 0\right) - \theta\right] 1(X_i \leq x).$$

A functional of T_n can be used as test statistic. For instance,

$$K_n = \max_{\{i=1,2,..,n\}}\left|\sqrt{n}T_n(X_i)\right|, \text{ or } C_n = \sum_{i=1}^{n} T_n(X_i)^2,$$

that resemble Kolmogorov-Smirnov and Crámer-v.Mises statistic respectively.
In addition to H_0 and H_1, we also consider Pitman's contiguous alternatives of the form

$$H_{1n} : F(m(X, \beta_0) + \ell_n s(X) | X) = \theta \ a.s. \text{ for some } \beta_0 \in B,$$

where $\ell_n = O\left(n^{1/2}\right)$ and $s : \mathbb{R}^d \to \mathbb{R}$ is such that $\Pr(s(X) \neq 0) > 0$. Such contiguous alternatives have been also considered by Zheng (1998), in this context, with $\ell_n = n^{-1/2} h^{-d/4} \to 0$ and $h \to 0$ as $n \to \infty$. Hence, according to the next result, our test is infinitely more efficient than Zheng's test under this type of contiguous alternatives. Define $S(x) = E[f(m(X, \beta_0) \mid X) s(X) 1(X \leq x)]$ and let G denote the distribution function of X.
Hereforth, we impose the usual smoothness assumptions on m and F, see e.g. Koenker and Bassett (1978, 1982), Pollard (1991) and Weiss (1994).

Theorem 2.1: *Under* H_0,

$$K_n \xrightarrow{d} K_\infty = \sup_{x \in \mathbb{R}^d} |T_\infty(x)| \text{ and } C_n \xrightarrow{d} C_\infty = \int T_\infty(x)^2 dG(x),$$

where T_∞ *is a Gaussian process centered at zero and with covariance structure,*

$$\Sigma(x_1, x_2) = E\left\{A(x_1, \beta_0) A(x_2, \beta_0)'\right\}, \tag{2}$$

$A(x, \beta) = [1(\varepsilon(\beta) \leq 0) - \theta] 1(X \leq x) + E\{f(m(X, \beta) \mid X) \dot{m}(X, \beta)' 1(X \leq x)\}' p(Y, X, \beta)$, *$f$ and $\dot{m}$ are the derivatives of F and m respectively and p is a centered, finite variance random variable depending on the first order asymptotic expansion of* $\hat{\beta}_n$.

Under H_{1n},

$$K_n \xrightarrow{d} \sup_{x \in I\!R^d} |T_\infty(x) + S(x)| \text{ and } C_n \xrightarrow{d} \int [T_\infty(x) + S(x)]^2 \, dG(x).$$

Finally under H_1, *with* β_0 *replaced by some other pseudoparameter value* β_1,

$$K_n \xrightarrow{p} \infty \text{ and } C_n \xrightarrow{p} \infty.$$

3 Bootstrap tests

Henceforth, we use standard bootstrap notation. In the paper it is shown that the residual based naive and wild bootstrap are inconsistent

The first bootstrap test is constructed from a "wild" resample of estimates of $A(x, \beta_0)$. Let $\hat{f}(\cdot \mid \cdot)$ be a uniformly consistent nonparametric estimate of $f(\cdot \mid \cdot)$ along $y = m(x, \beta_0)$, see Prakasa Rao (1983). The n posible estimates of $A(x, \beta_0)$ are

$$\begin{aligned} \hat{A}_i(x) &= \left[1(\varepsilon_i\left(\hat{\beta}_n\right) \leq 0) - \theta\right] 1(X_i \leq x) \\ &\quad + \left(n^{-1} \sum_{i=1}^{n} \hat{f}\left(m\left(X_i, \hat{\beta}_n\right) \mid X_i\right) \dot{m}\left(X_i, \hat{\beta}_n\right)' 1(X \leq x)\right)' \hat{p}_i\left(\hat{\beta}_n\right), \end{aligned}$$

and $\hat{p}_i$ is the feasible version of p_i. The "wild" bootstrap requires using a sequence of random variables such that A.4 holds. Then, the bootstrap statistics are

$$K_n^{A*} = \max_{\{i=1,2,..,n\}} \left|\sqrt{n} T_n^{A*}(X_i)\right| \text{ and } C_n^{A*} = \sum_{i=1}^{n} T_n^{A*}(X_i)^2,$$

where

$$T_n^{A*}(x) = \frac{1}{n} \sum_{i=1}^{n} \hat{A}_i(x) V_i$$

and $\{V_i\}$ is a sequence of independent bounded random variables such that $E(V_i) = 0$ and $E(V_i^2) = 1$ $i = 1, 2, ..., n$. This way of estimating the asymptotic distribution of tests statistics has been proposed by Su and Wei (1991) and Hansen (1996) among others.

Theorem 3.1: *Under* H_0 *or* H_{1n},

$$K_n^{A*} \xrightarrow{d^*} K_\infty \text{ and } C_n^{A*} \xrightarrow{d^*} C_\infty, \text{ in probability,}$$

under H_1, substituting β_0 by some other pseudoparameter value β_1,

$$K_n^{A*} \xrightarrow{d^*} \sup_{x \in I\!R^d} \left| T_\infty^1(x) \right| \text{ and } C_n^{A*} \xrightarrow{d^*} \int T_\infty^1(x)^2 \, dG(x), \text{ in probability,}$$

where T_∞^1 is a Gaussian process centered at zero and with covariance structure (2), with β_0 substituted by β_1.

The second bootstrap test can be described as follows: obtain the $\{(Y_i^{N*}, (X_i^{N*})')' : i = 1, ..., n, \}$ by sampling with replacement the vectors $\{(Y_i, X_i')' : i = 1, ..., n\}$. The bootstrap analog of $\hat{\beta}_n$ is given by $\hat{\beta}_n^{N*}$, see Hahn (1995) or Horowitz (1998).

Define

$$T_n^{N*}(x) = \frac{1}{n} \sum_{i=1}^n 1(\varepsilon_i^{N*}\left(\hat{\beta}_n^{N*}\right) \leq 0) 1\left(X_i^{N*} \leq x\right) - \frac{1}{n} \sum_{i=1}^n 1(\varepsilon_i\left(\hat{\beta}_n\right) \leq 0) 1\left(X_i \leq x\right),$$

i.e., we center the bootstrap process $n^{-1} \sum_{i=1}^n 1(\varepsilon_i^{N*}\left(\hat{\beta}_n^{N*}\right) \leq 0) 1\left(X_i^{N*} \leq x\right)$. The bootstrap statistics are

$$K_n^{N*} = \max_{\{i=1,2,..,n\}} \left| \sqrt{n} T_n^{N*}(X_i) \right| \text{ and } C_n^{N*} = \sum_{i=1}^n T_n^{N*}(X_i)^2.$$

Theorem 3.2: *Under H_0 or H_{1n},*

$$K_n^{N*} \xrightarrow{d^*} K_\infty \text{ and } C_n^{N*} \xrightarrow{d^*} C_\infty, \text{ in probability.}$$

Under H_1, substituting β_0 by some other pseudoparameter value β_1

$$K_n^{N*} \xrightarrow{d^*} \sup_{x \in I\!R^d} \left| T_\infty^1(x) \right| \text{ and } C_n^{N*} \xrightarrow{d^*} \int T_\infty^1(x)^2 \, dG(x), \text{ in probability,}$$

where T_∞^1 is a Gaussian process centered at zero and with covariance structure (2), with β_0 substituted by β_1.

References

Brunk, H. D. (1970) Estimation for isotonic regression, in *Nonparametric Techniques in Statistical Inference*, Ed. M.L. Puri, 177-197, Cambridge: Cambridge University Press. Vol. 88, 1310-1316.

Hahn, J. (1995) Bootstrapping quantile regression estimators. *Econometric Theory*, Vol. 11, 105-121.

Hansen, B. (1996) Inference when a nuisance parameter is not identified under the null hypothesis. *Econometrica*, Vol. 64, 413-430.

Horowitz, J.L. (1998) Bootstrap methods for median regression models. *Econometrica*, Vol. 66, 1327-1351.

Koenker, R. and G. Basset (1978) Regression quantiles. *Econometrica*, Vol. 46, 33-50.

Koenker, R. and G. Basset (1982) Robust test for heteroscedasticity based on regression quantiles. *Econometrica*, Vol. 50, 43-61.

Pollard, D. (1991) Asymptotics for least absolute deviations regression estimators. *Econometric Theory*, 7, 186-199.

Prakasa Rao, B.L.S. (1983) *Nonparametric functional estimation*, Londres: Academic Press.

Stute, W. (1997) Nonparametric model checks for regression. *Annals of Statistics,* Vol. 25, 613-641.

Su, J.Q. and L.J. Wei (1991) A lack of fit test for the mean function in a generalized linear model. *Journal of the American Statistical Association*, Vol. 86, 420-426.

Weiss, A. (1994) Estimating nonlinear dynamic models using least absolute error estimation. *Econometric Theory*, Vol. 7, 46-58.

Zheng, X. (1998) A consistent nonparametric test of parametric regression models under conditional quantile restrictions. *Econometric Theory,* Vol. 14, 123-138.

Mejora Continua y Optimización de Procesos.

Jorge Domínguez Domínguez

Centro de Investigación en Matemáticas

1 Introducción

Mejorar es un reto que se desea alcanzar en las distintas actividades y proyectos que se realizan todos los días, en particular cuando se trata de procesos industriales. La estadística desempeña un papel importante para alcanzar ese objetivo, ya que permite evaluar la existencia y significancia de una mejora, y en la planeación de la estrategia que se puede seguir para mejorar las características de un proceso.

En esta presentación nos enfocaremos a la mejora que se lleva a cabo dentro de un proceso industrial, hacemos esta acotación debido a la dinámica que ha adquirido la competencia del mercado en bienes y servicios. No sólo la mejora sino la mejora continua juega un papel importante en los medios de producción

Partimos del hecho de reconocer que la variación está alrededor de nosotros y presente en cada cosa que hacemos. Todo trabajo es una serie de procesos interconectados y la identificación, la caracterización, la cuantificación, el control y la reducción de la varianza, ofrecen la oportunidad de mejorar.

En los procesos industriales se observan los cambios y las mejoras mediante las cartas de control. Otras medidas que permiten evaluar la eficiencia de un proceso y las mejoras que sobre éste se hagan, son los índices de capacidad del proceso. Estos índices se denotan por C_p y C_{pk} y se definen por:

$$C_p = \min\{CPI, CPS\} \quad \text{donde } CPI = \frac{\overline{\overline{X}} - LEI}{3\hat{\sigma}} \qquad CPS = \frac{LES - \overline{\overline{X}}}{3\hat{\sigma}}$$

$$C_{pk} = C_p(1-k) \quad \text{donde } k = \frac{2\left|T - \overline{\overline{X}}\right|}{LES - LEI},$$

donde p está asociado al proceso, k es un índice de centralidad, *LES, LEI* son límites de especificación inferior y superior repectivamente, $\overline{\overline{X}}$ y $\hat{\sigma}$ son la media y la desviación estandar de la respuesta que mide la característica de calidad de interés, y T es un valor objetivo, es

decir, la característica ideal que debe tener un producto. Entre las mejoras de un proceso está el alcanzar un valor objetivo generalmente referido como target: T y la variabilidad alrededor de ese valor objetivo se desea mínima. Como se observa un proceso mejora cuando está centrado ($k \to 0$) o si la variabilidad disminuye, la finalidad de la mejora es centrar el proceso con la menor variabilidad.

Por lo anterior se puede decir que la mejora en un proceso industrial está altamente relacionada con la distancia entre el valor promedio de la respuesta y un valor establecido (T), y con la variabilidad. Así, el ideal será optimizar conjuntamente la media y la varianza, en el contexto de la metodología de superficie de respuesta (MSR) se han desarrollado varios métodos para alcanzar tal fin.

En este trabajo se presentará una breve descripción de los procedimientos recientes en optimización estadísticas que son de utilidad para la mejora continua en los procesos, además se muestra un procedimiento alternativo a los propuestos. Mediante un ejemplo se hace una comparación de los resultados que arrojan cada uno de los procedimientos. También se indica la importancia del método gráfico para obtener valores óptimos. Una discusión con mayor amplitud de estos resultados se reporta en Domínguez (2000).

2 Optimización y los procedimientos

Ideas generales. Las curvas de nivel son un método gráfico apropiado para estudiar la mejora continua, ya que permiten graficar curvas de nivel tanto para la media como para la varianza, luego se sobreponen las curvas. Con esta sobreposición, se crean varios escenarios con el propósito de encontrar un valor óptimo promedio cercano al valor objetivo y para mínimizar la varianza. A parte del método gráfico se han propuesto varias técnicas para la estimación conjunta de la media y varianza, en este apartado se describirán estos procedimientos.

Modelos. En un proceso se establecen los factores que son importantes para explicar las respuestas, dentro del procedimiento clásico de la MSR, un objetivo es ajustar un modelo a los datos que se obtienen al realizar el trabajo experimental, posteriormente se aplican técnicas de optimización al modelo ajustado. Por lo general, un modelo de segundo orden es el que se propone en la etapa final de la experimentación y se expresa por:

$$y = \beta_0 + x\beta + x'\Phi x + \varepsilon$$

donde $\beta' = (\beta_0, \beta_1, ..., \beta_k)$ es un vector de parámetros, x es una matriz $nx(k+1)$ donde n son los tratamientos por el número de réplicas y k son los factores del proceso, cada renglón de la matriz x es un vector y se expresa por $x_1 = (1, x_{i1}, ..., x_{ik})$ que corresponde al $i-ésimo$ tratamiento, el 1 representa al término constante $(i = 1, ..., n)$. Φ es una matriz simétrica de orden k, en la diagonal aparecen los parámetros que corresponde al efecto cuadrático y fuera de la diagonal están los parámetros que muestran el efecto de interacción. El vector y representa a las observaciones y ε es un vector aleatorio, el cual tiene una distribución de probabilidad normal.

Cuando en un experimento se realizan réplicas, se pueden ajustar dos modelos, uno para la media y otro para la variabilidad, los modelos estimados son :

$$\widehat{\mu}(x) = \widehat{\beta}_0 + x\widehat{\beta}' + x'Bx \quad \text{y} \quad \widehat{\sigma}(x) = \widehat{\alpha}_0 + x\widehat{\alpha}' + x'Ax$$

respectivamente para la media y la desviación estándar.

Optimización. El planteamiento general de optimización para la desviación estándar como función objetivo es :

optimizar :	$\widehat{\sigma}(x)$
(1) sujeto a :	$\widehat{\mu}(x) = T$
	$x \in R(x)$
(2) o	$\widehat{\mu}(x) > T$
(3)	$\widehat{\mu}(x) < T$

En la función objetivo se puede plantear la optimización para la media, es decir :

optimizar :	$\widehat{\mu}(x)$
(1) sujeto a :	$\widehat{\sigma}(x) = L$
	$x \in R(x)$

Ejemplo. Box-Draper (1987) p. 247, presentan un ejemplo sobre la capacidad de una imprenta para imprimir tinta de color en unas etiquetas. Se considera que tres factores en tres niveles tienen efecto en la impresión de la tinta, estos son:

factores	niveles	1	2	3
x_1 : velocidad		30	45	60
x_2 : presión		90	110	130
x_3 : distancia		12	20	28

Este diseño ha servido como referencia a diferentes autores para ilustrar los resultados que se obtienen al aplicar el método que proponen en la optimización de la media y la variabilidad, luego los compararan con los resutados obtenidos por otros autores. Aquí los usamos para hacer una comparación global de los procedimientos de optimización conjunta tratados en este trabajo. El diseño es un factorial completo 3^3. En un caso, la intención es minimizar la variabilidad, restringida a la media y a la región experimental.

minimizar :	$\widehat{\sigma}(x)$
(1) sujeto a :	$\widehat{\mu}(x) = 500$
	$x \in R(x)$

Modelos para $\widehat{\sigma}(x)$ y $\widehat{\mu}(x)$. Los modelos que se obtienen al ajustar la media y la desviación estándar son:

$$\begin{aligned}\widehat{\mu}(x) = \quad & 327.6 \quad +177.0x_1 \quad +109.4x_2 \quad +131.5x_3 + 32.0x_1^2 - 22.4x_2^2 \\ & \quad -29.1x_3^2 \quad +66.0x_1x_2 \quad +75.5x_1x_3 + 43.6x_2x_3\end{aligned}$$

$$\begin{aligned}\widehat{\sigma}(x) = \quad & 34.9 \quad +11.5x_1 \quad +15.3x_2 \quad +29.2x_3 + 4.2x_1^2 - 1.3x_2^2 \\ & \quad +16.8x_3^2 \quad +7.7x_1x_2 \quad +5.1x_1x_3 + 14.1x_2x_3\end{aligned}$$

Lista de procedimientos. El objetivo fundamental de cada uno de los siguientes procedimientos es encontrar una mejor respuesta común para la media y la variabilidad, por lo general se plantea optimizar una de ellas sujeta a ciertas características de la otra. En la secuencia del tiempo cada uno de los autores va diciendo que su método mejora al otro en algunos aspectos. En resumen resaltamos la característica de cada uno de ellos y al final pondremos una tabla comparativa para observar la cercanía de los resultados.

Los procedimientos son: *Vining - Myers* (Vining and Myers (1990)): Usan la respuesta dual. *Del Castillo - Montgomery* (Del Castillo and Montgomery (1993)): Usan técnicas estándar de programación no-líneal, este procedimiento es alternativo a la respuesta dual y consiste en el algoritmo Gradiente Generalizado Reducido. *Lin - Tu* (Lin and tu (1995)): Ellos señalan que el esquema de optimización propuesto por VM puede ser engañoso debido a la restricción que se le imponga a una de las respuestas. El planteamiento que proponen consiste en optimizar el Error Cuadrático Medio ECM. *Copeland - Nelson* (Copeland and Nelson (1996)): Precisan el problema de optimización del ECM poniendo una cota a la distancia entre la media y el

valor objetivo. *Kim - Lin* (Kim and Lin 1998): Una aproximación a la modelación usando conjuntos borrosos. Estas funciones pueden escribirse en las restricciones, o como una función objetivo. *Vining - Bohn* (Vining and Bohn (1998)): Una aproximación no paramétrica y semiparamétrica para la estimación conjunta. La idea es optimizar la función propuesta por Lin - Tu : $ECM(y)$ usando el *kernel* de regresión no paramétrico para estimar la varianza y un kernel de regresión no paramétrico separado del anterior para estimar la respuesta $\widehat{\mu}(x)$. Finalmente minimizan $ECM(y)$ mediante el método simplex.

Optimización usando curvas de nivel (CNi): 1.- Se establecen en el plano el conjunto de restricciones. En el caso de la respuesta dual, la región queda determinada por un modelo ajustado de segundo orden y la región experimental. 2.- Se sobrepone la función objetivo en la región de restricciones. Donde la función objetivo es el modelo de segundo orden el cual se quiere optimizar.

Modelo ponderado **(MP)**. Una de las ideas planteadas desde algún tiempo atrás, es la regresión en dos etapas. El propósito de esta estrategia es disminuir el impacto de la variabilidad en la estimación de un modelo de regresión, ver Harvey (1976) y Aitkin (1987). Apoyados en estas ideas, proponemos la siguiente estrategia para la optimización conjunta:

Etapa 1.- Ajustar el modelo $\mu(x)$
Etapa 2.- Ajustar el modelo $y(x) = \log(y - \widehat{\mu}(x))^2$

En este proceso usamos la idea de Chan-Mak (1995), en la cual se descompone el espacio de factores en $x = (X_1, X_2)$, donde X_1 describirá los factores que son significativos en la Etapa 2, es decir, afectan la variabilidad, X_2 se refiere a los factores que son importantes en la Etapa 1.

El procedimiento algorítmico es : 1.- Partir de los supuestos clásicos. 2.- Identificar los X_1 que afectan $y(x)$, ($y(x)$ modelo de segundo orden). 3.- Optimizar $\widehat{y}(X_1)$. 4.- Consideramos los pesos $w_i = \exp(\widehat{y}(X_1))$. 5.- Se reajusta el modelo $\mu(x) = x\beta + \varepsilon$, entonces $\widehat{\beta} = (x'wx)^{-1}x'wy$. 6.- Se ajusta el modelo $\widehat{\mu}_w(X_2)$ tal que $\widehat{\mu}_w(X_2)$ se aproxime a T. (6.1. Por análisis de coordillera, ó 6.2. Curvas de nivel).

Solución. Se aplica el algoritmo anterior al ejemplo de Box-Draper:

1. Se calculan los residuales $(y - \widehat{\mu}(x))$ del ajuste al modelo $\widehat{\mu}(x)$

2. Modelo ajustado $y(x) = \log(y - \widehat{\mu}(x))^2$ es :

$$\widehat{y}(x) = 6.35 + 0.182x_2 + 1.275x_2^2$$

3. Se optimiza el modelo anterior $X_1 : x_2$ y el óptimo es $x_2 = -0.07$

4. Se calculan los pesos $w_i = \exp(y(X_1))$

5. El nuevo modelo para el promedio es

$$\begin{aligned}\widehat{\mu}_w(x) = \quad 328.15 \quad & +177.1x_1 \quad +109.4x_2 \quad +131.2x_3 + 40.4x_1^2 - 22.4x_2^2 \\ & -38.2x_3^2 \quad +66.0x_1x_2 \quad +95.5x_1x_3 + 43.6x_2x_3\end{aligned}$$

6. Los valores óptimos $X_2 = (x_1, x_3)$ para $\widehat{\mu}_w(X_2)$. Posibles soluciones se obtienen graficando las curvas de nivel de la función $\widehat{\mu}_w$. Para fines de comparación se ha seleccionado uno de los mejores resultados, el cual se muestra en la tabla de resultados.

Presentación de resultados. En la tabla de abajo se reportan los resultados obtenidos por los diferentes autores, así podemos hacer una comparación de esos con la propuesta del modelo ponderado y las curvas de nivel.

Procedimientos	x_1	x_2	x_3	media	d.std.	var	ecm
VM (1990)	0.62	0.23	0.10	500.0	51.8	2679.7	2679.7
DM (1993)	0.98	0.03	-0.18	498.3	45.4	2037.6	2040.4
LT (1995)	1.00	0.07	-0.25	494.0	44.4	1974.0	2005.1
CN (1996)	0.98	0.03	-0.18	498.2	45.3	2035.7	2039.1
KL (1998)	1.00	0.06	-0.25	492.3	44.7	1951.8	2011.1
VB (1998)*	1.00	1.00	-0.07	500	7.0	49.10	49.1
MP (1999)#	1.00	-0.07	-0.12	499.1	23.2	568.79	2046
CNi	1.00	0.10	-0.25	499.0	45.0	2025.9	2026.4

* Modelo de regresión no paramétrico, solución particular para un acho de banda.

Este procedimiento reajusta el modelo inicial.

3 Conclusiones

No obstante los diferentes resultados, las curvas de nivel sobrepuestas contienen todas las soluciones análiticas. Las diferencias dan lugar a la necesidad de aplicar estos métodos a un problema real donde se puedan llevar a cabo pruebas confirmatorias. También, se

podría plantear un estudio de simulación para evaluar las ventajas reales de cada uno de los procedimientos.

Referencias

Aitkin, M. (1987) Modelling Variance Heterogeneity in Normal Regression Using GLIM. *Appl. Statist.* **36**, No. 3, 332-339.

Box, G. E. P. and Draper, N. R. (1987) *Empirical Model Building and Response Surfaces.* John Wiley & Sons, New York, NY.

Chan, L.K. and Mak, T.K.(1995) A Regression Approach for Discovering Small Variation around a Target. *Appl. Statist.* **44**, No. 3, 369-377.

Copeland, K. and Nelson, P. (1996) Dual Response Optimization via Direct Function Minimization. *Journal of Quality Technology* **28**, No. 3, 331-336.

Del Castillo, E. and Montgomery, D. (1993) A Nonlinear Programming Solution to the Dual Response Problem. *Journal of Quality Technology* **25**, No. 2, 199-204.

Domínguez, D. J. (2000) Procedimientos de Optimización en la Mejora Continua. *Comunicación Técnica CIMAT.*

Harvey, A.C. (1976) Estimating Regression Models with Multiplicative Heteroscedasticity. *Econometrica*, Vol.**44**, No, 3, 461-465.

Kim, K.J. and Lin, D.K.J. (1998) Dual Response Surface Optimization: A Fuzzy Modeling Approach. *Journal of Quality Technology* **30**, No. 1, 1-10.

Lin, D. and Tu, W. (1995) Dual Response Surface Optimization. *Journal of Quality Technology* **27**, 34-39.

Vining, G.G. and Myers, R. H. (1990) Combining Taguchi and Response Surface Philosophies: A Dual Response Approach. *Journal of Quality Technology* **22**, No. 1, 38-45.

Vining, G.G. and Bohn, R. H. (1998) Response Surfaces for the Mean and Variance Using a Nonparametric Approach. *Journal of Quality Technology* **30**, No. 3, 282-291.

Conteos Rápidos: Una Exploración Estadística

Guillermina Eslava Gómez
IIMAS, UNAM
Ignacio Méndez Ramírez
IIMAS, UNAM
Patricia Romero Mares
IIMAS, UNAM

1 Conteos rápidos

En el contexto de encuestas electorales, específicamente en conteos rápidos, se evalúan las varianzas para el estimador de razón, considerando diversos diseños muestrales y distintos tamaños de muestra.

Un conteo rápido es una muestra aleatoria de resultados de secciones (o rara vez casillas) electorales. Los resultados básicos con los que se cuenta es el número de votos a favor de cada uno de los partidos contendientes y el número de votos válidos. Se desea estimar la proporción nacional de votos a favor de cada uno de los partidos contendientes, generalmente sólo para los partidos mayoritarios. Es decir, el resultado poblacional (difiere ligeramente del resultado oficial por la impugnación y cancelación de algunas casillas) que desea estimarse es:

$$\frac{\sum_{i=1}^{N} \#\text{votos a favor del partido A en sec. i}}{\sum_{i=1}^{N} \#\text{votos válidos}}$$

Se debe estimar el numerador y también el denominador. Por esto se requiere de estimadores de razón, y no de proporciones donde el denominador es conocido.

$$\begin{aligned} Y &= \text{total de votos a favor del partido} \\ X &= \text{total de votos válidos} \\ R &= \frac{Y}{X} = \frac{\sum_{i=1}^{N} y_i}{\sum_{i=1}^{N} x_i} \\ y_i &= \text{número de votos a favor del partido A} \\ &\quad \text{en la sección } i \end{aligned}$$

x_i = número de votos válidos en la sección i

N = número de secciones electorales en la población.

Un estimador sesgado (con sesgo pequeño) pero consistente bajo muestreo aleatorio simple es la contraparte muestral:

$$\hat{R} = \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i}$$

n = número de secciones electorales en la muestra.

Se pueden usar diversos diseños muestrales:

a) simple aleatorio

b) con probabilidad proporcional al tamaño de la sección

c) estratificado

d) bietápico

e) combinaciones de los anteriores.

El costo de la encuesta, en lo que se refiere al levantamiento de la información, varía de acuerdo al tamaño y distribución de muestra. La intención de este trabajo es calcular las varianzas poblacionales del estimador de razón bajo diversos tamaños de muestra y bajo diversos diseños, esto como una guía para valorar los diseños, tamaños de muestra y relación a costo y precisión.
Se cuenta con los resultados de las elecciones presidenciales de 1994 desagregados por sección:
45,720,613 lista nominal
34,277,000 votos válidos
63,420 secciones electorales
300 distritos electorales (224 urbanos y 76 rurales)
32 Estados
5 circunscripciones

$R_{\text{PAN}} = 0.26$

$R_{\text{PRI}} = 0.5$

$R_{\text{PRD}} = 0.17$

$R_{\text{otros}} = 0.06$

a) Aleatorio simple:

$\hat{R}_{\text{mas}} = \sum_{i=1}^{n} y_i / \sum_{i=1}^{n} x_i$

$$\text{ECM}(\hat{R}_{\text{mas}}) = \left(1 - \frac{n}{N}\right) \frac{1}{n} \frac{N^2}{X^2} \left[\sum_{i=1}^{N} \frac{(y_i - Rx_i)^2}{N-1}\right].$$

b) Probabilidad proporcional al tamaño:

$$\hat{R}_{\text{ppt}} = \sum_{i=1}^{n} \frac{y_i}{p_i} / \sum_{i=1}^{n} \frac{x_i}{p_i}$$

$$\text{ECM}(\hat{R}_{\text{ppt}}) = \frac{1}{nX^2} \left[\sum_{i=1}^{N} \frac{1}{p_i} (y_i - Rx_i)^2\right].$$

c) Estratificado con varios tipos de estratificación:

$$\hat{R}_{\text{comb}} = \sum_{h=1}^{H} N_h \bar{y}_h / \sum_{h=1}^{H} N_h \bar{x}_h$$

$$\text{ECM}(\hat{R}_{\text{comb}}) = \frac{1}{\bar{X}^2} \sum_{h=1}^{H} \frac{N_h^2}{N^2} \left(1 - \frac{n_h}{N_h}\right) \frac{1}{n_h} \frac{1}{N_h - 1} \sum_{i=1}^{N_h} \left[(y_{h_i} - \bar{Y}_h) - R(x_{h_i} - \bar{X}_h)\right]^2$$

$$\hat{R}_{\text{sep}} = \sum_{h=1}^{H} \frac{N_h}{N} \hat{R}_h$$

$$\text{ECM}(\hat{R}_{\text{sep}}) = \frac{1}{\bar{X}^2} \sum_{h=1}^{H} \frac{N_h^2}{N^2} \left(1 - \frac{n_h}{N_h}\right) \frac{1}{n_h} \frac{1}{N_h - 1} \sum_{i=1}^{N_h} \left[y_{h_i} - R_h \bar{X}_{h_i}\right]^2.$$

d) Bietápico, upm=distritos, usm=secciones por "mas":

i)UPM por "mas"

$$\hat{R}_{\text{bmas}} = \sum_{i=1}^{n} M_i \bar{y}_i / \sum_{i=1}^{n} M_i \bar{x}_i$$

$$\mathrm{ECM}(\hat{R}_{\mathrm{bmas}}) \;=\; \left(1-\frac{n}{N}\right)\frac{N^2}{nX^2(N-1)}\sum_{i=1}^{N}(Y_i-rX_i)^2+\frac{N}{nX^2}\sum_{i=1}^{N}\frac{M_i^2}{m_i}\left(1-\frac{m_i}{M_i}\right)S_{w_i}^2$$

b) UPM por ppt

$$\hat{R}_{\mathrm{bppt}}=\sum_{i=1}^{n}\frac{M_i}{p_i}\bar{y}_i/\sum_{i=1}^{n}\frac{M_i}{p_i}\bar{x}_i$$

$$\mathrm{ECM}(\hat{R}_{\mathrm{bppt}}) \;=\; \frac{1}{X^2n}\sum_{i=1}^{N}\frac{1}{p_i}(Y_i-RX_i)^2+\frac{1}{X^2n}\sum_{i=1}^{N}\frac{1}{p_i}\frac{M_i^2}{m_i}\left(1-\frac{m_i}{M_i}\right)S_{w_i}^2$$

donde:

$$S_{w_i}^2=\frac{1}{M_i-1}\sum_{i=1}^{M_i}\left[y_{ij}-\bar{Y}_i-R(x_{ij}-\bar{X}_i)\right]^2.$$

En todos los casos:

$$\frac{|\mathrm{sesgo}(\hat{R})|}{\sqrt{\mathrm{Var}(\hat{R})}} \;\leq\; \mathrm{cv}(\hat{\bar{X}}).$$

2 Trabajo por hacer

Calcular varianzas para:

a) un muestreo con probabilidad proporcional al tamaño sin reemplazo.

b) estratificado bietápico.

c) otros diseños.

3 Comentarios y conclusiones

Resultó mejor, en términos de precisión, un muestreo estratificado por distrito (300 estratos). Se pueden considerar otros diseños cercanos que no dispersen tanto la muestra, como el de 5 circunscripciones y urbano/rural (10 estratos).

Si las elecciones no estan muy reñidas se puede optar por un diseño barato y tamaños de muestra entre 600 y 1000. Si la elección esta muy reñida se consideran diseños caros y tamaños de muestra de por lo menos 1000.

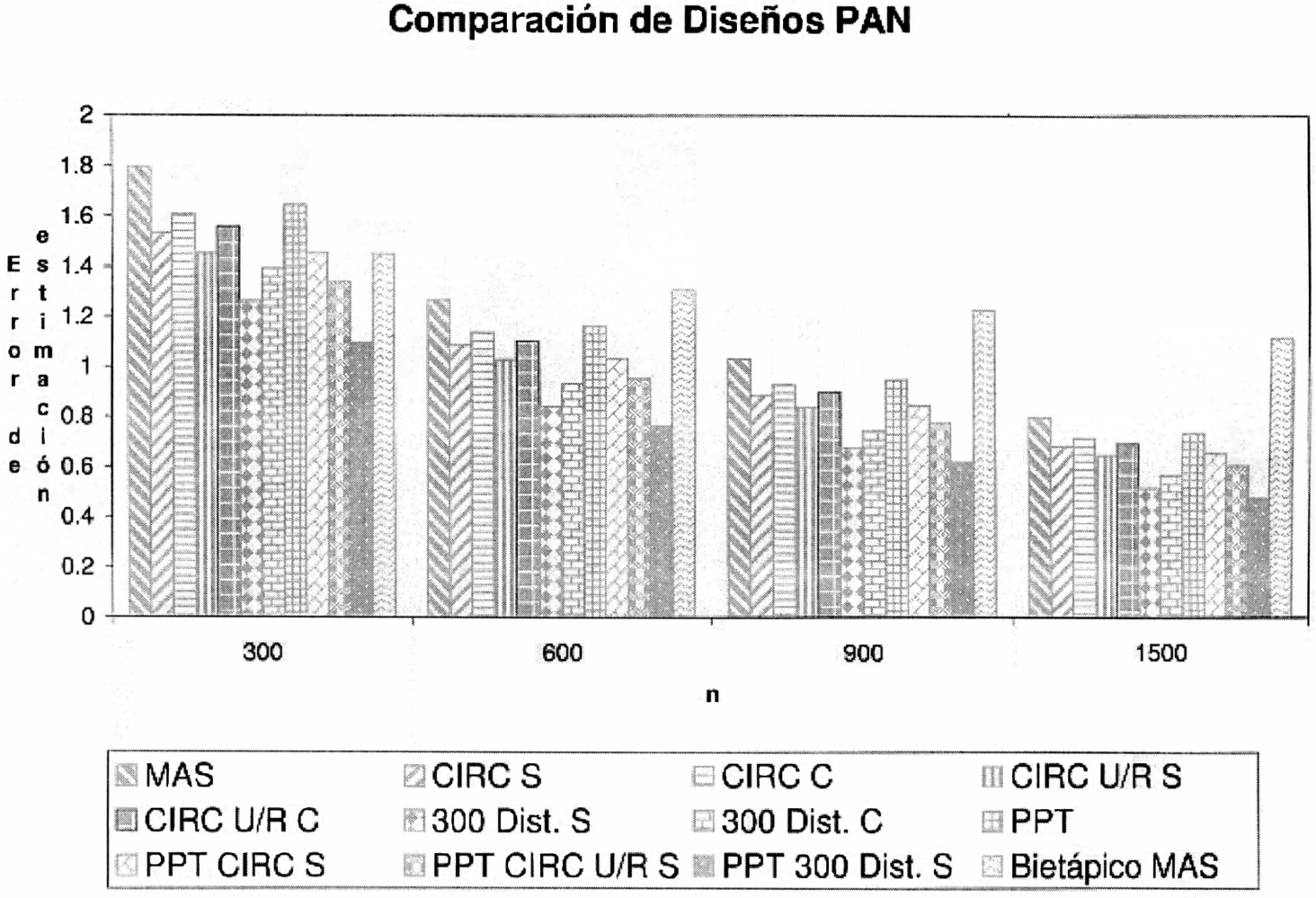

Comparación de Diseños PAN
Error de estimación
2
1.8
1.6
1.4
1.2
1
0.8
0.6
0.4
0.2
0
300
600
900
1500
n
MAS
CIRC S
CIRC C
CIRC U/R S
CIRC U/R C
300 Dist. S
300 Dist. C
PPT
PPT CIRC S
PPT CIRC U/R S
PPT 300 Dist. S
Bietápico MAS

Figura 1

Bibliografía

Cochran W.G. (1977) *Sampling Techniques.* J. Wiley. Trad. Ed Cecsa.

Hansen M.H., Hurwitz W.N., and Madow, W.G. (1953) *Sample Survey Methods and Theory* Vol II, J. Wiley.

Kish, L. (1965) *Survey Sampling.* N.Y.: Wiley. Trad. Ed. Trillas.

Raj R. (1968) *Sampling Theory.* Mc Graw Hill. Trad. Ed. Fondo de Cultura Económica.

Sukhatme, P.V., Sukhatme B.V. (1970) *Sampling Theory of Surveys with Applications.* Iowa State University Press. (tercera ed. 1984). Existe trad. al español.

Number of Connected Components of the Random Nearest Neighbors Graph

José M. González-Barrios
IIMAS, UNAM
Raúl Rueda
IIMAS, UNAM

1 Introduction

It has been of interest the study of Graph Theory under the assumption of vertices generated in a random environment, mainly due to the fact that they arise in several common situations. Many applications of graphs, such as the k-nearest neighbors graph, the minimal spanning tree and sphere of influence graph, have been proposed in the literature to solve problems in Statistics, such as detection of cluster structure, multivariate problems, and so on. In Computer Science they have been used in pattern recognition, computer vision, etc. (see references in Section 4). Without omitting, of course, their theoretical importance in Graph Theory. One of the important aspects about the study of these graphs, have to do with the connectivity, or the number of connected components of the graphs. In this work we find some results about the number of components of G_1, the graph of the nearest neighbors.

Given n points $X_1, X_2, \ldots, X_n$ in a metric space (S, d), the graph of the nearest points, denoted here by G_1 is constructed by joining each point X_i to its nearest neighbor, that is, we draw a directed edge between X_i and X_j if $d(X_i, X_j) = \min_{1 \leq k \leq n, k \neq i} d(X_i, X_k)$. It is clear that the resulting graph G_1 has n directed edges and n vertices. However the number of connected components of the resulting graph can not be specified in advance, all that is known is that if T_n is the number of connected components of G_1, generated by n points, then

$$1 \leq T_n \leq [n/2] \qquad \text{for all} \quad n \geq 2,$$

where $[m]$ denotes the greatest integer less or equal m. For basic definitions such as graph, edges, vertices, cycles, connected components, and so on. we refer the reader to Harary (1969) or any standard Graph Theory book.

Assume we generate n points $X_1, X_2, \ldots, X_n$ from the uniform distribution on $[0,1]$, and let us denote by $X_{n:1} \leq X_{n:2} \leq \cdots \leq X_{n:n}$ their order statistics. Now let us define

$$D_i := X_{n:i} - X_{n:i-1} \qquad \text{for} \qquad i = 2, 3, \ldots, n.$$

Then it is clear that the values of $D_i,\ i = 2, \ldots, n$, determine the graph G_1. If we define T_n as the number of connected components in the resulting graph G_1, we will see that we can find an expression of T_n in terms of the random variables $D_i,\ i = 2, \ldots, n$. In order to do so, we start with a more general theorem. Let us define

$$e_{i,j} = d(X_i, X_j) \qquad \text{for all} \qquad 1 \leq i, j \leq n,$$

where $\{X_i\}_{i=1}^n$ is a random uniform sample in the unit d-cube $[0,1]^d$. Then $e_{i,j} \neq e_{k,l}$ for all distinct pairs of different indices (i,j) and (k,l) among $1, 2, \ldots, n$, that is, all interpoint distances can be assumed to be different with probability one. We will denote by $\overline{(X_i, X_j)}$ a directed edge of a graph connecting X_i to X_j.

Theorem 1.1 *i) G_1 has no cycles containing more than two vertices.*
ii) The graph G_1 is connected if and only if there exists a unique pair (i,j) with $i, j \in \{1, 2, \ldots, n\}$ $i \neq j$ such that

$$\min\{e_{i,k} \,|\, k \in \{1, \ldots, n\} \backslash \{i\}\} = e_{i,j} = e_{j,i} = \min\{e_{j,k} \,|\, k \in \{1, \ldots, n\} \backslash \{j\}\}. \qquad (1)$$

iii) The number of connected components of G_1, equals the number of pairs (i,j) such that equation (1) holds.

Proof: Can be found in González-Barrios (1996).

Now let us go back to the case of dimension one. For a random sample in the real line, it is clear that given the order statistics $X_{n;1} < X_{n:2} < \cdots < X_{n:n}$, where we assume that matches do not occur, the nearest neighbor of $X_{n:i}$ can only be $X_{n:i+1}$ or $X_{n:i-1}$, for $2 \leq i \leq n-1$, and that the nearest neighbor of $X_{n:1}$ is $X_{n:2}$ and that of $X_{n:n}$ is $X_{n:n-1}$. We will assume that the random sample is generated from the uniform $(0,1)$ distribution, and we define the random variables $D_i = X_{n:i} - X_{n:i-1}$ for $2 \leq i \leq n$, known as "uniform spacings" in the literature, (see for example Pyke (1965)). In this classical paper it is proved, among many other things, that the set $\overline{D} := \{D_1, D_2, \ldots, D_{n+1}\}$ is formed by interchangeable random variables with the same distribution, where $D_1 := X_{n:1}$ and $D_{n+1} := 1 - X_{n:n}$, and since subsets of

interchangeable random variables are themselves interchangeable, then $\underline{D} := \{D_2, \ldots, D_n\}$ is a set of $n-1$ interchangeable random variables. Now if we define

$$Z_2 := 1_{\{D_2 < D_3\}},\ Z_i := 1_{\{D_i < \min\{D_{i-1}, D_{i+1}\}\}} \quad \text{for} \quad 3 \leq i \leq n-1, \quad Z_n = 1_{\{D_n < D_{n-1}\}} \text{ and,}$$

$$T_n = Z_2 + Z_3 + \cdots + Z_n.$$

A straightforward application of theorem 1.1, iii) implies that the random variable T_n counts the number of connected components of the graph G_1. In the next section we study some properties of T_n, as well as its asymptotic distribution.

2 Main results.

We start this section finding the first two moments of the random variable T_n.

The proofs of all the results stated in this paper can be found in González-Barrios and Rueda (1997).

Proposition 2.1 *For T_n defined as in the last paragraph of the previous section, the following holds*

i) For all $n \geq 3$, $E(T_n) = \frac{n}{3}$.

ii) For all $n \geq 5$, $Var(T_n) = \frac{2n}{45}$.

Recall that a sequence of random variables $\{X_n\}_{n=-\infty}^{\infty}$ is m-dependent if and only if for any integers a, b with $b - a > m$, $\{X_b, X_{b+1}, \ldots, X_{b+s}\}$ and $\{X_{a-r}, X_{a-r+1}, \ldots, X_a\}$ are independent sets of random variables, where r, s are nonnegative integers. From the proof of last Proposition we have:

Corollary 2.2 *The set of random variables $\{Z_2, Z_3, \ldots, Z_n\}$ is 2 dependent.*

Now we give a result about the connectivity of the nearest neighbors graph.

Proposition 2.3 *Let $\underline{X} = X_1, X_2, \ldots, X_n$ $n \geq 2$ be a sample from the uniform $(0,1)$ distribution, and let G_1 the graph of the nearest neighbors associated to $\underline{X}$. Then*

$$P(G_1 \textit{is a connected graph}) = \frac{2^{n-2}}{(n-1)!}.$$

To find the density of T_n is not an easy task, but for small n we can get the values of $P(T_n = k)$ for $1 \leq k \leq [n/2]$, using the result in the previous Proposition. For example for

$n = 2, 3,\ P(T_n = 1) = 1$, for $n = 4$ and $n = 5$ we have

$$P(T_4 = k) = \begin{cases} 2/3 & k = 1 \\ 1/3 & k = 2 \\ 0 & \text{otherwise} \end{cases} \qquad P(T_5 = k) = \begin{cases} 1/3 & k = 1 \\ 2/3 & k = 2 \\ 0 & \text{otherwise.} \end{cases}$$

For $n = 6$ and $n = 7$

$$P(T_6 = k) = \begin{cases} 2/15 & k = 1 \\ 11/15 & k = 2 \\ 2/15 & k = 3 \\ 0 & \text{otherwise} \end{cases} \qquad P(T_7 = k) = \begin{cases} 2/45 & k = 1 \\ 26/45 & k = 2 \\ 17/45 & k = 3 \\ 0 & \text{otherwise.} \end{cases}$$

For larger values of n it is possible to evaluate $P(T_n = k)$, but there seems to be no easy algorithm to do it. However, in the following section we obtain some asymptotic results for T_n.

3 Asymptotic results of T_n

Let $X_1, X_2, \ldots, X_n$ be random variables and define

$$S_n = \frac{\sum_{i=1}^n X_i}{n}.$$

There are well known results about conditions that relax the independent identically distributed assumptions, which allow strong laws of the large numbers. We now mention a theorem which can be found in Serfling (1980), page 27, with related extensions and proof in Serfling (1970).

Theorem A *Let $\{X_n\}_{n\geq 1}$ be random variables with means $\{\mu_n\}_{n\geq 1}$, variances $\{\sigma_n^2\}_{n\geq 1}$ and covariances $Cov(X_i, X_j)$ satisfying*

$$Cov(X_i, X_j) \leq \rho_{j-i}\sigma_i\sigma_j \qquad for \qquad 1 \leq i \leq j,$$

where $0 \leq \rho_k \leq 1$ for all $k \geq 0$. If the series $\sum_{i=1}^\infty \rho_i$ and $\sum_{i=1}^\infty \sigma_i^2(\log i)^2/i^2$ are both convergent, then

$$S_n - \frac{1}{n}\sum_{i=1}^n \mu_i \xrightarrow{wp1} 0. \tag{2}$$

We can apply this last Theorem to get

Proposition 3.1 *Let T_n be the number of connected components of G_1 the graph of the nearest neighbors of a uniform $(0,1)$ sample $X_1, X_2, \ldots, X_n$. Then*

$$\frac{T_n}{n} - \sum_{i=1}^{n} \mu_i = \frac{T_n}{n} - \frac{n}{3} \xrightarrow{wp1} 0.$$

Some other results, such as the law of iterated logarithm, can be obtained for the random variable T_n (see Serfling (1968)). Now we prove a central limit theorem, using a result proved in Hoeffding and Robbins (1948), as stated in Serfling (1968).

Theorem B *Let $\{X_i\}_{i=1}^{\infty}$ be a sequence of random variables, such that the sequence $\{X_i\}$ is m dependent and*

$$E(X_i) = 0 \qquad \textit{for all} \quad i = 1, 2, \ldots, \tag{3}$$

$$E(W_a^2) \sim A^2 \qquad \textit{uniformly in } a \qquad \textit{as} \quad n \to \infty, \tag{4}$$

$$E|X_i|^{2+\delta} \leq M \qquad \textit{for some} \quad \delta > 0 \quad \textit{and} \quad m < \infty, \tag{5}$$

where W_a denotes the normed sum $n^{-1/2} \sum_{i=a+1}^{a+n} X_i$. Then

$$P\left(\frac{\sum_{i=1}^{n} X_i}{(nA^2)^{1/2}} \leq z \right) \xrightarrow{n \to \infty} \frac{1}{(2\pi)^{1/2}} \int_{-\infty}^{z} \exp(-\frac{t^2}{2}) dt \quad \textit{for all } z \in I\!R\,. \tag{6}$$

From this Theorem we can obtain

Theorem 3.2 *Let be T_n the number of connected components of G_1 the graph of the nearest neighbors of a uniform $(0,1)$ sample $X_1, X_2, \ldots, X_n$. Rewrite $T_n = Z_2 + Z_3 + \cdots + Z_n$ as in the previous section. If we define $Y_i = Z_i - E(Z_i)$ for $i = 2, 3, \ldots, n$, then for $A^2 = 2/45$ equation (13) holds for the random variables Y_i.*

4 Applications and final remarks

Many authors have proposed to use graphs to solve some statistical problems. In particular the use of the connectivity of the k-nearest neighbors, besides edge lenghts of related graphs such as the minimal spanning tree, or even the nearest neighbor graph have been proposed to detect cluster structures in a random sample (see for example González-Barrios (1996), Steel and Tierney (1986), Tabakis (1992) and Zahn (1971)). These methods have proved to be quite efficient due to the nonparametric structure of theses graphs under mild assumptions.

In computer science it has been of importance the study of the influence graphs from a probabilistic perspective, which are closely related to the nearest neighbors graph (see Avis and Horton (1985), Dwyer (1995)), as stated in Chalker et al (1997) "...a type of proximity graphs for use in Pattern recognition, computer vision and other low-level vision tasks." More recently a non parametric test for equal distributions in higher dimension has been proposed in González-Barrios (1999), which is based on connectivity properties of the nearest neighbors graph. Also a multivariate two sample test has been suggested in Schilling (1986) using nearest neighbors.

In the case of higher dimensions, González-Barrios (1996) gives some theoretical results, which include the fact that for n points coming from the uniform distribution in the unit d-cube, the graph of the k nearest neighbors is connected with high probability for $k = O(\log(n))$. In that paper we also see that the value of k required for connectivity of G_k, the graph of the k-nearest neighbors, decreases when the dimension d increases. In fact based on simulations it is clear that for large values of d, the graph of the k-nearest neighbors is connected for k as small as 2, with large probability. Therefore if we are interested in the number of connected components of G_1 in dimension $d \geq 2$, for the uniform distribution in the unit d-cube, which we denote by T_n^d, we can always think of it as being bounded above by T_n as defined in this paper. To examplify that this result holds, we give in Table 1 the results of 500 simulations of T_n^d for $n = 50$ and for dimensions $d = 1, 2, 3, 4, 5, 10, 50, 100, 500, 1000, 5000, 10000$ which clearly show the decreasing nature of T_{50}^d as d increases to infinity.

Table 1: Table of frequencies of T_{50}^d for different dimensions and 500 simulations.

dim	1	2	3	4	5	10	50	100	500	1000	5000	10000
T_{50}^d												
1												
2												
3								2		2	2	3
4							1	2	6	7	6	6
5							4	5	7	17	21	16
6							10	22	35	42	41	44
7						3	27	57	64	71	81	79
8						8	72	85	107	103	102	124
9				1	5	30	100	114	99	96	97	98
10		1		7	14	45	119	79	93	85	77	74
11		2	14	26	31	98	79	78	65	48	51	32
12		11	25	35	85	118	53	37	13	20	17	19
13	33	41	73	94	120	100	21	15	9	7	5	3
14	69	79	96	128	112	56	10	4	2	2		2
15	122	110	111	115	70	31	4					
16	143	116	108	62	41	7						
17	75	83	45	25	18	4						
18	39	45	23	7	4							
19	17	10	5									
20	2	2										
mean	16.6	15.5	14.8	14.1	13.5	12.1	9.8	9.2	8.8	8.6	8.4	8.3

References.

Avis, D. and Horton, J. (1985) Remarks on the Sphere of Influence Graph, in *Discrete Geometry and Convexity*, J.E. Goodman et al. eds., New York Academy of Sciences, 323-327.

Chalker, T.K., Godbole, A.P., Radcliff, J. and Ruehr, O.G. (1997) On the Size of a Random Sphere of Influence. preprint.

Dwyer, R.A. (1995) The Expected Size of the Sphere-of-Influence Graph. *Comput. Geom. Th. and Appl.* **5**, 155-164.

González-Barrios, J.M. (1996a) The Connectivity of the Random k-Nearest Neighbors Graph. *Aportaciones Matematicas* **12**, 107-118.

González-Barrios, J.M. (1996b) Clustering and the k-Nearest Neighbors Graph. preprint.

González-Barrios, J.M. (1999) A New Test for Detecting Equal Distributions in Higher Dimensions. In preparation.

González-Barrios, J.M. and Rueda, R. (1997) Number of Connected Components of the Random Nearest Neighbor Graph. *Preimpreso del IIMAS, UNAM* **No. 64**, México.

Harary, F. (1969) *Graph Theory.* Addison-Wesley, Reading, Mass.

Hoeffding, W. and Robbins, H. (1948) The Central Limit Theorem for Dependent Random Variables. *Duke Math. J.* **15**, 773-780.

Pyke, R. (1965) Spacings (with discussion). *J. R. Statist. Soc.* B **27**, 395-449.

Schilling, M.F. (1986) Multivariate Two Sample Test Based on Nearest Neighbors. *J. Amer. Statist. Assoc.* **81**, 799-806.

Serfling, R.J. (1968) Contributions to Central Limit Theory for Dependent Variables. *Ann. Math. Statist.* **39**, 4, 1158-1175.

Serfling, R.J. (1970) Convergence Properties of S_n under Moment Restrictions. *Ann. Math. Statist.* **41**, 4, 1235-1248.

Serfling, R.J. (1980) *Approximation Theorems of Mathematical Statistics.* Wiley Ser. in Prob. and Math. Statist., John Wiley and Sons, New York.

Steel, J.M. and Tierney, L. (1986) Boundary Domination and the Distribution of the Largest Nearest Neighbor Link in Higher Dimension. *J. of Appl. Prob.* **23**, 524-528.

Tabakis, E. (1992) Asymptotic and Computational Problems in Single-Link Clustering. PhD Thesis, M.I.T.

Zahn, C.T. (1971) Graph-Theoretical Methods for Detecting and Describing Gestalt Clusters. *IEEE Trans. on Comp.* **C-20**, No. 1, 68-86.

Modelación a Nivel Nacional del Impacto de la Campaña "5 al Día" en el Estado de Veracruz

Rafael Alberto Guajardo Panes
Subcoordinación Censal Estatal, INEGI
Mario Miguel Ojeda Ramírez
Laboratorio de Investigación y Asesoría Estadística, Universidad Veracruzana
Iñigo Verdalet Guzmán
Instituto de Ciencias Básicas, Universidad Veracruzana

1 Resumen

Se presenta el análisis de los datos agregados a nivel municipal, que resultaron de una encuesta repetida diseñada para elaborar un diagnóstico alimentario y evaluar el impacto de una campaña publicitaria diseñada con el fin de motivar el incremento del consumo de frutas y verduras. Se construyeron indicadores socioeconómicos y de perfiles de consumo de alimentos y se realizaron comparaciones por zonas geográficas. Con el propósito de explicar el impacto de la campaña y el cambio de dieta en las familias, se ajustó un modelo de regresión bivariada, demostrando que el conocimiento de la campaña está asociado al nivel socioeconómico y que se conoció más la campaña en municipios con alto índice de consumo de frutas y verduras, además de que ésta tuvo menor impacto en municipios con alto consumo de verduras y bajo consumo de frutas.

2 Introducción

La relación entre indicadores socioeconómicos y hábitos alimentarios se ha estudiado ampliamente (FAO, 1993). Se han reportado evidencias para México (Avila et al. (1995)) y para Veracruz (Verdalet (1994) y Verdalet et al. (1996)). En estos estudios se ha encontrado que el consumo de frutas y verduras es un condicionante importante de la salud y el estado nutricional de la familia, pero también se ha destacado que éste se presenta asociado a aspectos culturales. No sólo es la disponibilidad de los alimentos lo que garantiza su consumo.

El gobierno del Estado de Veracruz instrumentó la campaña "5 al día", que consistió en un esquema publicitario para motivar el consumo de frutas y verduras usando todos los medios de comunicación masiva. Esta acción fue precedida de una evaluación diagnóstica sobre el

consumo de estos alimentos en una muestra de familias veracruzanas. Después de seis meses de campaña se aplicó una encuesta sobre la misma muestra de familias y se evaluó el impacto de la campaña.

El objetivo central de este trabajo es evaluar como influyen los hábitos de alimentación de frutas y verduras y el nivel socioeconómico de grupos de población sobre el impacto de la campaña, tanto en el reconocimiento de los mensajes publicitarios como en el cambio de hábitos en el consumo de frutas y verduras.

3 Metodología

El esquema de muestreo usado fue el de conglomerados estratificado en cuatro etapas (Ver Tabla 1). El procedimiento consistió en elegir áreas geoestadísticas básicas (AGEB) de cada municipio seleccionando; en cada AGEB seleccionada se enumeraron conglomerados (manzanas o grupos de al menos cinco viviendas) y se seleccionó una muestra de éstos; y, finalmente, en los conglomerados seleccionados se enumeraron y seleccionaron viviendas. En todas las etapas posteriores a la selección de municipios se utilizó muestreo aleatorio simple. Se aplicó un cuestionario diseñado de acuerdo a los objetivos del estudio. Más detalles sobre el método de muestreo se pueden ver en Verdalet et al. (1998). Las entrevistas se realizaron en dos ocasiones: (1) antes de la campaña y (2) después de la campaña.

Zona	Población		Muestra	
	Número	Porcentaje	Municipios	Viviendas
Norte	48	23	17	1,053
Centro	110	53	43	2,075
Sur	52	24	10	1,241
Total	**210**	**100**	**70**	**4,369**

Tabla 1. Distribución de municipios por zonas y tamaños de muestra respectivos.

Se utilizaron datos agregados a nivel municipal que son promedios de frecuencias de consumo y porcentajes calculados sobre el total de viviendas encuestadas por municipio. Se tiene un total de 31 variables de consumo de frutas y verduras, 2 variables de impacto y 3 variables socioeconómicas. Con los datos de frecuencias se realizó un análisis de componentes principales y se seleccionaron los siete primeros componentes que acumularon el 77% de la variación total. Se agregó un índice socioeconómico, que fue el primer componente principal de las variables: porcentaje de viviendas con agua potable, con luz eléctrica y con drenaje, el cual explicó el 78% de la variación total. Las variables de impacto se evaluaron integralmente para cada familia, realizando un análisis de varias preguntas en el cuestionario

que se referían a detalles de la campaña y a la dieta de la familia. De tal manera que se evaluó si la familia conocía efectivamente la campaña, y si había evidencia de un cambio en su dieta respecto al consumo de frutas y verduras. La descripción general de las variables que se utilizaron para el análisis definitivo en este trabajo aparece en la Tabla 2. Cada una de estas variables es un índice simple y describe, en una escala de menor a mayor, el perfil de consumo de frutas y verduras que se mencionan. De manera análoga se interpreta el índice socioeconómico.

Tipo	Indicador	Descripción
Consumo	FACCON1	Índice que representa en la parte alta de la escala a los municipios cuyas familias tienen en promedio un alto consumo de todas las frutas y verduras. Excepto nopales y quelites.
	FACCON2	Índice que representa en la parte más alta de la escala a los municipios cuyas familias tienen un promedio alto en el consumo de limón, papa, chayote, zanahoria, hojas verdes, pepino, brócoli y col; y un consumo bajo promedio de mandarina y pera.
	FACCON3	Índice que representa en la parte alta de la escala a los municipios cuyas familias registran un alto consumo promedio de ejote, col y quelites; y un bajo consumo promedio de naranja, manzana, jitomate, cebolla y chile.
	FACCON4	Índice que representa en la parte alta de la escala a los municipios cuyas familias registran un alto consumo promedio de durazno, piña, sandía y elote, y un bajo consumo promedio de tuna, chile, hojas verdes, nopales y brócoli.
	FACCON5	Índice que representa, en la parte alta de la escala a los municipios cuyas familias registran un alto consumo promedio de mandarina, mango, chile, papa y nopal; y un bajo consumo promedio de plátano, manzana y tuna.
	FACCON6	Índice que representa en la parte alta de la escala a los municipios cuyas familias registran un alto consumo promedio de naranja, mandarina, uva, zanahoria y brócoli; y un bajoo consumo promedio de manzana, pera, ciruela, chile, elote y ejote.
	FACCON7	Índice que representa en la parte alta de la escala a los municipios cuyas familias registran un alto consumo promedio de guayaba, chile, col y quelites; y un bajo consumo promedio de pera, melón, calabaza y nopales.
Socioeconómico	FACSOECO	Índice socioeconómico relacionado directamente con el porcentaje de hogares que cuentan con servicios de agua, luz y drenaje.
Impacto	SI_MODIFI	Proporción de familias que modificaron su dieta.
	SI_CONOCE	Proporción de familias que afirmaron conocer el Programa "5 al día".

Tabla 2. Descripción de las variables utilizadas para el análisis de regresión.

Se realizó una serie de análisis exploratorios e inferenciales comparativos por zona a partir de análisis de la varianza. Para tener una perspectiva inicial de la relación entre las variables, se obtuvo la matriz de correlaciones, tanto a nivel global como para cada zona.

Se aplicó un análisis de regresión multivariada en el que las variables de impacto (SI-

CONOCE, SI-MODIFI) fueron las variables dependientes y las variables independientes fueron los indicadores de consumo de frutas y verduras junto con el indicador socioeconómico. La forma del modelo en notación matricial es: $\mathbf{Y} = \mathbf{X}\beta + \mathbf{E}$, donde $\mathbf{Y}$ es la matriz de datos (68×2) de la variable respuesta, $\mathbf{X}$ es la matriz de datos (68×9) de las variables explicatorias, β es la matriz (9×2) de coeficientes y $\mathbf{E}$ es la correspondiente matriz de errores aleatorios.

Una vez obtenido el modelo ajustado por mínimos cuadrados ordinarios, se realizó un análisis diagnóstico sobre los residuos y se obtuvo un modelo reducido. Para establecer diferencias por zona, el modelo reducido fue corrido para cada zona y se realizaron comparaciones formales de los parámetros utilizando variables indicadoras en el modelo global múltiple para la variable respuesta SI-CONOCE.

4 Resultados

Fueron detectadas diferencias significativas ($p \leq 0.01$) por zona en los hábitos de consumos de frutas y verduras. Los municipios de la zona sur observaron mayores consumos de frutas y verduras en general, aunque en la zona norte el índice en consumo promedio de mandarina, mango, chile, papa y nopal fue mayor.

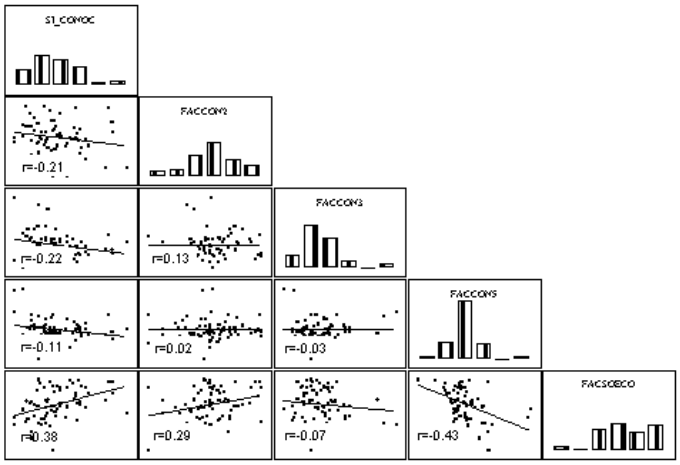


Figura 1. Gráfica de escalera con coeficientes de correlación de las principales variables en el estudio.

Se detectó, así mismo, una diferencia significativa en el nivel socioeconómico de los municipios de la zona centro.

Fue identificada una escasa relación entre las variables de impacto, exceptuando en la zona sur, donde en los municipios con mayor número de familias que conocieron el programa,

mayor número de ellas modificó la dieta. Sin embargo, el conocimiento de la campaña fue escaso (entre 23% y 31% en promedio por zona) y la modificación de la dieta siempre mayor y uniforme entre zonas (entre 38% y 39%).

En la Figura 1 se observa que las relaciones más fuertes entre las variables representadas es SI-CONOCE vs FACSOECO y FACCON5 vs FACSOECO.

Se obtuvo que el modelo multivariado fue significativo ($p = 0.027$); sin embargo, el nivel de correlación entre las variables respuesta no lo fue ($p = 0.3139$). La Tabla 3 presenta los resultados de las pruebas de hipótesis sobre los coeficientes de regresión por separado, tanto de manera univariada como multivariada.

Prueba para el efecto de:	Variable Dependiente	Coeficientes estimados	Valor p	
			Univariado	Multivariado
FACCON1	*SI_MODIFI* *SI_CONOCE*	$\hat{\beta}_{11} = -1.431$ $\hat{\beta}_{12} = 1.018$	0.639 0.609	0.755
FACCON2	*SI_MODIFI* *SI_CONOCE*	$\hat{\beta}_{21} = -3.319$ $\hat{\beta}_{22} = -5.644$	0.259 0.004	0.015
FACCON3	*SI_MODIFI* *SI_CONOCE*	$\hat{\beta}_{31} = -1.431$ $\hat{\beta}_{32} = -3.471$	0.652 0.098	0.252
FACCON4	*SI_MODIFI* *SI_CONOCE*	$\hat{\beta}_{41} = 1.671$ $\hat{\beta}_{42} = -0.671$	0.569 0.726	0.773
FACCON5	*SI_MODIFI* *SI_CONOCE*	$\hat{\beta}_{51} = -0.535$ $\hat{\beta}_{52} = 1.779$	0.885 0.464	0.741
FACCON6	*SI_MODIFI* *SI_CONOCE*	$\hat{\beta}_{61} = -1.582$ $\hat{\beta}_{62} = -1.496$	0.564 0.405	0.641
FACCON7	*SI_MODIFI* *SI_CONOCE*	$\hat{\beta}_{71} = -5.878$ $\hat{\beta}_{72} = 0.430$	0.037 0.812	0.100
FACSOECO	*SI_MODIFI* *SI_CONOCE*	$\hat{\beta}_{81} = 1.556$ $\hat{\beta}_{82} = 8.019$	0.665 0.001	0.005

Tabla 3. Resultados de las pruebas de hipótesis para cada uno de los coeficientes en los modelos univariado y multivariado (Se excluye el término constante).

Al realizar el análisis de regresión múltiple por zona, aplicando el método de selección de variables paso a paso hacia delante, para la variable si conoce se obtuvieron los modelos reducidos que aparecen en la Tabla 4.

ZONA	VARIABLES					R^2
	CONSTANTE	FACCON2	FACCON3	FACCON5	FACSOECO	
NORTE	20.617	-10.122	-19.383	-10.951	8.652	0.752
	(5.731)	(4.138)	(7.554)	(2.898)	(3.028)	
CENTRO	34.574	-6.536	-7.426	11.302	11.879	0.374
	(3.201)	(3.319)	(2.829)	(4.578)	3.941	
SUR	31.401	-7.549	-7.304	6.568	----	0.479
	(4.905)	(3.497)	(2.794)	(4.265)	----	

Tabla 4. Modelos de regresión múltiple reducidos por zona para la variable SI_CONOCE.

Se obtuvo que el modelo con mayor ajuste fue el de la zona norte, el cual es a su vez diferente a los de las zonas centro y sur en el coeficiente asociado a la variable FACCON5, que es el perfil de consumo de frutas y verduras que observó una distinción clara para esta zona.

5 Conclusiones

La zona centro cuenta con un mayor índice de desarrollo socioeconómico. Por otro lado, se observó que la zona sur mostró un claro patrón de asociación entre el conocimiento del programa "5 al día" y el haber modificado la dieta, lo cual no se observó para las zonas norte y centro. Por otro lado, el nivel socioeconómico no fue un factor determinante en el conocimiento del programa para la zona sur, y sí para la zona centro y zona norte. Cabe destacar que el indicador FACCON5 influyó de manera diferente en la zona norte, donde se registra un mayor consumo de mandarina, mango, chile, papa y nopal; y un bajo consumo promedio de plátano, manzana y tuna, y ésto influyó de manera diferente en la proporción de familias que dijeron haber conocido el programa.

El impacto de la campaña si se explica por el nivel socioeconómico, pero no en los municipios donde hay mayor consumo de frutas y verduras en general. Por otro lado, altos consumos de verduras combinadas con alto consumo de ejote, col y quelites y con alto consumo de tomate, cebolla y chile, con bajo consumo de frutas en general, produjeron los más bajos niveles de conocimiento del programa.

Referencias

Avila, C. A., Chávez, V. A., Shamah, L. T. y Madrigal, F. H. (1995) La Desnutrición Infantil en el Medio Rural Mexicano: Análisis de las Encuestas Nacionales de Alimentación. *Salud Pública* **35**, 658-666.

FAO, OMS y OPS (1993) Magnitud y Tendencias de los Problemas Nutricionales. *Conferencia Internacional sobre Nutrición; Situación Alimentaria y Nutricional de América Latina.* Chile.

Verdalet, I. (1994) Diagnóstico Alimentario y Nutricional de la Población Residente en el Municipio de Xalapa, Veracruz. *Gaceta* Febrero, Universidad Veracruzana, México.

Verdalet, I., Ojeda, M. M. y Méndez, V. M. (1996) *Diagnóstico Nutricio y Alimentario de las Familias del Estado de Veracruz.* Reporte Global, Universidad Veracruzana, México.

Verdalet, I., Ojeda, M. M., Silva, E. R. y López, L. (1998) *Impacto de La campaña "5 al día" en el Estado de Veracruz.* Reporte Global, Universidad Veracruzana, México.

Estimación Conjunta de una Serie de Tiempo Ajustada y de sus Efectos Deterministas Lineales

Víctor M. Guerrero
Instituto Tecnológico Autónomo de México

1 Introducción

Los efectos deterministas se incluyen en los modelos de series de tiempo para incorporar la información provista por la variación del calendario, por observaciones aberrantes y/o por intervenciones. La detección, modelación y estimación de efectos del calendario han sido estudiadas por, entre otros, Bell y Hillmer (1983). Los mismos temas, correspondientes a efectos de observaciones aberrantes, fueron examinados por Tsay (1988), Chen y Liu (1993) y Gerlack et al. (1999), mientras que los referidos al análisis de intervención se encuentran en Box y Tiao (1975) y Guerrero (1991). En este artículo no se tratan más esos temas.

El principal problema que aquí se considera es la estimación de la serie ajustada por efectos deterministas. En la literatura sobre el tema, no se ha puesto atención a este problema ya que, aparentemente, los analistas piensan que una vez que los efectos han sido estimados, los valores ajustados de la serie quedan estimados adecuadamente al sustraerle a cada valor observado sus efectos deterministas correspondientes. La siguiente sección muestra que un enfoque más apropiado debe considerar la estimación conjunta de los parámetros de los efectos y la serie ajustada, lo cual impone una restricción a los estimadores. Primero se encuentra una solución óptima pero no-factible y después se sugiere un procedimiento factible, el cual se justifica asintóticamente. En un documento de trabajo no publicado (véase Guerrero, 1999) se muestran algunos ejemplos, tanto teóricos como empíricos, que indican un beneficio adicional del método propuesto, esto es, que los parámetros de los efectos deterministas resultan estimados con un incremento sustancial en eficiencia, respecto de los estimadores irrestrictos.

2 Conceptos básicos

Sea $\{Z_t\}$ una serie de tiempo finita que puede representarse como

$$Z_t = \mathbf{X}_t'\beta + e_t \quad \text{para } t = 1, ..., N \tag{1}$$

donde $\mathbf{X}_t$ es un vector columna de observaciones de k variables independientes fijas, observadas en el tiempo t, y β es un vector de parámetros. $\{e_t\}$ es una serie de errores aleatorios que sigue un modelo Auto-Regresivo Integrado de Promedios Móviles (ARIMA)

$$\phi(B)d(B)e_t = \theta(B)a_t \tag{2}$$

donde B denota al operador de retraso tal que $Be_t = e_{t-1}$, $d(B)$ es un operador de diferencias de orden d que vuelve estacionaria a $\{d(B)e_t\}$, y $\{a_t\}$ es un proceso de ruido blanco, con media cero y varianza σ_a^2. Además $\phi(B) = 1 - \phi_1 B - ... - \phi_p B^p$ y $\theta(B) = 1 + \theta_1 B + ... + \theta_q B^q$ son polinomios primos con raíces fuera del círculo unitario. Cuando $d(B) \neq 1$, $\{Z_t\}$ es no-estacionaria y, para que (2) esté bien definida, se supondrá que el proceso inició en un punto finito del pasado con condiciones iniciales fijas.

El vector $\mathbf{X}_t$ puede incluir variables del tipo $X_{it} = (1 - \delta B)^{-1} P(T)_t$, donde $P(T)_t$ es una función de pulso que toma el valor 1 cuando $t = T$ y es cero en otro caso. Entonces, como lo hizo Tsay (1988), X_{it} se asocia con un Cambio Transitorio si $\delta \in (0, 1)$, con un Efecto Aditivo si $\delta = 0$ o con un Cambio de Nivel si $\delta = 1$. Similarmente se puede especificar un Efecto Innovativo al definir$X_{it} = [\phi(B)d(B)]^{-1}\theta(B)P(T)_t$. Efectos más complejos de intervenciones también pueden incorporarse al definir a X_{it} apropiadamente, como lo indican Box y Tiao (1975). De igual forma, diversos tipos de efectos de calendario, como los que consideran Bell y Hillmer (1983), se pueden representar de manera lineal y, por ende, están cubiertos por el modelo (1).

El valor ajustado (no-observable) de la serie se define como

$$Z_t^{(A)} = Z_t - \mathbf{X}_t'\beta \quad \text{para} \quad t = 1, ..., N. \tag{3}$$

Considérese ahora el predictor óptimo de un paso hacia adelante de Z_t, en el sentido de Error Cuadrático Medio (ECM) mínimo, que está dado por $E(Z_t|\mathbf{Z}_{t-1})$, con $\mathbf{Z}_{t-1} = (Z_1, ..., Z_{t-1})'$ para $t = 2, ..., N$. Como $Z_t - E(Z_t|\mathbf{Z}_{t-1}) = a_t$, se sigue que $Z_t^{(A)} - E(Z_t^{(A)}|\mathbf{Z}_{t-1}) = a_t$ y $E(Z_t^{(A)}|\mathbf{Z}_{t-1}) = Z_t - \mathbf{X}_t'\beta - a_t$. Se define entonces el vector de predicciones un paso hacia adelante como $E(\mathbf{Z}_N^{(A)}|Z_1, \mathbf{Z}_2, ..., \mathbf{Z}_{N-1}) = (E(Z_1^{(A)}), E(Z_2^{(A)}|Z_1), ..., E(Z_N^{(A)}|\mathbf{Z}_{N-1}))'$ con $\mathbf{Z}_N^{(A)} = \mathbf{Z}_N - X\beta$, de tal forma que

$$E(\mathbf{Z}_N^{(A)}|Z_1, \mathbf{Z}_2, ..., \mathbf{Z}_{N-1}) = \mathbf{Z}_N - X\beta - \mathbf{a}_N \tag{4}$$

donde $\mathbf{Z}_N^{(A)}$ y $\mathbf{a}_N$ se definen de igual forma que $\mathbf{Z}_N$, y $X = (\mathbf{X}_1, \mathbf{X}_2, ..., \mathbf{X}_N)$. Así, es claro ahora que $\mathbf{Z}_N^{(A)}$ y β están ligados y, por lo tanto, deben estimarse simultáneamente, teniendo en cuenta la restricción impuesta por (3). Para ello supóngase que $\mathbf{b}$ es un estimador irrestricto de β tal que

$$\mathbf{b} = \beta + \mathbf{u} \quad \text{con} \quad E(\mathbf{u}) = \mathbf{0}_k \quad \text{y} \quad Var(\mathbf{u}) = \Sigma_u, \tag{5}$$

donde $\mathbf{0}_k$ denota el vector cero de dimensión k (el primer paso del método sugerido más adelante indica cómo obtener $\mathbf{b}$ y Σ_u). De esta forma se obtiene el siguiente sistema de ecuaciones

$$\begin{pmatrix} E(\mathbf{Z}_N^{(A)}|Z_1, \mathbf{Z}_2, ..., \mathbf{Z}_{N-1}) \\ \mathbf{b} \end{pmatrix} = \begin{pmatrix} \mathbf{Z}_N^{(A)} \\ \beta \end{pmatrix} + \begin{pmatrix} -\mathbf{a}_N \\ \mathbf{u} \end{pmatrix} \tag{6}$$

restringido por (3), esto es

$$\mathbf{Z}_N = \begin{pmatrix} I_N & , & X \end{pmatrix} \begin{pmatrix} \mathbf{Z}_N^{(A)} \\ \beta \end{pmatrix}, \tag{7}$$

con I_N la matriz identidad de dimensón N, $E(\mathbf{a}_N|Z_1, \mathbf{Z}_2, ..., \mathbf{Z}_{N-1}) = \mathbf{0}_N$, $Var(\mathbf{a}_N|Z_1, \mathbf{Z}_2, ..., \mathbf{Z}_{N-1}) = \sigma_a^2 I_N$ y $E(\mathbf{a}_N\mathbf{u}'|Z_1, \mathbf{Z}_2, ..., \mathbf{Z}_{N-1}) = 0$.

3 Estimación restringida óptima

Una aplicación de Mínimos Cuadrados Restringidos (véase Judge et al., 1980, p. 56) produce el siguiente resultado.

Teorema 3.1. *Supóngase que el sistema de ecuaciones (6)-(7) es válido con* $E(\mathbf{Z}_N^{(A)}|Z_1, \mathbf{Z}_2, ..., \mathbf{Z}_{N-1})$, $\mathbf{b}$,σ_a^2 *y* Σ_u *conocidos. Sea* X *una matriz de rango completo, entonces el Estimador Restringido Optimo (Lineal con ECM Mínimo) de* $\mathbf{Z}_N^{(A)}$ *y el Estimador Lineal Insesgado con Varianza Mínima de* β *son*

$$\hat{\mathbf{Z}}_N^{(A)} = E(\mathbf{Z}_N^{(A)}|Z_1, \mathbf{Z}_2, ..., \mathbf{Z}_{N-1}) + A_Z[\mathbf{Z}_N - E(\mathbf{Z}_N^{(A)}|Z_1, \mathbf{Z}_2, ..., \mathbf{Z}_{N-1}) - X\mathbf{b}] \tag{8}$$

y

$$\hat{\beta} = \mathbf{b} + A_\beta[\mathbf{Z}_N - E(\mathbf{Z}_N^{(A)}|Z_1, \mathbf{Z}_2, ..., \mathbf{Z}_{N-1}) - X\mathbf{b}]. \tag{9}$$

$$A_Z = \sigma_a^2(\sigma_a^2 I_N + X\Sigma_u X')^{-1}, \ y \ \ A_\beta = \Sigma_u X'(\sigma_a^2 I_N + X\Sigma_u X')^{-1}. \tag{10}$$

Además, la matriz de ECM $\Sigma_Z = E[(\hat{\mathbf{Z}}_N^{(A)} - \mathbf{Z}_N^{(A)})(\hat{\mathbf{Z}}_N^{(A)} - \mathbf{Z}_N^{(A)})']$, *la matriz de varianza-covarianza* $\Sigma_\beta = Var(\hat{\beta})$ *y la matriz de covarianza* $\Sigma_{Z\beta} = E[(\hat{\mathbf{Z}}_N^{(A)} - \mathbf{Z}_N^{(A)})(\hat{\beta} - \beta)']$ *están dadas por*

$$\Sigma_Z = (I_N - A_Z)\sigma_a^2, \ \Sigma_\beta = (I_k - A_\beta X)\Sigma_{\{u\}} \ y \ \Sigma_{Z\beta} = -A_Z X\Sigma_u. \tag{11}$$

El Teorema 3.1 presupone el conocimiento de diversas cantidades que se desconocen en la práctica. No obstante, un supuesto razonable en aplicaciones prácticas es que tales cantidades pueden estimarse a partir de datos observados. Por ello, $\mathbf{b}$, $\hat{\Sigma}_u$ y $\hat{\sigma}_a^2$ se supondrán conocidos de aquí en adelante. De hecho sus estimaciones pueden obtenerse una vez que el modelo (1)-(2) se ha estimado como lo indican Bell y Hillmer (1983), Tsay (1984) o Chen y Liu (1993). En tal caso, también se contará con los residuos estimados

$$\hat{a}_t = \hat{\theta}(B)^{-1}\hat{\phi}(B)d(B)(Z_t - \mathbf{X}_t'\mathbf{b}) \text{ para } t = p + d + 1, ..., N, \tag{12}$$

donde el número real de observaciones usadas al estimar el modelo ARIMA es $n = N - (p + d)$. Entonces pueden estimarse los estimadores restringidos como se indica a continuación. Desafortunadamente, las propiedades de estos nuevos estimadores no son idénticas a las de (8)-(9). Sin embargo, el resultado que sigue establece su equivalencia asintótica.

Teorema 3.2. *Supóngase que el sistema (6)-(7) se mantiene válido con* N *reemplazada por* n, $\mathbf{Z}_n = (Z_{p+d+1}, ..., Z_N)'$ *en lugar de* $\mathbf{Z}_N$ *y* $\mathbf{a}_n$ *(definido similarmente) en lugar de* $\mathbf{a}_N$. *Sea* X *una matriz de rango completo definida en concordancia y sea* $(-\mathbf{a}_n', \mathbf{u}')'$ *distribuído como una Normal. Si, conforme* $n \to \infty$, $\mathbf{b} \xrightarrow{p} \beta$, $\hat{\Sigma}_u \xrightarrow{p} \Sigma_u$, $\hat{\sigma}_a^2 \xrightarrow{p} \sigma_a^2$ *y* $\hat{\mathbf{a}}_n - \mathbf{a}_n \xrightarrow{p} \mathbf{0}_n$, *entonces los Estimadores Restringidos Estimados*

$$\hat{\hat{\beta}} = \mathbf{b} + \hat{\mathbf{A}}_\beta \hat{\mathbf{a}}_n \quad y \quad \hat{\hat{\mathbf{Z}}}_n^{(A)} = \mathbf{Z}_n - X\hat{\hat{\beta}}, \tag{13}$$

donde $\hat{A}_\beta = \hat{\Sigma}_u X'(\hat{\sigma}_a^2 I_n + X\hat{\Sigma}_u X')^{-1}$, *son equivalentes asintóticamente a los Estimadores Restringidos Optimos (8)-(9), con* $\sqrt{n}(\hat{\hat{\beta}} - \beta) \xrightarrow{d} N(\mathbf{0}_k, \Sigma_\beta)$ *y* $\sqrt{n}(\hat{\hat{\mathbf{Z}}}_n^{(A)} - \mathbf{Z}_n^{(A)}) \xrightarrow{d} N(\mathbf{0}_n, \Sigma_Z)$.

Adicionalmente Σ_β, Σ_Z y $\Sigma_{Z\beta}$ se pueden estimar consistentemente mediante

$$\hat{\Sigma}_\beta = (I_k - \hat{A}_\beta X)\hat{\Sigma}_u \quad , \hat{\Sigma}_Z = X\hat{\Sigma}_\beta X' \quad y \quad \hat{\Sigma}_{Z\beta} = -X\hat{\Sigma}_\beta. \tag{14}$$

4 Conclusiones

El método que se propone utilizar en la práctica para estimar una serie ajustada, es bietápico y complementa a los métodos en uso actualmente. Las propiedades asintóticas de los estimadores resultantes son idénticas a las de los Estimadores Restringidos Optimos, en condiciones que se alcanzan razonablemente en la práctica. Aunque el objetivo original pueda ser sólo la estimación de la serie ajustada, el que se incremente la eficiencia en la estimación de los efectos deterministas puede justificar el uso del método con este fin exclusivamente.

Referencias

Bell, W.R. y Hillmer, S.C. (1983) Modeling Time Series with Calendar Variation. *Journal of the American Statistical Association*, **78**, 526-534.

Box, G.E.P. y Tiao, G.C. (1975) Intervention Analysis with Applications to Economic and Environmental Problems. *Journal of the American Statistical Association*, **70**, 70-79.

Chen, C. y Liu, L-M. (1993) Joint Estimation of Model Parameters and Outlier Effects in Time Series. *Journal of the American Statistical Association*, **88**, 284-297.

Gerlack, R., Carter, C. y Kohn, R. (1999) Diagnostics for Time Series Analysis. *Journal of Time Series Analysis*, **20**, 309-330.

Guerrero, V.M. (1991) *Análisis Estadístico de Series de Tiempo Económicas*. México: UAM-Iztapalapa.

Guerrero, V.M. (1999) Joint Estimation of an Adjusted Time Series and its Linear Deterministic Effects. Documento de Trabajo DE-C99.1. Departamento de Estadística, ITAM.

Judge, G.G., Griffiths, W.E., Hill, R.C. y Lee, T.C. (1980) *The Theory and Practice of Econometrics*, New York: Wiley.

Tsay, R.S. (1984) Regression Models with Time Series Errors. *Journal of the American Statistical Association*, **79**, 118-124.

Tsay, R.S. (1988) Outliers, Level Shifts and Variance Changes in Time Series. *Journal of Forecasting*, **7**, 1-20.

Proceso Estocástico de la Pesca del Atún Usando el Modelo de Líneas de Espera

José de Jesús Lara Tejeda

Facultad de Ciencias, Universidad Autónoma de Baja California.

Hector Manzo Monroy

Facultad de Ciencias Marinas, Universidad Autónoma de Baja California.

1 Resumen

La teoría de Líneas de Espera presenta una posibilidad de modelar actividades humanas para obtener expresiones analíticas que involucren variables de naturaleza aleatoria. La pesca del atún de cerco en el Océano Pacífico tiene componentes de naturaleza aleatoria como son: navegar, deriva, búsqueda, lance, salida y llegada a puerto, etc (Hilborn y Walters (1992)). Un tratamiento como proceso estocástico de esa actividad pesquera constituye, por lo tanto, un caso para El Modelo de Líneas de Espera.

El tratamiento estocástico correspondiente permite desarrollar las expresiones analíticas para estimar probabilísticamente la cantidad promedio de buques en la búsqueda y lance, así mismo, estimar la cantidad promedio de buques en el proceso de encontrar señales que permitan llegar a la fase de servicio del lance. A estas cantidades puede estimarseles el tiempo promedio de espera y de estancia total en las actividades pesqueras.

El informe diario de la pesca es la fuente principal de información para trabajar esta modelación.

2 Planteamiento del problema de la pesca.

El problema de pesquería se plantea de la siguiente manera: Una cantidad m de buques se encuentra en dos posibles situaciones mútuamente excluyentes:

1.- El buque está navegando, se encuentra a la deriva o se localiza en un puerto.

2.- El buque realiza la fase de pesca.

En la fase de pesca se involucran dos actividades secuenciadas: búsqueda de cardúmen y actividad del lance. Cuando el buque termina un lance, regresa a navegar o se sitúa a la deriva; de estas dos anteriores puede suceder que el buque proceda a localizarse posteriormente en un puerto.

El modelo de líneas de espera esta constituido por tres etapas: la población, la espera y el servicio. En la etapa de población las actividades de la pesca que la conforman son: la salida y llegada de puerto, navegar y deriva; en la etapa de espera se encuentran las actividades: buscar, caza, mamifero, pájaros, brisa, otraseñas, chapoteo y palo; en la etapa de servicio las actividades que la forman son: lance con mamífero, lance de atún, lance con palo y termina lance.

En este modelo el buque se encuentra, al empezar, en la etapa de la población, es decir, que esta en el puerto, hace salida, asi como entrada a puerto, navega y esta a la deriva. En un momento dado el buque decide pescar y pasa a la etapa de espera en la cual está buscando señales que le permitan hacer el lance correpondiente y por lo tanto realiza una o varias de las actividades mencionadas.

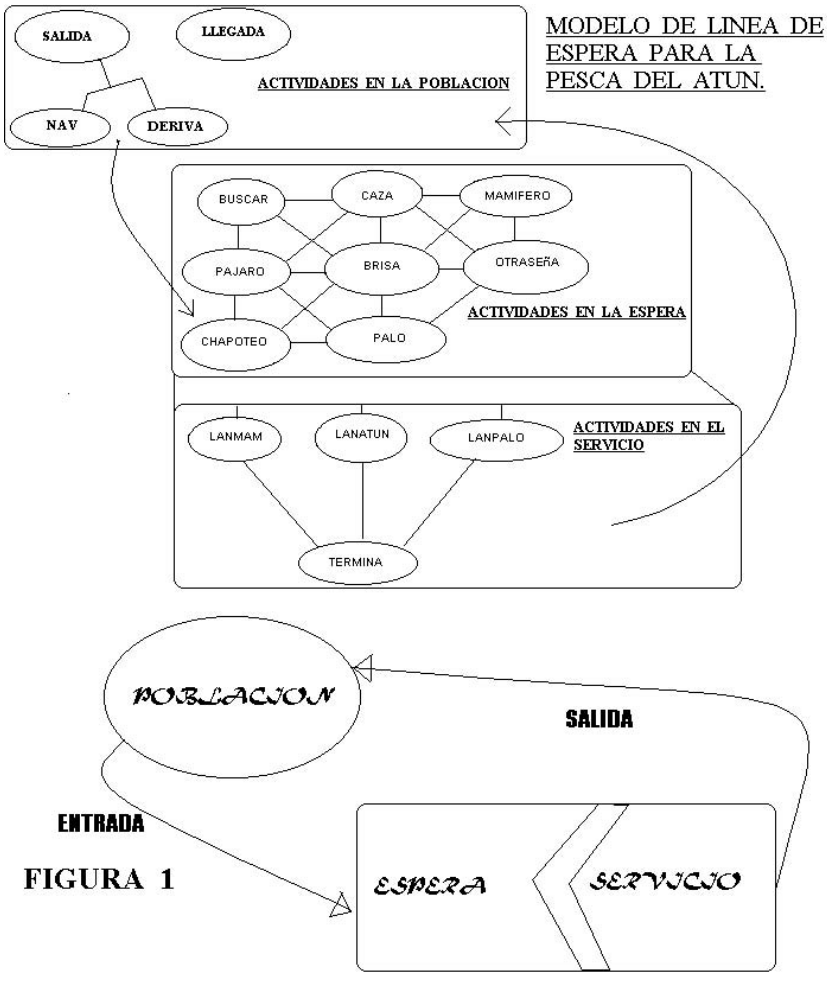


FIGURA 1

Cuando ya tiene localizado un cardumen, pasa a la etapa de servicio llevando a cabo un lance; esta etapa en particular para la pesca es un autoservicio. La figura 1 presenta una

gráfica de la relación del Modelo de Líneas de Espera con el agrupamiento de las actividades de la pesca en las etapas de este modelo.

3 Distribución de probabilidad de la presencia de buques en el proceso de pesca

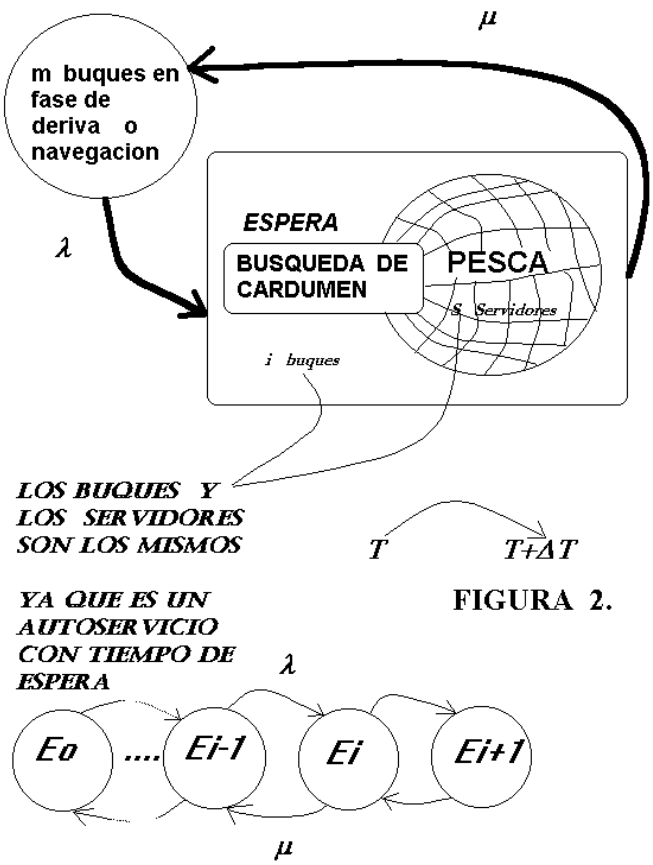


FIGURA 2.

Sea i la cantidad de buques que se encuentran en la fase de pesca, con esto se define el estado E_i de una cadena de Markov que en la figura 2 se representa gráficamente para este problema, de manera que dicho proceso puede estar en el estado E_{i-1}, en E_i o en E_{i+1}. Considerando las probabilidades de llegar al estado $p(E_i)$ en particular, la probabilidad de que un buque entre al proceso de pesca es $\lambda_i \cdot \Delta t$ y la probabilidad de que un buque salga es $\mu_i \cdot \Delta t$, el desarrollo conduce a la expresión general de la probabilidad de tener i buques buscando señales de cardumen o pescando (Prawda (1980)):

$$P_i = P_0 \cdot \frac{(\lambda_0 \cdot \lambda_1 \cdot \ldots \cdot \lambda_{i-1})}{(\mu_1 \cdot \mu_2 \cdot \ldots \cdot \mu_i)},$$

donde λ es la tasa con la que los buques pasan al proceso de pesca, μ es la tasa de servicio con la que los buques realizan los lances.

En forma particular el ritmo como los buques pasan de la población al proceso de pesca permite suponer que $\lambda_i = (m - i) \cdot \lambda$, donde m es la cantidad de buques que forman a la flota pesquera. Para el caso de la pesquería, los buques que esperan el servicio de pesca al pasar a pescar, ellos mismos se dan el servicio (autoservicio), por lo que no se da el caso de que $i > S$ y se considera que $\mu_i = i\mu$. En un autoservicio, en general, no se tiene tiempo de espera (Taha (1995)), pero en este caso existe una fase en que el buque busca señales de cardumen por lo que este problema es un autoservicio con tiempo de espera. Lo anterior permite llegar a:

$$P_i = \frac{m\lambda(m-1)\lambda(m-2)\lambda \ldots (m-i+1)\lambda}{1\mu 2\mu \ldots i\mu} \cdot P_0$$

lo que fácilmente se traduce en:

$$P_i = C_i^m \left(\frac{\lambda}{\mu}\right)^i P_0$$

4 Determinación de parámetros aleatorios

Para determinar cuantos buques se encuentran en la fase *busqueda-lance* se utiliza la esperanza matemática: $W = \sum_{i=0}^{m}(i \cdot P_i)$.

El tiempo promedio en que un buque permanece en *busqueda-lance* (Espera-Servicio) estaría dado por $Tw = \frac{W}{\lambda}$

El tiempo promedio en que un buque permanece en la búsqueda de cardumen sería dado por $T_L = Tw - \frac{1}{\mu}$.

La cantidad de buques en la fase de búsqueda sería: $L = \lambda \cdot T_L$.

5 Aplicación

De la información del diario de la pesca de 3 meses de una flota de 20 buques, estadísticamente se calculó el tiempo promedio entre llegadas en 2.37 horas con una desviación estandard de 0.42; para el teimpo entre lances se obtuvo 2.41 horas con desviación estandard de 0.39. Con estos datos se determinó la tasa promedio para entrar a la fase *busqueda-lance* $\lambda = \frac{1}{2.37} = 0.4219$; la tasa de autoservicio $\mu = \frac{1}{2.41} = 0.4149$; probabilidad de que todos los buques estén

navegando, a la deriva o en puerto,

$$P_0 = \frac{1}{\sum_{i=0}^{m} \left(C_i^m \left(\frac{\lambda}{\mu} \right)^i \right)} = 8.061 \times 10^{-7}.$$

Dada la distribución de probabilidad : $P_i = C_i^{20}(1.017)^i \cdot 8.061 \times 10^{-7}$, la cantidad de buques en la fase *busqueda-espera* es 10.084 que considerando $1-\alpha = 0.924$ el intervalo de confianza es 6 a 14 buques. La estimación de buques en navega o a la deriva es $m \cdot P_0 + (m-1) \cdot P_1 + ... + (m-i) \cdot P_i = 9.916$.

El tiempo promedio de estancia en la fase *busqueda-lance* es $Tw = 23.898$ horas; comparándolo con datos estadísticos de la información fuente que da 26 horas con desviación estandard de 8 horas se considera una buena estimación del Modelo de Líneas de Espera.

El tiempo en la fase de búsqueda es $T_L = Tw - \frac{1}{\mu} = 21.488$; la cantidad de buques en la fase de búsqueda es $\lambda \cdot T_L = 9.067$; y los buques promedio en lance es $\lambda \cdot (2.41) = 1.017$.

Bibligrafía

Hilborn, R. and Walters, C.J. (1992) *Quantitative Fisheries Stock Assessment. Choice, Dynamics & Uncertainty.* Chapman and Hall.

Prawda, J. (1980) *Método y Modelos de Investigación de Operaciones. Vol. I* Limusa México.

Taha, H.A. (1995) *Investigación de Operaciones.* Alfaomega, México

Introducción al Bootstrap y sus Aplicaciones

Gabriel Nuñez Antonio
Instituto Tecnólogico Autónomo de México

1 Introducción.

Un aspecto importante en el proceso de inferencia estadística es el de conocer el comportamiento probabilístico de alguna función de los datos con el fin de evaluar el desempeño de algun estimador. Dependiendo de los supuestos distribucionales que se tengan y de la forma funcional de la estadística de interés, será posible conocer la distribución de muestreo de dicha estadística. Las técnicas de remuestreo dan la posibilidad de aproximar la distribución de una estadística independientemente si se conoce o no la distribución de los datos involucrados. En este trabajo se presentan los conceptos fundamentales de una de estas técnicas de remuestreo, conocida como *bootstrap*.

2 Descripción general del problema.

Sean $Y_1, ..., Y_n$ variables aleatorias independientes e idénticamente distribuidas con función de densidad de probabilidad, f.d.p.,$f(\mathbf{y}|\theta)$ y función de distribución acumulada, f.d.a., $F(\mathbf{y}|\theta)$.
Objetivo: Usar la información muestral para realizar inferencias sobre algún parámetro θ, usando alguna estadística T.

3 Soluciones bootstrap

La idea principal de los métodos bootstrap es el remuestreo a partir de los datos originales, ya sea en forma directa o vía un modelo ajustado; con la finalidad de obtener muestras replicadas a partir de las cuales se pueda evaluar la variabililad de las cantidades de interés sin incurrir en errores de cálculos analíticos. Los métodos bootstrap también se puden aplicar en problemas simples para verificar las características de las medidas de incertidumbre, para relajar supuestos, o para dar rápidas soluciones aproximadas. Un ejemplo de lo anterior es el remuestreo aleatorio para estimar la distribución permutacional de alguna estadística de prueba no paramétrica.

Es verdad que en muchas aplicaciones se puede confiar ampliamente en un modelo paramétrico particular y en el correspondiente análisis clásico basado en dicho modelo. Aún así, puede ser de utilidad investigar que tanto se puede inferir sin asumir los supuestos de un modelo

paramétrico particular. Esto es la esencia de la robustez del análisis estadístico realizado. El bootstrap no paramétrico permite hacer esto.

De acuerdo a lo anterior los métodos bootstrap se pueden aplicar tanto en los casos en los que se cuente con un modelo probabilístico bien definido para los datos, como en los casos en los que no se disponga de dicho modelo.

3.1 Simulación paramétrica

Suponga que se tiene un modelo paramétrico particular $F(\mathbf{y}|\theta)$ para la distribución de los datos $Y_1, ..., Y_n$. Cuando el parámetro es estimado por $\hat{\theta}$, a menudo pero no necesariamente su estimador de máxima verosimilitud, y este es substituido en el modelo se obtiene el *modelo ajustado*, con f.d.a. $\hat{F}(\mathbf{y}) = F(\mathbf{y}|\hat{\theta})$, el cual se puede usar para averiguar propiedades de T, algunas veces en forma exacta. Se empleará Y^* para denotar la v.a. distribuida de acuerdo al modelo ajustado $\hat{F}$.

3.2 Simulación no paramétrica

Si no se tiene un modelo paramétrico específico, pero se puede asumir que $Y_1, ..., Y_n$ son variables aleatorias independientes e idénticamente distribuidas de acuerdo a una función de distribución desconocida F. Se usará la función de distribución empírica $\hat{F}$ para estimar la f.d.a. desconocida F. $\hat{F}$ se usará de la misma forma que en un modelo paramétrico, se obtendrán cálculos teóricos de ser posible, de otro modo se realizarán simulaciones de conjuntos de datos y cálculos empíricos para descubrir las propiedades de interés. Las simulaciones a partir de la función de distribución empírica (f.d.e.) son directas, lo anterior debido a que la f.d.e. asigna probabilidades iguales a cada una de las observaciones originales $Y_1, ..., Y_n$. Así, cada y^* se muestrea aleatoria e independientemente de estos datos. Por lo tanto, la muestra simulada será una m.a. tomada con remplazo de los datos. Este procedimiento de muestreo es llamado *Bootstrap no paramétrico.*

4 Inferencias usando bootstrap

Si se quieren obtener intervalos de $(1-\alpha)$ de confianza para θ, es posible (en algunos casos) mostrar que T es aproximadamente Normal con media $\theta+\beta$ y varianza ν, donde β es el sesgo de T. Si β y ν fueran conocidos, entonces $P[T \leq t|F] \approx \Phi(t-(\theta+\beta)/\nu^{1/2})$. Y un intervalo de confianza de $(1-\alpha)$ aproximado para θ sería $(t-\beta-\nu^{1/2}Z_{1-\alpha/2},\ t-\beta-\nu^{1/2}Z_{\alpha/2})$.

Sin embargo, en la práctica el sesgo y la varianza no son conocidos. Por lo que para usar

la aproximación Normal se deben remplazar por sus respectivos estimadores. Hay que notar que

$$\beta = b(F) = E[T|F] - t(F), \qquad \nu = \nu(F) = V[T|F].$$

Si F es estimada por $\hat{F}$, la cual podría ser la f.d.e. o una distribución paramétrica ajustada. Entonces, los correspondientes estimadores del sesgo y la varianza se pueden obtener como

$$\beta = b(\hat{F}) = E[T|\hat{F}] - t(\hat{F}), \qquad \nu = \nu(\hat{F}) = V[T|\hat{F}].$$

4.1 Intervalos de confianza

La mayor aplicación para la distribución y los cuantiles de un estimador T está en el cálculo de intervalos de confianza. Existen diversas maneras de usar bootstrap en este contexto, aquí sólo se presentan dos métodos básicos.

La aproximación más simple es usar una aproximación Normal a la distribución de T. Esto significa estimar los límites del intervalo, para lo cual sólo se requieren estimadores bootstrap del sesgo y la varianza. Sin embargo, la aproximación Normal no siempre resulta adecuada. Si se usa bootstrap para estimar los cuantiles de $T - \theta$, entonces un intervalo de confianza para θ tendrá los siguientes límites

$$\left(t - (t^*_{[(R+1)(1-\alpha/2)]} - t), t - (t^*_{[(R+1)(\alpha/2)]} - t)\right) \tag{1}$$

donde$(t^*_{[(R+1)(1-\alpha/2)]} - t)$ y $(t^*_{[(R+1)(\alpha/2)]} - t)$ son los cuantiles bootstrap de $T - \theta$. Los límites (1) son referidos como límites de confianza bootstrap básicos. Su precisión depende de R, por supuesto, y se podría tomar $R \geq 100$ para tener mayor seguridad. Sin embargo, la precisión de estos límites también depende del grado de concordancia de la distribución de $T^* - t$ y la de $T - \theta$.

Si la distribución de $T - \theta$ no depende de parámetros desconocidos, se puede definir

$$Z = (T - \theta)/V^{1/2}$$

donde V es un estimador de la $V[T|F]$. Aquí resulta más adecuado estimar los cuantiles de Z por medio de replicas de la estadística bootstrap estandarizada $Z = (T^* - \theta)/V^{1/2*}$ donde T^* y V^* están basados sobre una m.a. simulada, $Y^*_1, ..., Y^*_k$. Si el modelo es paramétrico, las Y^*_j, son generadas de la distribución paramétrica ajustada, y si el modelo es no paramétrico

las Y_j^* son generadas a partir de la f.d.e. En cualquier caso se emplea la $(R+1)\alpha$- ésima estadística de orden de los valores simulados $Z_1^*, ..., Z_k^*$. Si $Z^*_{[(R+1)\alpha]}$ estima a Z_α . Entonces, el intervalo bootstrap estudentizado para θ tiene los siguientes límites

$$(t - \nu^{1/2} Z^*_{[(R+1)(1-\alpha/2)]}, t - \nu^{1/2} Z^*_{[(R+1)\alpha/2]}) \tag{2}$$

5 Ejemplo

La tabla 1 presenta los datos, obtenidos de Hand et al. (1994), de un índice de monóxido de carbono registrado en siete pacientes con varicela, los datos muestran la medición del índice en el momento de admisión a un hospital (X) y después de una semana de estancia en el hospital (Y). El objetivo es estimar el cambio promedio en el índice.

Paciente	X	Y
1	40	73
2	50	52
3	56	80
4	58	85
5	60	64
6	62	63
7	66	60

Tabla 1. *Indice de monóxido de carbono en 7 pacientes*

Una prueba estadística clásica para verificar si las diferencias $(X-Y)$ son normales, realizada en S-plus, presenta los siguientes resultados:

Estadística	g.l.	valor p	I.C. al 0.95
2.0892	6	0.0817	(-2.079, 26.363)

¿Son las diferencias Normales?, ¿Son apropiados los intervalos de confianza, basados en una pivotal T-Student?.

Comparación de la varianza del estimador bootstrap t^* con la varianza estimada de t:

$$\hat{V}(t) = 33.782331, \qquad V(t^*) = 27.23971$$

Intervalo de Confianza Bootstrap estudentizado, para la diferencia de medias:

$$(0.785108,\ 26.794124)$$

Las conclusiones de este ejemplo se dejan al interesado.

6 Conclusiones

En este trabajo se ejemplificó que el bootstrap se puede aplicar en situaciones donde no se puede suponer que la distribución de muestreo de la estadística de interés es normal.

El bootstrap también puede ser útil cuando la distribución de muestreo no tiene una solución analítica, por ejemplo, la diferencia de dos medianas muestrales.

En tales situaciones, más que una aproximación clásica de intervalos de confianza, uno puede considerar métodos bootstrap para intervalos de probabilidad, algunos de los cuales se presentaron en este trabajo. Mooney y Duval (1992) presentan una discusión de estos.

Si se confía en los análisis de inferencia clásicos, se puede emplear bootstrap no paramétrico para evaluar la robustez del análisis realizado, en cuanto a la violación de los supuestos requeridos por algún modelo.

Aunque existe en la literatura bastante material sobre métodos bootstrap, en este trabajo se presentan las ideas básicas de esta metodología de remuestreo como un paso inicial para abordar problemas en áreas donde este tipo de técnicas pueden ser la única opción viable para realizar inferencias estadísticas. Por mencionar algunas, en economía y demografía existen índices y medidas de desigualdad las cuales son funciones no lineales de cierto grupo de v.a. En tales casos los estimadores por intervalo proporcionados por la teoría asintótica pueden resultar inadecuados y las propiedades de estos intervalos para muestras pequeñas no son conocidas.

Existen otras áreas que involucran problemas donde el parámetro de interés esta acotado, por lo que la aplicación de resultados asintóticos estándar puede producir estimadores por intervalos que se sobrepasen de las respectivas cotas.

El bootstrap se presenta como un método alternativo para obtener intervalos de probabilidad, además los intervalos bootstrap no son demasiado caros en términos computacionales y son fáciles de calcular. Si bien los métodos bootstrap tienen sus desventajas, dada las ventajas potenciales estos métodos aparecen como una herramienta que vale la pena considerar para hacer inferencias estadísticas.

Referencias.

Cochran, W. G. (1977) *Sampling Techniques.* 3a. edición. John Wiley.

Davison, A. C. y Hinkley, D. V. (1997) *Bootstrap Methods and their Application.* Cambridge University Press.

Efron, B. (1979) Bootstrap Methods: Another Look at the Jacknife. *Ann. of Statist.* **7**, 1-26.

Efron, B. and Tibshirani, R. (1993) *An Introduction to the Bootstrap.* Chapman and Hall.

Hand, D. J., Daly F., Lunn, A. D., McConway, K. J. y Ostrowski E. (1994). *A Handbook of Small Data Seta.* Chapman and Hall.

Hinkley, D.V. (1988) Bootstrap Methods (with Discussion). *J. of the Royal Statist. Soc., B*, **50**, 312-337, 355-370.

Mooney Z. C. y Duval D. R. (1992) Bootstrap inference: A preliminar Monte Carlo evaluation. Presented at the annual meeting of the American Political Science Association, Chicago.

Las Componentes de Varianza y su Desarrollo Computacional Aplicado a Ensayos de Híbridos de Maíz

Emilio Padrón Corral
Centro de Investigación en Matemáticas Aplicadas U.A. de C.
Emilio Olivares Sáenz
Facultad de Agronomía U.A.N.L.

1 Resumen

A partir de los cuadrados medios de efectos principales, interacciones y sus particiones, usando los datos porcentuales de mala cobertura en 164 híbridos de maíz, evaluados en tres localidades Celaya, Gto., Río Bravo, Tamps., y Gómez Palacio, Dgo. México, respectivamente, se estimaron los componentes de varianza obteniendo no sólo la magnitud relativa de las varianzas, sino también los porcentajes de la suma de las varianzas estimadas de cada uno de los efectos del modelo, así como su programa computacional.

2 Introducción

En experimentación agrícola es frecuente la necesidad de comparar tratamientos en diferentes localidades (Padrón (1996)), Por lo tanto en este trabajo nuestro objetivo es estimar los componentes de varianza de un modelo desbalanceado y desarrollarlo computacionalmente, para ello se cuenta con programas computacionales específicos para este tipo de análisis estadístico (Olivares (1998)). Aquí se estudia el comportamiento de genotipos de maíz en diferentes localidades, y se logra el verdadero efecto de sus varianzas contando para ello con la herramienta fundamental de las esperanzas de cuadrados medios, en base a la técnica del análisis de varianza (ANOVA) como se aprecia en (Searle (1987)), también se nos comenta en (Searle et al. (1992)), como las sumas de cuadrados del análisis de varianza para datos desbalanceados siguen siendo las mismas que para datos balanceados excepto que en lugar de tener n se debe tener n_i, y en lugar de N se debe tener $\sum n_i$, y el que los datos sean desbalanceados no elimina la posibilidad de obtención de estimadas negativas de σ_τ^2 (componentes de varianza para tratamientos) en el análisis de varianza. La teoría desarrollada en este trabajo se aplicó a un experimento de campo titulado "Selección y Evaluación Agronómica de Líneas S_2 y S_3 de Maíz (Zea mays L.) en tres Ambientes Contrastantes " (Martínez (1993)). Esta

investigación forma parte del programa de mejoramiento genético del Instituto Mexicano del Maíz (Mario E. Castro Gil) de la U.A.A.A.N. y consta de tres localidades, Celaya, Guanajuato. Río Bravo, Tamaulipas. y Gómez Palacio, Durango. Además de 164 híbridos donde 160 son experimentales y 4 comerciales, utilizados estos últimos como testigos. También está a la disposición un programa computacional el cual se desarrolló en lenguaje Visual Basic para correr en el ambiente Windows y cuando se ejecuta aparece la pantalla inicial donde aparecen los titulos, posteriormente se pasa a la pantalla principal en donde se presenta una ventana para escribir la información básica del número de ambientes, genotipos, repeticiones y testigos, posteriormente se oprime la tecla continuar dentro de la ventana y se prepara una hoja de trabajo para los datos, los cuáles son capturados considerando el orden: localidades, genotipos, repeticiones; los datos del primer grupo (híbridos) deben de capturarse primero (los híbridos son los primeros h tratamientos y posteriormente se capturan los t testigos). Los datos pueden ser guardados en un archivo. Después de capturar los datos se oprime la tecla continuar y se presenta el análisis de varianza y las medias de ambientes, genotipos, genotipos por ambientes, híbridos, testigos. También se presenta una tabla con las estimaciones de las componentes de varianza. Los resultados del programa fueron comparados con el análisis estadístico realizado con calculadora de escritorio, obteniendo los mismos resultados.

3 Descripción del problema

Dado un modelo estadístico con varios factores agronómicos e interacción, el ANOVA se descompondrá como sigue: Genotipos se partirá en cruzas , testigos, y se hace el contraste de cruzas contra testigos y las respectivas interacciones con localidad. Para cada uno de ellos se desarrollan las sumas de cuadrados y se obtienen las esperanzas de cuadrados medios para conocer sus correspondientes componentes de varianza estimadas y por lo tanto, saber en cuánto están contribuyendo de acuerdo a los resultados del experimento de campo, además de comprobarlo computacionamente.

4 Metodología

Se aplicará la técnica de la esperanza de cuadrados medios en el modelo que a continuación se presenta:

$$Y_{ijk} = \mu + L_i + R_k i + G_j + (LG)_{ij} + E_{ijk}$$

donde

$$\begin{aligned} i = &\quad 1, 2, 3, \ldots, t \quad \text{localidades} \\ j = &\quad 1, 2, 3, \ldots, r \quad \text{genotipos} \\ k = &\quad 1, 2, 3, \ldots, l \quad \text{repeticiones} \end{aligned}$$

Y_{ijk} : Variable aleatoria observable de la i-ésima localidad en el j-ésimo genotipo de la k-ésima repetición

μ : Media general

L_i : Efecto de la i-ésima localidad

$R_k i$: Efecto de la k-ésima repetición dentro de la i-ésima localidad

G_j : Efecto del j-ésimo genotipo

$(L * G)_{ij}$: Efecto conjunto de la i-ésima localidad y del j-ésimo genotipo

E_{ijk} : Componente aleatoria asociada con la i-ésima localidad en el j-ésimo genotipo de la k-ésima repetición

De acuerdo a dicho modelo, el cuadrado medio de la interacción Localidad*Genotipo es el apropiado cuadrado medio del error para probar genotipos. Además se supone que las esperanzas de efectos son cero, es decir,

$$E[L_i] = E[R_k i] = E[G_j] = E[(LG)_{ij}] = E[E_{ijk}] = 0$$

También se supone que las esperanzas de productos cruzados de sus diferentes efectos son cero, y se cumple que

$$E[L_i^2] = \sigma_L^2 \qquad E[R_k i]^2 = \sigma_{R/L}^2 \qquad E[G_j^2] = \sigma_G^2 \qquad E[(LG)_{ij}^2] = \sigma_{LG}^2 \qquad E[E_{ijk}^2] = \sigma_e^2$$

5 Resultados

A continuación se obtienen las esperanzas de cuadrados medios obtenidos de cada una de las fuentes de variación correspondientes al modelo, así como de sus particiones, en este caso sólo se presenta el resultado final.

$$\begin{aligned} E[CM(Loc)] &= rg\sigma_L^2 + r\sigma_{LG}^2 + g\sigma_{R/L}^2 + \sigma_e^2 \\ E[CM(Rep/Loc)] &= g\sigma_{R/L}^2 + \sigma_e^2 \\ E[CM(Gen)] &= rl\sigma_G^2 + r\sigma_{LG}^2 + \sigma_e^2 \\ E[CM(Cruza)] &= rl\sigma_C^2 + r\sigma_{LC}^2 + \sigma_e^2 \end{aligned}$$

$$
\begin{aligned}
E[CM(Tes)] &= rl\sigma_T^2 + r\sigma_{LT}^2 + \sigma_e^2 \\
E[CM(Cruza \text{ vs } Test)] &= rl(\sigma_C^2 + \sigma_T^2 - \sigma_G^2) + r(\sigma_{LC}^2 + \sigma_{LT}^2 - \sigma_{LG}^2) + \sigma_e^2 \\
E[CM(Gen \times Loc)] &= r\sigma_{LG}^2 + \sigma_e^2 \\
E[CM(Cruza \times Loc)] &= r\sigma_{LC}^2 + \sigma_e^2 \\
E[CM(Tes \times Loc)] &= r\sigma_{LT}^2 + \sigma_e^2 \\
E[CM((Cruza \text{ vs } Tes) \times Loc)] &= r(\sigma_{LC}^2 + \sigma_{LT}^2 - \sigma_{LG}^2) + \sigma_e^2 \\
E[CM(Error)] &= \sigma_e^2
\end{aligned}
$$

En la (Tabla 1) se presentan cada uno de los valores de las componentes de varianza estimadas (C.V.E) de efectos principales e interacciones, así como de sus particiones para la variable por ciento de mala cobertura del experimento de Martínez (1993). En este trabajo se obtuvo no sólo la magnitud relativa de las varianzas, sino también los porcentajes de la suma de las varianzas como se observa en la (Tabla 2).

Tabla 1. Estimados de Componentes de Varianza por Fuente de Variación.

F.V.	**G.L.**	**C.V.E.**
localidad	2	117.408
rep/Loc	3	2.088
Genotipos	163	7.918
cruzas	159	7.470
Testigos	3	13.650
Cruzas vs testigos	1	61.969
Gen * Loc	326	12.355
cruzas * Loc	318	11.848
testigos * Loc	6	36.691
(cruza vs test) * Loc	2	19.971
error	489	40.083
total	983	331.451

Tabla 2. Porcentajes de Componentes de Varianza para los efectos considerados en el modelo.

σ^2_L	$\frac{117.408\times100}{331.451}$	35.42%
$\sigma^2_{R/L}$	$\frac{2.088\times100}{331.451}$	0.62%
σ^2_G	$\frac{7.918\times100}{331.451}$	2.38%
σ^2_C	$\frac{7.470\times100}{331.451}$	2.25%
σ^2_T	$\frac{13.650\times100}{331.451}$	4.11%
$\sigma^2_{C \text{ vs } T}$	$\frac{61.969\times100}{331.451}$	18.69%
σ^2_{G*L}	$\frac{12.355\times100}{331.451}$	3.72%
σ^2_{C*L}	$\frac{11.848\times100}{331.451}$	3.57%
σ^2_{T*L}	$\frac{36.691\times100}{331.451}$	11.06%
$\sigma^2_{(C \text{ vs } T)*L}$	$\frac{19.971\times100}{331.451}$	6.02%
σ^2_e	$\frac{40.083\times100}{331.451}$	12.09%

6 Conclusiones

En lo que respecta a este trabajo, se obtuvieron los grados de libertad y el estadístico de prueba apropiado para probar las hipótesis nulas correspondientes (Tabla 1). Y se observa que fueron los efectos de localidad los que más contribuyeron en la respuesta (Tabla 2). Esto significa que la localidad de Celaya, Gto., fue más rendidora que la de Río Bravo, Tamps., y Gómez Palacio, Dgo. (Martínez (1993)). Como se observa en las dos tablas anteriores, los estimadores de los componentes de varianza permiten conocer los grados de variación así como la magnitud relativa de los efectos en el modelo. Es evidente que la principal fuente de variación fué debido a las localidades (Tablas 1 y 2), sin embargo, el fitomejorador requiere de la estimación de la varianza de los genotipos para predecir el comportamiento en ensayos posteriores. El programa de computación es fácil de usar y presenta resultados confiables en la comparación de grupos de tratamientos en experimentos con varias localidades.

Referencias

Martínez, P.C. (1993) Selección y Evaluación Agronómica de Líneas S_2 y S_3 de Maíz (Zea mays L.) en tres Ambientes Contrastantes. *Tesis de Licenciatura U.A.A.A.N.*

Olivares, S.E. (1998) PAREST: Un Programa Computacional para el Análisis de Parámetros de Estabilidad por el Método de Eberhart y Russell. *XVII Congreso de Fitogenética*, p. 535.

Padrón, C.E. (1996) *Diseños Experimentales con Aplicación a la Agricultura y la Ganadería*. Trillas.

Searle, S.R. (1987) *Linear Models for Unbalanced Data*, John Wiley and Sons, Inc. N.Y.

Searle, S.R., Casella, G. and McCulloch, C.E. (1992) *Variance Components*. John Wiley and Sons, Inc. N.Y.

Una Prueba de Homogeneidad de Varianzas

Blanca Rosa Pérez Salvador
Universidad Autónoma Metropolitana-Iztapalapa
Sergio De los Cobos Silva
Universidad Autónoma Metropolitana-Iztapalapa
Miguel Angel Gutiérrez Ándrade
Universidad Autónoma Metropolitana-Azcapotzalco

1 Introduction

La adecuada aplicación de algunas técnicas estadísticas como la regresión líneal y el análisis de varianza supone la igualdad de la varianza de las observaciones de los diferentes grupos en el análisis. De aquí que el tener una prueba de homogeneidad de varianzas tenga especial interés en el estudio de los modelos lineales.

Varias pruebas de homogeneidad de varianzas se han desarrollado, debido básicamente a que la estadística de prueba basado en el cociente de verosimilitud

$$\lambda = \frac{\prod_{i=1}^{k} (\hat{\sigma}^2)^{n_j/2}}{\left(\sum_{i=1}^{k} \hat{\sigma}^2/n_j\right)^{\sum n_j/2}}.$$

no tiene una función de distribución fácil de deducir. Barlett (1946) desarrolló una prueba al demostrar que para n_j suficientemente grande, la variable aleatoria $-2\ln(\lambda)$ tiene una función de distribución aproximadamente igual a una *ji-cuadrada* con $k-1$ grados de libertad, de esta manera la región de rechazo está dada por la desigualdad $-2ln(\lambda) > \chi^2_\alpha$. La estadística de la prueba de Cochram es el cociente

$$C = \frac{\max\{s_i^2\}}{\sum(n_i-1)s_i^2/\sum(n_i-1)}.$$

Esta variable aleatoria se distribuye de acuerdo a la función de distribución F de Snedekor con $n_{\max}-1$ y $\sum(n_i-1)$ grados de libertad. Una de las desventajas de estas dos pruebas es su alta sensibilidad a la falta de normalidad de los datos.

Una prueba robusta se debe a Levene (1960), la estadística de prueba que propone es:

$$W_0 = \frac{\sum_i n_i(\overline{z}_{i\cdot} - \overline{z}_{\cdot\cdot})^2/(k-1)}{\sum_i \sum_j (z_{ij} - \overline{z}_{\cdot\cdot})^2 / \sum_i (n_i - 1)}$$

donde $z_{ij} = |x_{ij} - \overline{x}|$, $\overline{z}_{i\cdot} = \sum_j z_{ij}/n_i$ y $\overline{z}_{\cdot\cdot} = \sum_j \sum_i z_{ij} / \sum n_i$. Esta estadística se distribuye de acuerdo a una ley F de Snedekor con $k-1$ y $\sum(n_i - 1)$ grados de libertad.

En este trabajo se propone una prueba alternativa basada en las estadísticas de orden de las varianzas muestrales. La estadística de prueba es relativamente simple de construir, pero su función de distribución no es simple de expresar, aunque para algunos casos particulares es posible obtenerla por integración numérica o de manera analítica.

2 Una prueba de homogeneidad de varianzas

2.1 La estadística de prueba

Considere que se tienen k muestras independientes con n_i observaciones en la muestra i, $i = 1, 2, 3, \ldots, k$. Se sabe que, independientemente de la función de distribución que tengan las variables aleatorias x_{ij}, un estimador insesgado de la varianza σ_i^2 es $s_i^2 = \frac{\sum_{=1}^{n_i} i(x_i - \overline{x})^2}{n_i - 1}$. Entonces, para cualquier función de distribución de x_{ij}, en el caso que $\sigma_1^2 = \sigma_2^2 = \sigma_3^2 = \ldots = \sigma_k^2$ se tiene que $s_1^2, s_2^2, s_3^2, \ldots, s_k^2$ forman una muestra aleatoria tal que $E(s_i^2) = \sigma_i^2$. De esta manera, para probar

$$H_0 : \ \sigma_1^2 = \sigma_2^2 = \sigma_3^2 = \ldots = \sigma_k^2$$

contra

$$H_1 : \text{ al menos una } \sigma_i^2 \text{ es diferente}$$

se puede dar una estadística de prueba basada en los valores extremos de esta muestra. La estadística de prueba es el cociente

$$B = \frac{s_{(n)}^2}{s_{(1)}^2},$$

donde $s_{(n)}^2 = \max_i\{s_i^2\}$ y $s_{(1)}^2 = \min_i\{s_i^2\}$. Se tendrá evidencia para rechazar H_0 cuando B tenga un valor "grande".

La región de rechazo de la prueba está dada por la relación $\frac{s_{(n)}}{s_{(1)}} > r$ con r un valor tal que

$$P\left(\frac{s_{(n)}}{s_{(1)}} > r \mid H_0 \text{ es cierta}\right) = 1 - F_B(r|\sigma_1^2 = \sigma_2^2 = \ldots = \sigma_k^2) = \alpha.$$

La función de densidad de B es:

$$\begin{aligned} F_B(r \mid \sigma_1^2 = \sigma_2^2 = \ldots = \sigma_k^2) &= P\left(\frac{s_{(n)}}{s_{(1)}} < r\right) \\ &= P\left(\cup_{i=1}^{k}\left(\cap_{j\neq i}\{1 \leq \frac{s_j^2}{s_i^2} \leq r\}\right)\right) \\ &= \sum_{i=1}^{k} P\left(\cap_{j\neq i}\{1 \leq \frac{s_j^2}{s_i^2} \leq r\}\right) \\ &= \sum_{i=1}^{k}\int_0^\infty \prod_{j\neq i}\left(F_j\left(r\frac{n_i-1}{n_j-1}t_i\right) - F_j\left(\frac{n_i-1}{n_j-1}t_i\right)\right) f_i(t_i)dt_i, \end{aligned} \tag{1}$$

donde $F_j(x)$ es la función de distribución de $\frac{(n_j-1)s_j^2}{\sigma_j^2}$.

Esta expresión se puede calcular algebraicamente o con integración numérica para casos particulares de diferentes funciones de distribución $F_i(x)$, en particular se puede hacer esto cuando se supone que los datos se distribuyen de acuerdo a una ley normal.

2.2 Caso de normalidad

Cuando se supone normalidad en los datos, se tiene que $F_j(x)$ es función de distribución *ji-cuadrada* con $n_j - 1$ grados de libertad. En este caso, se puede probar que si n_j es un número impar,

$$F_j(x) = \frac{1}{(n_j-3)/2)!2^{(n_j-3)/2}}A(x) \tag{2}$$

donde

$$A(x) = 2^{(n_j-1)/2}(n_j-3)/2)! - \left(2x^{(n_j-3)/2} + 2\sum_{i=0}^{(n_j-3)/2}\frac{(n_j-3)/2)!}{((n_j-3)/2-i)!}x^{(n_j-3)/2-i}\right)e^{-\frac{1}{2}x}$$

y cuando n_j es par

$$F_j(x) \approx \frac{1}{(\Gamma(n_j-1)/2)2^{(n_j-3)/2}}B(x) \tag{3}$$

donde

$$B(x) = 2^{(n_j-1)/2}\Gamma(n_j-1)/2) - \left(2x^{(n_j-3)/2} + 2\sum_{i=0}^{(n_j-3)/2}\frac{\Gamma((n_j-1)/2)}{\Gamma((n_j-3)/2-i)}x^{(n_j-3)/2-i}\right)e^{-\frac{1}{2}x}$$

$$+\frac{\pi}{2}e^{-(1-\frac{1}{2\sqrt{2}})x}$$

Al sustituir (2) o (3) en (1) y desarrollar el producto en el integrando, se tendrá una suma de términos que tienen primitivas algebraicas. Para casos particulares, se puede deteminar el valor de r y reportarlo en tablas. Así, cuando todas las muestras tienen el mismo tamaño, se tiene que

$$F_B(r \,|\, \sigma_1^2 = \sigma_2^2 = \ldots = \sigma_k^2) = k \int_0^\infty (F(rt) - F(t))^{k-1} f(t) dt$$

El valor de r para algunos casos particulares de igual tamaño de las muestras para un nivel de significancia igual a 0.05, se reportan en la siguiente tabla.

Tamaño de cada muestra n_i	Número de muestras k	Valor de r
5	4	6.5
5	10	8.7
10	4	4.1
10	10	5.7

2.3 Sobre la potencia de la prueba

En el caso que la hipótesis nula no sea cierta, se puede considerar sin perder generalidad que

$$\sigma_1^2 \leq \sigma_2^2 \leq \sigma_3^2 \leq \ldots \leq \sigma_k^2 \qquad \sigma_1^2 < \sigma_k^2,$$

y entonces la potencia de la prueba es

$$P\left(\frac{s_{(n)}^2}{s_{(1)}^2} > r \,|\, \sigma_1^2 < \sigma_n^2\right) \geq P\left(\frac{s_n^2}{s_1^2} > r \,|\, \sigma_1^2 < \sigma_n^2\right) = 1 - F\left(\frac{\sigma_1^2}{\sigma_n^2} r\right),$$

donde $F(x)$ es la función de distribución de $\frac{s_n^2}{s_1^2}$.

3 Conclusión

Se está presentando una prueba alternativa de la homogeneidad de varianzas cuya región de rechazo se puede calcular para algunos casos particulares. La estadística de prueba es simple de construir y es posible elaborar una tabla de los valores críticos para los casos de igual tamaño de las muestras y elaborar un pograma de cómputo para calcularlos en casos más generales. Se espera realizar un análisis comparativo con la prueba de Levene para determinar las bondades de esta prueba.

Bibliografía

Barlett, M.S. and Kendal, D.G.(1946) The Statistical Analysis of Variance-Heterogeneity and the Logarithic Transformation. *Supplement to the Journal of the Royal Statistical Society*, Ser. B, **1**, pp 128-138.

Levene H. (1960) Robust Tests for Equality of Variances. Olkin, ed.; *Contribution of Probability and Statistics*, Palo Alto, Calif; Stanford University, Press; pp 278-292.

Un Índice de Exposición de Contaminación Atmosférica en un Estudio de Efectos a la Salud del Sistema de Vigilancia Epidemiológica en la Zona Metropolitana de la Ciudad de México

Silvia Ruiz-Velasco Acosta
IIMAS, UNAM

1 Introducción

La contaminación atmosférica es un problema complejo en la Ciudad de México. Esto ha llevado a medidas de control no del todo satisfactorias. La Secretaría de Salud participa cuantificando los efectos de la contaminación en la salud (respiratoria). El sistema de vigilancia epidemiológica ambiental tiene como objetivo el evaluar los daños a la salud de altos niveles de contaminación, por tal motivo a partir de 1996 se han realizado entrevistas aleatorias diarias en los radios de incidencia de 5 monitores de la Red Automática de Monitoreo Atmosférico. El objetivo de este estudio es medir el efecto de la contaminación en síntomas agudos, en población abierta, esto a través de modelos de regresión Poisson. Dado que los diversos contaminantes están altamente correlacionados, el considerar el efecto de más de un contaminante simultáneamente puede ocasionar problemas de multicolinealidad, aunque es posible considerar el efecto de un contaminante en altos niveles de otro, introduciendo este por medio de una variable indicadora. Este trabajo crea un índice de contaminación por medio de componentes principales comunes, sin considerar la dependencia en el tiempo de cada contaminante. Se evalúa el comportamiento de este índice en los distintos monitores y se utiliza para medir el efecto de la contaminación en la salud corrigiendo por variables demográficas.

Los contaminantes que se consideran dañinos para la salud son: Partículas Suspendidas Respirables (PM_{10} y $PM_{2.5}$), Ozono (O_3), Bióxido de Azufre (SO_2), Bióxido de Nitrógeno (NO_2), Monóxido de Azufre (CO), Oxido de Nitrógeno (NO_x). Algunos de los efectos conocidos en cuanto a exposición a partículas son: incrementa padecimientos cardiacos previos, incrementa padecimientos pulmonares previos, alergias, tos. La exposicion a monóxido de carbono puede causar insuficiencia respiratoria e incluso la muerte. Algunos de los efectos de exposición a ozono son: reducción de funciones pulmonares, tos, resequedad e irritación

de garganta, agravamiento de enfermedades crónicas. Mientras que la esposicion bióxido de azufre puede causar asma.

El programa de Contingencias Ambientales, estableció los niveles máximos permitidos, cuando se sobrepasan (una o varias veces) esos niveles se denominan días de contingencia ambiental. Estos niveles fueron convertidos a una medida denominada IMECA. Para cada uno de estos contaminantes los niveles máximos permitidos o normas son: $PM_{10} < 150 ug/m^3$ promedio de 24 hrs; $CO < 11 ppm$ promedio de 8 hrs; $NO_2 < 210 ppb$ promedio horario máximo; $O_3 < 110 ppb$ promedio horario máximo; $SO_2 < 130 ppb$ promedio móvil de 24 hrs.

2 Estudio

El objetivo del estudio es crear un índice de contaminación ambiental que permita estudiar la variación en el comportamiento de los síntomas sub-agudos agudos y crónicos de la población residente en seis zonas diferentes del área metropolitana de la Ciudad de México y estableces umbrales para estándares. En este trabajo sólo se reportan resultados para síntomas agudos. El área de estudio seleccionada corresponde a todas las viviendas en un radio de dos kilometros de los monitores localizados en: Merced, Pedregal, Plateros, Cerro de la Estrella, Xalostoc y Tlanepantla. Es decir, el marco muestral corresponde a todas las viviendas en un radio de dos kilómetros de cada uno de estos monitores.

Cada día de estudio se tomo una muestra de las viviendas localizadas alrededor de cada monitor y se realizo una entrevista a todos lo habitantes de la casa que estuvieran presentes en ese momento y que fueran mayores de 12 años, en el caso de menores la madre (o responsable) contesto la entrevista.

Los datos utilizados en este trabajo corresponden al período del 1 de enero de 1996 al 31 de diciembre de 1997. Un total de 148,885 entrevistas fueron realizadas en dicho período. Las preguntas concernientes a salud respiratoria aguda, es decir, si el dia de hoy presentaban ese síntoma fueron: tos con flema, tos seca, dolor de garganta, falta de aire, silbidos en el pecho, afonia, resequedad de nariz, catarro, comezón de ojos, ardor de ojos, infección de ojos, dolor de oídos, secreción de oídos, dolor de cabeza, fiebre, dolor de articulaciones, diarrea, sueño. Concernientes a la salud respiratoria sub-agudas fueron, si en la última semana se presento: conjuntivitis, catarro, respiratoria aguda, faringitis y bronquitis

A su vez se definieron otras medidas agudas: ERA (Enfermedad Respiratoria Alta), ERB (Enfermedad Respiratoria Baja) y Ojos (Irritaciones en los Ojos). En cuanto a sub-agudas se definieron: Alta (Infecciones Respiratorias Altas), Baja (Infecciones Respiratorias Bajas)

y Aguojo (Infecciones Agudas de los Ojos).

Para la construcción del índice de contaminación, se emplearon componentes principales comunes.

3 Componentes principales

Los componentes principales son combinaciones lineales de las variables originales que cumplen que el primer componente tiene la varianza mayor y que son no correlacionados. En otras palabras lo que queremos en encontrar una combinación lineal de $Y = a'X$, tal que su varianza sea lo más grande posible, sujeto a $a'a = 1$. Si suponemos una variable aleatoria X tal que:

$$E(X) = 0,$$

$$E(XX') = \Psi.$$

La descomposición espectral de Ψ está dada por

$$\Psi = B \wedge B$$

donde $B(\beta_1, \cdots, \beta_p)$ y $\wedge = diag(\lambda_1, \cdots, \lambda_p)$

Si a es una combinación lineal de las $\beta' s$, se puede demostrar que su varianza siempre es menor o igual que λ_1 y que sólo es igual cuando $a = \beta_1$. Es decir, los componentes principales son combinaciones lineales de las variables, en donde los coeficientes corresponden a los vectores propios y las varianzas a los valores propios de la matriz de varianzas y covarianzas de las variables originales, entonces el primer componente estara definido por el valor propio mayor.

4 Componentes principales comunes

Los componentes principales comunes, están definidos como combinaciones lineales de las variables originales, en distintas poblaciones, de manera que cumplen:

$$\Psi_i = B \wedge_i B,$$

es decir, suponemos que los coeficentes de las combinaciones lineales de las variables originales son las mismas en las distintas poblaciones, pero su varianza es diferente. Una manera de estimar estos componentes es por máxima verosimilitud, suponiendo que $S_i \tilde{W}(n_i, \Psi_i/n_i)$. Una vez obtenidos podemos probar la hipótesis $H_0 = \Psi_i = B \wedge_i B$ con la estadística de prueba

$$\Sigma n_i \log\left(\frac{det\Psi_i}{detS_i}\right).$$

Pero en la situación que tenemos: la distribución de las Si no es Wishart y no existe una transformación a normalidad que sea común y que pase una prueba de normalidad. Sin embargo podemos obtener los coeficientes de los componentes usando las ecuaciones de la verosimilitud, ya que no vamos a hacer inferencia sobre los coeficientes en si, si no lo que buscamos es la construccion de un índice. En este trabajo se presenta el índice considerando el nivel máximo de ozono, NOX y $NO2$; el promedio de 24 hrs de $PM10$ y $SO2$ y el máximo del promedio móvil de 8 hrs de CO. El componente está dado por:

$$\begin{aligned} CPCD &= 0.433 * COPRO + 0.420 * PM10PRO + 0.420 * SO2PRO + \\ &\quad 0.435 * NO2MAX + 0.294 * NOXMAX + 0.428 * O3MAX \end{aligned}$$

El porcentaje de varianza que este componentes explica en los distintos monitores y su valor medio es:

	Monitor 1	Monitor 2	Monitor 3	Monitor 4	Monitor 5	Monitor 7
% varianza	0.7856	0.7884	0.7629	0.7735	0.8140	0.7997
media	1451.89	1266.78	1443.81	1359.75	943.99	1042.95

Una manera de estudiar la relación de los contaminantes con la salud respiratoria es por medio de un modelo de regresión Poisson. Utilizando como variable respuesta el número de individuos que manifestaron tener el síntoma referido y como variable explicativa el índice de contaminación, corrigiendo por temperatura mínima, el monitor, un polinomio de la fecha como una variable de tiempo y si hubo contingencia ambiental o no, además de utilizar como "offset" el número de entrevistas realizadas:
El modelo ajustado es:

$$\text{In}(casos) = \alpha + \beta_1 CPCD + \beta_2 temp + \beta_3 con + \beta_m + \beta_1 t + \beta_{12} t^2 + \beta_{13} t^3 + \text{In}(total)$$

La interpretación de estos modelos es en términos de $\exp(\beta_1)$ lo que representa el incremento en el riesgo de tener el síntoma en cuestión cuando CPCD aumenta en una unidad, manteniendo todas las demás variables constantes. Entonces $\exp(\beta_1 x)$ representa el incremento en el riesgo de tener el síntoma cuando CPCD aumenta en x. Algunos de los ajustes son:

ERA		cambio en 200 (95%I.C.)			cambio en 300 (95%I.C.)	
CPC (lineal)	1.0127	1.0081	1.0174	1.0192	1.0122	1.0262
CPC_1 (lineal)	1.0105	1.0064	1.0146	1.0157	1.0096	1.0220
CPC_2 (lineal)	1.0086	1.0045	1.0126	1.0129	1.0067	1.0190

ERB

CPC (lineal)	1.0158	1.0087	1.0229	1.0238	1.0130	1.0346
CPC_1 (lineal)	1.0104	1.0041	1.0167	1.0157	1.0062	1.0252
CPC_2 (lineal)	1.0100	1.0038	1.0162	1.0150	1.0057	1.0244

OJOS

CPC (lineal)	1.0167	1.0114	1.0219	1.0251	1.0172	1.0331
CPC_1 (lineal)	1.0543	1.0236	1.0859	1.0825	1.0355	1.1316
CPC_2 (lineal)	1.0101	1.0054	1.0147	1.0151	1.0082	1.0221

SUEÑO

CPC (lineal)	1.0202	1.0139	1.0295	1.0305	1.0209	1.0446
CPC_1 (lineal)	1.0093	1.0026	1.0160	1.0140	1.0038	1.0242
CPC_2 (lineal)	1.0429	1.0315	1.0545	1.0651	1.0476	1.0828

Dolor de cabeza

CPC (lineal)	1.0145	1.0075	1.0214	1.0218	1.0113	1.0323
CPC_1 (lineal)	1.0109	1.0048	1.0171	1.0164	1.0071	1.0258
CPC_2 (lineal)	1.0064	1.0003	1.0125	1.0096	1.0004	1.0189

Realizando pruebas post-ajuste, se puede ver que no existen observaciones influyentes ni discrepantes, una gráfica de residuales deja ver que su comportamiento es adecuado, aunque en el algunos síntomas existe una mejora al considerara polinomios de tercer grado en el componente. En todos los casos existe sobre-dispersión, sin embargo ajustando un modelo de regresión binomial negativo, es posible ver que las conclusiones no se cambian.

También es posible extender los resultados de componentes principales para series de tiempo a un contexto de componentes principales comunes en series de tiempo.

5 Anexo

Durante el período de estudio los valores de los contaminantes fueron:

	mo	pp	mn	mnx	mc	ps
1	95.18(40.39)	143.54(59.04)	84.09(48.69)	302.63(130.37)	44.40(17.38)	26.00(14.02)
2	113.51(50.54)	91.15(50.19)	94.23(52.37)	236.93(134.48)	41.06(18.33)	23.28(12.22)
3	116.36(47.10)	92.22(41.05)	97.09(50.45)	255.87(120.09)	43.10(14.67)	20.51(13.62)
4	119.15(42.66)	83.48(32.61)	59.42(27.72)	95.58(69.14)	36.82(13.40)	15.27(11.05)
5	168.09(51.36)	64.85(16.93)	94.78(39.77)	204.19(10.10)	28.01(9.12)	15.06(5.72)
7	134.71(58.72)	63.17(35.97)	107.07(56.99)	195.63(95.63)	33.34(16.01)	14.83(7.92)

Procedimiento para Estratificación Mediante Particiones Sucesivas en Función de Sumas de Cuadrados Dentro de Estratos

Francisco Sánchez Villarreal
Facultad de Ciencias, UNAM

1 Introducción

Es conocido para los muestristas las ganancias en eficiencia que se obtienen al estratificar adecuadamente una población antes de proceder a la selección de unidades de muestreo. El objetivo subyacente es identificar y controlar las fuentes de varianza para lograr el máximo de precisión para un mismo tamaño de muestra. Si se dispone de una variable de estratificación medida en nivel intervalar o de razón y con elevada correlación con la variable objetivo, existen diversos métodos para estratificar. Uno de los más populares por su sencillez de aplicación es el de Dalenius-Hodge (1959). Este procedimiento procede con la formación de estratos a partir de la construcción de una tabla de frecuencias de los valores de la variable de estratificación, valores que se asocian a los elementos de la población para obtener puntos de corte que permitan definir una estructura de estratos óptima para afijaciones de muestra de Neyman (Cochran). El supuesto de estratos numerosos y pequeños permite suponer además uniformidad dentro de los estratos y en términos prácticos de cálculo, proceder con la suma acumulada de las raíces cuadradas de las frecuencias de la tabla previamente construida con intervalos de igual amplitud o ponderados si se tienen amplitudes desiguales, para determinar por división del total de la suma de raíces cuadradas de frecuencias en tantos estratos como se quiera y así identificar los límites de los estratos. El procedimiento de Dalenius, como muchos otros en estadística fueron ideados en tiempos en que existían serias limitaciones de cálculo para los investigadores, pues se contaba con primitivas máquinas de calcular de escritorio y el acceso a computadoras era muy restringido. La época actual carece de esas restricciones y en el escritorio de cualquier investigador en general se dispone de una computadora bastante poderosa. Ello ha llevado a superar los obstáculos de cálculo que restringían la utilización de técnicas más elaboradas en diversos tópicos del cómputo estadístico.

2 Efectos de la estratificación.

Las ganancias debidas a la estratificación se obtienen a medida que se logra maximizar

la varianza entre estratos y simultáneamente minimizar la varianza dentro de estratos. La variación total se puede descomponer en dos fuentes: la suma de cuadrados dentro de estratos y la suma de cuadrados entre estratos, como se ilustra en la siguiente igualdad:

Sumas de Cuadrados

Total		Dentro de Estratos		Entre Estratos
$\sum_{h=1}^{L}\sum_{i=1}^{N_h}(y_{hi}-\overline{Y})^2$	=	$\sum_{h=1}^{L}\sum_{i=1}^{N_h}(y_{hi}-\overline{Y}_h)^2$	+	$\sum_{h=1}^{L}\sum_{i=1}^{N_h}(\overline{Y}_h-\overline{Y})^2$

La ganancia aportada por la estratificación se puede apreciar en la fórmula que relaciona la varianza del estimador de la media, estimada por muestreo aleatorio simple con la varianza del estimador de la media obtenida por muestreo estratificado con afijación proporcional de la muestra.

$$V\left(\overline{y}_{\text{mas}}\right) = V\left(\overline{y}_{\text{prop}}\right) + \frac{\sum_{h=1}^{L}\sum_{i=1}^{N_h}(\overline{Y}_h-\overline{Y})^2}{nN}$$

Como se puede apreciar, a medida que la varianza entre estratos sea mayor y como consecuencia, la varianza dentro de estratos sea menor, se tendrá mayor eficiencia al utilizar muestreo estratificado.

Maximizar la varianza entre estratos se puede lograr también si se minimiza la varianza dentro de estratos, esto es, minimizar la siguiente suma de cuadrados:

$$\sum_{h=1}^{L}\sum_{i=1}^{N_h}\left(y_{hi}-\overline{Y}_h\right)^2$$

Cuando se estratifica en forma univariada, en general la variable de estratificación es una variable que presenta correlación alta con la variable objetivo o se dispone de datos de un período anterior de la variable objetivo. Así, el supuesto básico es que al lograr una buena estratificación a partir de la variable de estratificación, se logrará un efecto similar con la variable objetivo.

3 Procedimiento de particiones sucesivas

En forma empírica la minimización de la suma de cuadrados dentro de estratos se logra con el siguiente procedimiento que llamaremos de particiones sucesivas.

- Se parte de una serie ordenada descendente o ascendente de los valores de la variable de estratificación.

- Se supone que inicialmente se tienen dos grupos o estratos, uno que incluye los N elementos de la población y otro que incluye 0 elementos.

- Se calcula la aportación de cada uno de los dos grupos a la suma de cuadrados, en el paso inicial o final, uno de los dos estratos incluye la suma de cuadrados total y el otro tiene contribución nula.

$$\sum_{i=1}^{N_h} \left(y_{hi} - \overline{Y}_h\right)^2$$

- Se excluyen, uno cada vez, los elementos del primer estrato y se incluyen simultáneamente en el segundo grupo. En el paso K el primer estrato tiene $N-K$ elementos y el segundo grupo K elementos. El proceso continua hasta que el primer grupo tenga 0 elementos y el segundo tenga asociados los N elementos.

- En cada paso se suman las aportaciones de los dos grupos a la suma de cuadrados:

$$\sum_{i=1}^{N_1} \left(y_{1i} - \overline{Y}_1\right)^2 + \sum_{i=1}^{N_2} \left(y_{2i} - \overline{Y}_2\right)^2$$

- Se observará que la suma de cuadrados disminuye desde los extremos hasta un punto, generalmente alejado del centro del grupo de observaciones, en el cual la suma de cuadrados es mínima.

- Ese punto será la frontera para definir dos estratos.

El investigador puede proceder en forma similar con el grupo que haga la mayor aportación a la suma de cuadrados, si es que el número de elementos no es demasiado pequeño, en cuyo caso convendrá censar el estrato, se tendrían entonces 3 estratos. Desde luego se puede proceder a particionar los dos grupos iniciales y entonces tener 4 estratos o más si se considera conveniente.

El procedimiento es engorroso en apariencia, pues se puede lograr en forma relativamente fácil con la ayuda de las funciones estadísticas para cálculo de varianzas de una hoja electrónica o con ayuda de un programa específico que un programador capacitado puede lograr con relativa facilidad.

Cuadro 1

FASE 1 DE ESTRATIFICACION									FASE 2 DE ESTRATIFICACION		
Estrato	Clave	Nombre de Entidad	Población Total 1995	Numeración Auxiliar	Numeración Auxiliar	Sumas de Cuadrados Grupo 1	Sumas de Cuadrados Grupo 2	Sumas de Cuadrados Dentro de Grupos	Sumas de Cuadrados Grupo 1	Sumas de Cuadrados Grupo 2	Sumas de Cuadrados Dentro de Grupos
					33	-	196,359,583,049,256	196,359,583,049,256			
1	15	Estado de México	11,707,964	1	32	-	190,409,292,653,824	190,409,292,653,824			
1	9	Distrito Federal	8,489,007	2	31	5,180,842,083,925	109,390,846,688,252	114,571,688,772,176			
1	30	Veracruz	6,737,324	3	30	12,712,446,503,313	73,101,646,008,198	85,814,092,511,511			
1	14	Jalisco	5,991,176	4	29	19,403,725,272,337	53,328,627,441,673	72,732,352,714,010			
1	**21**	**Puebla**	**4,624,365**	**5**	**28**	**29,812,100,343,143**	**38,556,885,723,235**	**68,368,986,066,378**			
2	11	Guanajuato	4,406,568	6	27	37,838,005,838,610	31,841,968,867,602	69,679,974,706,212	-	31,841,968,867,602	31,841,968,867,602
2	16	Michoacán	3,870,604	7	26	46,193,173,613,096	25,754,945,030,095	71,948,118,643,190	-	25,754,945,030,095	25,754,945,030,095
2	7	Chiapas	3,584,786	8	25	53,869,571,310,454	21,685,040,977,699	75,554,612,288,153	143,628,704,648	21,685,040,977,699	21,828,669,682,347
2	19	Nuevo León	3,550,114	9	24	60,000,920,788,480	18,415,941,744,193	78,416,862,532,674	348,091,664,648	18,415,941,744,193	18,764,033,408,841
2	20	Oaxaca	3,228,895	10	23	66,348,682,753,298	14,994,846,101,190	81,343,528,854,489	470,426,108,936	14,994,846,101,190	15,465,272,210,126
2	12	Guerrero	2,916,567	11	22	72,988,299,241,186	12,423,867,311,370	85,412,166,552,556	782,049,724,239	12,423,867,311,370	13,205,917,035,609
2	8	Chihuahua	2,793,537	12	21	79,089,338,349,083	10,578,707,030,248	89,668,045,379,331	1,330,997,568,553	10,578,707,030,248	11,909,704,598,802
2	28	Tamaulipas	2,527,328	13	20	85,479,418,462,973	8,892,348,052,131	94,371,766,515,104	1,878,726,349,538	8,892,348,052,131	10,771,074,401,668
2	25	Sinaloa	2,425,675	14	19	91,424,723,400,839	7,699,640,006,506	99,124,363,407,345	2,670,737,118,299	7,699,640,006,506	10,370,377,124,805
2	24	San Luis Potosí	2,200,763	15	18	97,610,981,899,220	6,603,957,961,556	104,214,939,860,776	3,446,372,035,716	6,603,957,961,556	10,050,329,997,272
2	5	Coahila	2,173,775	16	17	103,146,232,745,096	5,838,022,275,617	108,984,255,020,713	4,448,560,266,836	5,838,022,275,617	10,286,582,542,453
2	13	Hidalgo	2,112,473	17	16	108,296,677,811,472	5,027,216,860,502	113,323,894,671,974	5,315,796,525,614	5,027,216,860,502	10,343,013,386,116
2	2	Baja California	2,112,140	18	15	112,876,236,177,096	4,225,678,113,356	117,101,914,290,451	6,141,728,509,745	4,225,678,113,356	10,367,406,623,101
2	26	Sonora	2,085,536	19	14	117,079,238,765,717	3,310,293,514,773	120,389,532,280,490	6,841,129,058,553	3,310,293,514,773	10,151,422,573,326
3	27	Tabasco	1,748,769	20	13	122,246,481,142,728	2,310,065,576,995	124,556,546,719,723	**7,480,970,968,663**	**2,310,065,576,995**	**9,791,036,545,658**
3	31	Yucatán	1,556,622	21	12	127,767,664,233,360	1,777,562,132,095	129,545,226,365,455	8,625,903,735,044	1,777,562,132,095	10,403,465,867,139
3	17	Morelos	1,442,662	22	11	133,298,203,190,640	1,426,378,670,492	134,724,581,861,132	10,034,761,648,544	1,426,378,670,492	11,461,140,319,036
3	10	Durango	1,431,748	23	10	138,395,911,984,675	1,145,853,190,488	139,541,765,175,163	11,536,626,198,668	1,145,853,190,488	12,682,479,389,156
3	32	Zacatecas	1,336,496	24	9	143,480,646,371,306	816,329,042,648	144,296,975,413,954	12,896,239,292,856	816,329,042,648	13,712,568,335,504
3	22	Querétaro	1,250,476	25	8	148,530,285,687,374	523,898,584,326	149,054,184,271,700	14,325,842,679,259	523,898,584,326	14,849,741,263,585
3	18	Nayarit	896,702	26	7	154,809,752,678,685	252,234,052,723	155,061,986,731,409	15,809,719,471,070	252,234,052,723	16,061,953,523,794
3	29	Tlaxcala	883,924	27	6	160,684,702,201,652	203,954,078,627	160,888,656,280,279	18,071,539,691,051	203,954,078,627	18,275,493,769,679
3	1	Aguascalientes	862,720	28	5	166,237,711,874,949	143,444,500,285	166,381,156,375,234	20,163,699,114,835	143,444,500,285	20,307,143,615,120
3	23	Quitana Roo	703,536	29	4	172,143,527,217,419	66,402,471,003	172,209,929,688,422	22,131,685,936,856	66,402,471,003	22,198,088,407,859
3	4	Campeche	642,516	30	3	177,940,928,674,921	35,943,730,595	177,976,872,405,516	24,378,562,692,060	35,943,730,595	24,414,506,422,655
3	6	Colima	488,028	31	2	184,095,247,838,178	6,331,950,578	184,101,579,788,756	26,621,181,811,475	6,331,950,578	26,627,513,762,053
3	3	Baja California Sur	375,494	32	1	190,409,292,653,824	-	190,409,292,653,824	29,150,158,347,583	-	29,150,158,347,583
				33		196,359,583,049,256	-	196,359,583,049,256	31,841,968,867,602	-	31,841,968,867,602
			91,158,290								

4 Ejemplo con entidades federativas

Como ejemplo para ilustrar en forma resumida el procedimiento, se tomaron los datos de población total por entidad federativa reportados por el INEGI en el Conteo de Población y Vivienda de 1995. Los datos de las 32 entidades, ordenados de acuerdo al número de habitantes, presentan una distribución muy asimétrica, como suelen ser en la práctica las poblaciones sometidas a muestreos (Cuadro 1).

En el mismo Cuadro 1 se puede observar la Fase 1 del procedimiento, que resultará en dos estratos. Una vez ordenados los estados de mayor a menor población, se procede a calcular las aportaciones de cada grupo a la suma de cuadrados dentro de estratos o grupos. La suma de cuadrados total es 190,409,292,653,824 unidades cuadráticas. Este valor se observa en los extremos del Cuadro 1. La suma de cuadrados desciende en la columna del extremo derecho del cuadro y su valor mínimo (68,368,986,066,378) lo alcanza al incluir en el primer grupo del Estado de México a Puebla. Los 27 estados restantes configuran entonces el segundo grupo o estrato. El primer grupo de 5 estados aporta 29,812,100,343,143 y el segundo 38,556,885,723,235.

Cuadro 2

FASE 3 DE ESTRATIFICACION								
Estrato	Clave	Nombre de Entidad	Población Total 1995	Numeración Auxiliar	Numeración Auxiliar	Sumas de Cuadrados Grupo 1	Sumas de Cuadrados Grupo 2	Sumas de Cuadrados Dentro de Grupos
1	15	Estado de México	11,707,964	1	5	-	29,812,100,343,143	29,812,100,343,143
1	9	Distrito Federal	8,489,007	2	4	5,180,842,083,925	7,783,128,927,130	12,963,971,011,055
4	30	Veracruz	6,737,324	3	3	12,712,446,503,313	2,296,501,627,769	15,008,948,131,081
4	14	Jalisco	5,991,176	4	2	19,403,725,272,337	934,086,154,861	20,337,811,427,197
4	21	Puebla	4,624,365	5	1	29,812,100,343,143	-	29,812,100,343,143

En la Fase 2 (Cuadro 1) se incluyen los 27 estados del segundo grupo original. Al repetir el procedimiento se observa que la suma de cuadrados alcanza su menor valor en 9,791,036,545,658 y que corresponde a la posición del estado de Sonora. Si en este punto se suspende el proceso se tendrían 3 estratos. El primero con 5 elementos, el segundo con 14 y el tercero con 13 elementos.

Si se continúa el proceso una etapa más (Cuadro 2), se procede a dividir el primer estrato (por ser el que más aporta a la suma de cuadrados). El resultado es la formación de un estrato que contiene al Estado de México y al Distrito Federal y otro que contiene a Veracruz, Jalisco y Puebla. Finalmente se detiene el proceso en 4 estratos con 2, 3, 14 y 13 elementos. Es

evidente que el proceso se puede continuar aunque la ganancia marginal disminuye en cada etapa.

Por diferencia de la suma de cuadrados inicial de 190,409,292,653,824, la suma de cuadrados entre estratos asciende a 167,654,285,097,112. Esto es el 88 % de la variación total se debe a variació entre estratos y el 12% corresponde a variación dentro de estratos (22,755,007,556,712). Para mayor detalle consulte el Cuadro 3.

Cuadro 3

Estrato	Aportación a la Suma de Cuadrados	Porcentaje
1	5,180,842,083,925	2.7%
2	7,783,128,927,130	4.1%
3	7,480,970,968,663	3.9%
4	2,310,065,576,995	1.2%
Suma de Cuadrados Dentro de Estratos	22,755,007,556,712	12.0%
Suma de Cuadrados Entre Estratos	167,654,285,097,112	88.0%
Suma de Cuadrados Total	190,409,292,653,824	100.0%

En el ejemplo se utilizó como variable de estratificacin el número de habitantes, el efecto sobre la variable que sea objeto de la estimación será mayor a medida que haya mayor correlación entre ambas.

Calibración en Muestreo

Mónica Tinajero Bravo
IIMAS, UNAM
Guillermina Eslava Gómez
IIMAS, UNAM

1 Introducción

Con el fin de mejorar las estimaciones obtenidas a partir de una encuesta es común utilizar información auxiliar en la etapa de estimación, por ejemplo, estimadores de razón, estimadores de regresión. Estos estimadores son el resultado de ajustar el ponderador original mediante un factor, y pertenecen a una familia más general conocida como estimadores obtenidos mediante la calibración de los pesos muestrales. Estos nuevos pesos deben cumplir dos condiciones: i) estar tan cerca como sea posible de los pesos muestrales originales, de acuerdo a una medida de distancia, ii) satisfacer un conjunto de restricciones. Estas se refieren a la información auxiliar, proveniente de censos, registros administrativos, estadísticas vitales, etc., así como en algunos casos al rango en el que deben estar los ponderadores calibrados. Un planteamiento formal del problema es el siguiente:

Notación,

$U = \{1, \cdots, N\}$	Una población finita con N elementos.
$s \subset U$	Una muestra obtenida bajo un diseño de muestreo probabilístico.
$p(s)$	Probabilidad de que la muestra s sea seleccionada.
$\pi_k = P(k \in s)$	Probabilidad de que la unidad k pertenezca a la muestra.
$d_j = 1/\pi_k$	Peso de la unidad k asociado al diseño de muestreo, también conocido como factor de expansión.
y_k	Es el valor de la variable de interés y para el elemento k de la pob.
$\mathbf{x}_k = (x_{k1}, \cdots, x_{kN})'$	Vector auxiliar de valores para el elemento k.
$\mathbf{t}_k = \sum_U \mathbf{x}_k$	Total poblacional para el vector de variables auxiliares $\mathbf{x}$.
w_k	Peso muestral calibrado de la unidad muestral k.

La calibración puede aplicarse, en principio, a la estimación de cualquier parámetro, sin embargo el planteamiento más utilizado corresponde a la estimación de totales.

Entonces, el objetivo es estimar el total poblacional $t_y = \sum_U y_k$, mediante $\hat{t}_{yw} = \sum_{k \in s} w_k y_k$, donde los nuevos pesos se obtienen minimizando la distancia promedio $E\{\sum_s F^*(w_k, d_k) =$

$\sum_s \frac{d_k}{q_k} F(g_k)\}$, sujeta a las restricciones $\mathbf{t}_x = \sum_{k \in s} w_k \mathbf{x}_k$ (ecuaciones de calibración).

La razón entre el ponderador calibrado y el original, $g_k = w_k/d_k$, es el factor de ajuste conocido como 'factor-g' y $1/q_k$ es un peso positivo conocido no relacionado con d_k. En la práctica es común usar $q_k = 1$. Resolviendo el problema de minimización mediante multiplicadores de Lagrange, se obtiene:

$$w_k = d_k g_k = d_k g(q_k \mathbf{x}_k' \lambda)$$

donde $g(z)$ es la función inversa de $f(z) = \frac{\partial F(z)}{\partial z}$ y λ debe satisfacer el conjunto de ecuaciones:

$$\sum_s w_k \mathbf{x}_k = \sum_s d_k g(q_k \mathbf{x}_k' \lambda) \mathbf{x}_k = \mathbf{t}_x. \tag{1}$$

Una aplicación importante de esta técnica es cuando los elementos del vector de totales corresponden a los totales marginales de una tabla de contingencia.

Los estimadores calibrados fueron desarrollados por Deville y Särndal (1992). Sin embargo, uno de los primeros trabajos en el caso de tablas de contingencia corresponde a Deming y Stephan (1940), cuyo objetivo era estimar las frecuencias de las celdas de una tabla de contingencia de 2 ó 3 dimensiones. Deville y Särndal (1992) propusieron una clase de medidas de distancia y posteriormente Deville et al. (1993), las aplican en el contexto en el que la información auxiliar corresponda a los totales marginales de tablas de contingencia. Estevao et al. (1995) desarrollaron un sistema computacional, en la Oficina de Estadística de Canadá, para la estimación de totales, razones de totales, promedios y proporciones, utilizando el estimador de regresión generalizado. Finalmente, Huang y Fuller (1978), y Singh y Mohl (1996) han desarrollado estimadores similares, los cuales mantienen las propiedades mencionadas anteriormente, pero donde el método numérico de minimización es diferente.

2 Medidas de distancia

La medida de distancia es hasta cierto punto arbitraria, motivo por el cual se han propuesto varias alternativas, entre las que destacan estan las sugeridas por Deville y Särndal (1992), y por Singh y Mohl (1996). En la tabla 1 se presentan cada una de ellas, así como la expresión para los ponderadores calibrados.

Es posible que el sistema de ecuaciones (1) sea no lineal, por lo que su solución requiere de procedimientos iterativos. La distancia más utilizada corresponde a la de mínimos cuadra-

dos generalizados. No obstante, Deville, et al. (1993), aplicaron la de mínimos cuadrados generalizados, *raking ratio*, mínimos cuadrados generalizados restringida, y *logit* a la encuesta

Tabla 1. Funciones de distancia y pesos calibrados

Dis.	$F^*(w_k, d_k)$	Peso calibrado (w_k)
1	$\frac{(w_k-d_k)^2}{2d_k q_k}$	$d_k(q_k \mathbf{x}_k'\lambda + 1)$
2	$\frac{1}{q_k}\{w_k \log(\frac{w_k}{d_k}) - w_k + d_k\}$	$d_k \exp(q_k \mathbf{x}_k'\lambda)$
3	$\frac{2(\sqrt{w_k}-\sqrt{d_k})^2}{q_k}$	$d_k(1 - \frac{q_k}{2}\mathbf{x}_k'\lambda)^{-2}$
4	$\frac{1}{q_k}\{-d_k \log(\frac{w_k}{d_k}) + w_k - d_k\}$	$d_k(1 - q_k \mathbf{x}_k'\lambda)^{-1}$
5	$\frac{(w_k-d_k)^2}{2w_k q_k}$	$d_k(1 - 2q_k \mathbf{x}_k'\lambda)^{-1/2}$
6	$\begin{cases} \frac{(w_k-d_k)^2}{2d_k q_k} & L < \frac{w_k}{d_k} < U \\ \infty & \text{en otro caso} \end{cases}$	$\begin{cases} d_k(q_k \mathbf{x}_k'\lambda + 1) & \frac{L-1}{q_k} < \mathbf{x}_k'\lambda < \frac{L-1}{q_k} \\ d_k L & \mathbf{x}_k'\lambda \le \frac{L-1}{q_k} \\ d_k U & \mathbf{x}_k'\lambda \ge \frac{U-1}{q_k} \end{cases}$
7	$\frac{d_k}{Aq_k}\left[(\frac{w_k}{d_k} - L)\log\left(\frac{\frac{w_k}{d_k}-L}{1-L}\right) + (U - \frac{w_k}{d_k})\log\left(\frac{U-\frac{w_k}{d_k}}{U-1}\right)\right]$	$\frac{d_k(U-1)L+(1-L)U\exp(Aq_k\mathbf{x}_k'\lambda)}{(U-1)+(1-L)\exp(Aq_k\mathbf{x}_k'\lambda)}$
8	$\frac{(w_k^{(v)}-d_k)^2}{d_k a_k^{(v-1)*}}$	$d_k(1 + a_j^{(v-1)*}\mathbf{x}_k'\lambda^{(v)})$
9	$\frac{(w_k^{(v)}-w_k^{(v-1)})^2}{w_k^{(v-1)}}$	$w_k^{(v-1)}(1 + \mathbf{x}_k'\lambda^{(v)})$

1: Mínimos cuadrados generalizados. 2: Raking ratio. 3: Hellinger. 4:Entropía mínima. 5:Mínimos cuadrados generalizados modificada. 6:Mínimos cuadrados generalizados restringida. 7:Logit, donde $A = \frac{U-L}{(1-L)(U-1)}$. 8:Huang-Fuller modificada, donde

$a_k^{(v-1)*} = a_k^{(v-1)} \ldots a_k^{(1)} a_k^{(0)}$, con $a_k^{(0)} = 1; a_k^{(v-1)} = 1$ si $\xi^{(v-1)} < .5$,

$a_k^{(v-1)} = 1 - \beta(\xi_k^{(v-1)} - .5)^2$ si $.5 \le \xi_k^{(v-1)} < 1, a_k^{(v-1)} = \frac{1-\beta/4}{\xi_k^{(v-1)}}$ si $\xi_k^{(v-1)} \ge 1$;

$\xi_k^{(v-1)} = \frac{g_k^{(v-1)}-1}{L'-1}$ si $g_k^{(v-1)} \le 1, \xi_k^{(v-1)} = \frac{g_k^{(v-1)}-1}{U'-1}$ en otro caso;

$L' = \alpha L + 1 - \alpha, U' = \alpha U + 1 - \alpha; 0 < \alpha, \beta < 1$ arbitrarios.

9:Contracción-minimización. Las primeras siete distancias fueron propuestas por Deville y Särndal (1992) y las dos últimas por Singh y Mohl (1996).

de condiciones de vida en Francia, 1990. Singh y Mohl (1996), adicionalmente a las funciones anteriores, utilizaron la distancia modificada de Huang-Fuller y la de contracción-minimización, a los datos de la Encuesta de Gasto Familiar en Canada, 1990. Stukel et al. (1996) también compararon estas seis medidas, simulando datos a partir de la Encuesta de Fuerza Laboral en Canada, 1990.

A continuación se ilustrará esta técnica bajo las condiciones siguientes: i) La información auxiliar corresponde a los totales marginales de una tabla de contingencia de 2 dimensiones. ii) Los ponderadores a calibrar corresponden a los de la Encuesta Nacional de Ingreso Gasto en los Hogares (ENIGH92). iii) La fuente de información auxiliar es el XI Censo General de Población y Vivienda, 1990. iv) La calibración se hará a nivel del hogar y no de individuos, utilizando variables censales del hogar. v) Para encontrar la solución al sistema de ecuaciones se elaboró un programa en MATLAB.

3 Fuentes de información

Las fuentes de información que se utilizaron fueron la ENIGH92 y el Censo de Población y Vivienda, 1990. La ENIGH92 se lleva a cabo por el INEGI y tiene como objetivo proporcionar información sobre:

a) La distribución, monto y estructura del ingreso y el gasto de los hogares mexicanos.

b) Las características sociodemográficas de los miembros del hogar y la condición de actividad y características ocupacionales de los miembros de 12 años y más.

c) Las características de infraestructura de la vivienda y de equipamiento del hogar.

La encuesta fue diseñada para generar resultados a nivel nacional así como para los estratos urbano (zonas de alta densidad) y rural (zonas de baja densidad). El diseño de muestreo fue estratificado, polietápico y con probabilidad proporcional al tamaño. Los estratos se conformaron de acuerdo a su densidad de población en urbano y rural.

El número efectivo de viviendas en muestra fue de 6,335 para el área urbana, 4,195 para el área rural sumando 10,530 a nivel nacional y proporciona información que puede clasificarse en los rubros siguientes:

Principales rubros que integran la ENIGH 1992

Rubro	**Nivel**	**Ejemplos**
Caract. sociodemográficas	Personas	Edad, sexo, perfil educativo, caract. ocupacionales
Infraestructura y servicios	Vivienda	Disponibilidad de agua, drenaje, material en muros
Ingreso gasto	Hogar	Salarios, rentas, intereses consumo comida, educación

Por otra parte, el XI Censo General de Población y Vivienda, 1990 genera información correspondiente a características sociodemográficas, y de las viviendas (materiales predominantes en la vivienda, disponibilidad de servicios, número de ocupantes, etc.) así como otras variables.

4 Resultados

Con el objetivo de ilustrar la técnica para generar estimaciones mediante pesos calibrados se utilizaron las variables siguientes.

i) Disponibilidad de drenaje y disponibilidad de luz eléctrica.

ii) Disponibilidad de agua entubada y material predominante en pisos.

Es importante mencionar que en el caso de tablas de contingencia los factores-g tienen la forma

$$g_k = g(\mathbf{x}_k'\lambda) = g(u_i + v_j),$$

donde i corresponde a la categoría i de una variable, y j corresponde a la categoría j de la otra variable. Los factores de ajuste obtenidos para cada celda se muestran en las tablas 2 y 3.

Tabla 2. Factores de ajuste utilizando disponibilidad de drenaje y de luz eléctrica

Categoría	Mín. Cuad. Gen.	Raking Ratio	Hellin.	Entro.	Logit	Mín. Cuad. Rest.	H-F	Cont. min.
Calle, luz	0.9642	0.9643	0.9644	0.9645	0.9638	0.9638	0.9638	0.9638
Calle, no luz	1.6767	1.6146	1.5581	1.4861	1.8834	1.9000	1.8926	1.8870
Fosa, luz	0.6342	0.6468	0.6527	0.6573	0.6121	0.6164	0.6078	0.6085
Fosa, no luz	1.3468	1.0831	0.9601	0.8640	1.8104	1.7193	1.8995	1.8866
Suelo, luz	0.4480	0.4838	0.4960	0.5036	0.5009	0.5000	0.5134	0.5038
Suelo, no luz	1.1606	0.8101	0.6913	0.6166	0.6427	0.6520	0.5205	0.6151
No dren., luz	1.1862	1.1695	1.1627	1.1577	1.1883	1.1859	1.1880	1.1899
No dren., no luz	1.8988	1.9582	1.9825	2.0004	1.8914	1.9000	1.8925	1.8857

Tabla 3. Factores de ajuste utilizando disponibilidad de agua entubada y material predominante en pisos

Categoría	Mín. Cuad. Gen.	Raking Ratio	Hellin.	Entropía	Mín. Cuad. Mod.	Logit Logit	Mín. Cuad. Rest.	H-F
Dentro vivienda, tierra	0.9228	0.9093	0.9035	0.8983	0.8897	0.8738	0.8886	0.8551
Dentro vivienda, cemento	0.8496	0.8514	0.8524	0.8533	0.8553	0.8564	0.8550	0.8579
Dentro vivienda, otro piso	0.8913	0.8904	0.8898	0.8892	0.8877	0.8879	0.8881	0.8878
Fuera vivienda, tierra	1.2449	1.2489	1.2508	1.2525	1.2558	1.2300	1.2350	1.2271
Fuera vivienda, cemento	1.1717	1.1694	1.1682	1.1669	1.1641	1.1741	1.1722	1.1750
Fuera vivienda, otro piso	1.2134	1.2229	1.2285	1.2349	1.2500	1.2322	1.2350	1.2315
Otro agua, tierra	1.2022	1.2035	1.2039	1.2043	1.2047	1.2212	1.2152	1.2271
Otro agua, cemento	1.1290	1.1268	1.1258	1.1249	1.1231	1.1025	1.1103	1.0958
Otro agua, otro piso	1.1707	1.1784	1.1829	1.1880	1.1996	1.2273	1.2148	1.2315

Como puede observarse, las diferentes medidas se comportan de manera parecida. Sin embargo, en el primer caso (tabla 2) la distancia de entropía toma los valores más extremos. Además, al explorar algunas medidas de la distribución de los factores de ajuste, se aprecia que la medida de entropía dio lugar a factores mayores en comparación con las demás.

En el segundo ejercicio (tabla 3), las diferencias entre los factores originados por las diversas medidas son pequeñas, observándose que la distancia modificada de Huang-Fuller es la que toma valores más extremos.

Tabla 4. Porcentajes estimados utilizando el ponderador original

Disponibilidad de luz eléctrica

Disp. de drenaje	Sí	No	No especif.	Total
Conect. al de la calle	54.1	0.1		54.2
Fosa séptica	12.6	0.6		13.2
Desagüe al suelo, río o lago	4.8	0.5		5.3
No dispone	20.2	5.7		25.8
No especificado			1.5	1.5
Total	91.6	6.9	1.5	100.0

Es importante notar que en las categorías, celdas en la tabla de contingencia, con menor número de casos es donde las diferencias son mayores. Por otra parte, con la finalidad de ver

cómo se modifican las estimaciones para las celdas de la tabla de contingencia, se estimaron las proporciones en cada una de estas celdas utilizando los ponderadores ajustados. En las tablas 4 y 5 se puede observar que existen diferencias entre las estimaciones sin usar los pesos calibrados y las que los usan. Dado que las estimaciones para las diversas medidas fueron muy parecidas, sólo se presentan las originadas por la función de mínimos cuadrados.

Tabla 5. Porcentajes estimados utilizando el ponderador calibrado usando la distancia de mínimos cuadrados generalizados

Disponibilidad de luz eléctrica

Disponibilidad de drenaje	Sí	No	No especif.	Total
Conectado al de la calle	52.2	0.2		52.3
Fosa séptica	8.0	0.8		8.8
Desage al suelo, río o lago	2.2	0.6		2.7
No dispone	23.9	10.8		34.7
No especificado			1.5	1.5
Total	86.2	12.3	1.5	100.0

5 Comentarios y trabajo por realizar

Pero, ¿cuál medida usar? La respuesta no es única, cada medida tiene ventajas y desventajas ya sea en términos computacionales, de existencia de una solución y distribución de los ponderadores calibrados (rango, variación, etc.). Desde el punto de vista computacional la distancias de Huang-Fuller y de Contracción-Minimización requieren de mayor trabajo, y el algoritmo más sencillo corresponde a la de mínimos cuadrados generalizados. Por otra parte, las distancias de mínimos cuadrados generalizados y *raking ratio* garantizan la existencia de una solución al sistema de ecuaciones. Por ejemplo, en el primer ejercicio no se encontró solución para la distancia de mínimos cuadrados generalizados modificada, y en el segundo para la de contracción-minimización. No obstante, falta trabajo por realizar con la finalidad de obtener observaciones y conclusiones más certeras. Se pretende continuar con el trabajo siguiente.

i) Utilizar la información de más de dos variables auxiliares, del censo, para calibrar.

ii) Explorar otras variables de la ENIGH para ser calibradas.

iii) Comparar las diversas medidas en términos de la varianza del estimador.

iv) Explorar el caso en que las variables auxiliares correspondan a características de los individuos y no de los hogares.

Referencias

Deming, W. E. y Stephan, F. F. (1940) On a Least Squares Adjustment of a Sampled Frequency Table when Expected Marginal Totals are Known, *The Annals of Mathematical Statistics*, **11**, 427-444.

Deville J. C., Särndal C. E., Sautory O. (1993) Generalized Raking Procedures in Survey Sampling, *Journal of the American Statistical Association*, **88**, 1013-1020.

Deville J. C., Särndal C. E., (1992) Calibration Estimators in Survey Sampling, *Journal of the American Statistical Association*, **87**, 376-382.

Estevao, V., Hidirolou, M.A. y Särndal, C.E. (1995) Methodological Principles for a Generalized Estimation System at Statistics Canada. *Journal of Official Statistics*, **11**, 181-204.

Huang , E.T., Fuller, W.A. (1978) Nonnegative regression estimation for sample survey data. *Proceedings of the Social Statistics Section, American Statistical Association*, 300-305.

Singh A. C., Mohl C.A. (1996) Understanding Calibration Estimators in Survey Sampling, *Survey Methodology*, bf 22, 107-115.

Stukel D.M., Hidiroglou M. A., Särndal C.E. (1996) Variance Estimation for Calibration Estimators: a Comparison of Jackknifing versus Taylor Linearization, *Survey Methodology*, **22**, 117-125.

Assessing and Modeling Rater Agreement

Alexander von Eye[1]
Michigan State University
Christof Schuster
University of Notre Dame

1 Abstract

Routinely, rater agreement is assessed using Cohen's (1960) κ which indicates the degree to which two raters agree beyond chance agreement. Although this measure is useful, it does not allow one to explain the structure of the agreement between two or more raters. This article presents a general log-linear model that allows researchers to capture several facets of the structure of the cross-tabulation of two or more raters' judgements. In addition, this general model allows one to test specific hypotheses concerning the structure of the cross-tabulation. Data examples are taken from a study in language research where raters evaluate the concreteness/abstractness of respondents' interpretations of proverbs.

2 Introduction

This article is concerned with the assessment and modeling of rater agreement. In the standard case, two raters rate n objects using I categorical or metric categories. These ratings can be cross-classified. The resulting table can then be subjected to a statistical analysis of the agreement between the two raters. In this article we propose a general model for the analysis of cross-classifications of two or more raters' judgements. Existing models are shown to be special cases of this general model. New models are derived. Section 2 of this article presents Cohen's (1960) κ. Section 3 presents a general log-linear model. Section 4 applies this model to a data example.

[1]Address correspondence concerning this article to Alexander von Eye, Michigan State University, Department of Psychology, 119 Snyder Hall, East Lansing, Michigan 48824-1117. The authors are indebted to Neal Schmitt for helpful comments on earlier versions of this article. Alexander von Eye's work on this article was supported in part by NIAAA grant # 2RO1 AA07065.

3 Cohen's (1960) coefficient κ for rater agreement

Consider the $I \times I$ cross-classification of two raters where each rater used the same I rating categories. A maximum-likelihood estimator of the agreement between these two raters is

$$\kappa = \frac{n \sum_i f_{ii} - \sum_i f_{i\cdot} f_{\cdot i}}{n^2 - \sum_i f_{i\cdot} f_{\cdot i}}, \tag{1}$$

(Cohen, 1960), where f_{ii} is the observed frequency in Cell (i, i), $f_{i\cdot}$ and $f_{\cdot j}$ are marginal frequencies and n is the number of objects that the two raters processed. The coefficient κ has been extensively discussed and significance tests have been proposed. κ indicates the degree to which the two raters agree beyond chance. Typically, the chance model is that of independence between the two raters. Other models have been discussed (von Eye & Soerensen, 1991). The characteristics of κ are well known and include

1. $\kappa \leq 1$; the smallest possible value of κ is $1 - n(n - \sum_i f_{ii})^{-1}$;

2. $\kappa = 0$ only if the number of disagreement cases in the off-diagonal cells is the same as the number of agreement cases in the diagonal cells; κ can be zero even if the raters' judgements are not independent;

3. $\kappa = 1$ only if the number of cases in the disagreement cells is zero;

4. κ is defined only if at least two categories are used by both raters, that is, if the probability, p_{ii}, is greater than zero for at least two cells.

5. If the number of cases in the off-diagonals is non-zero, the maximum value of κ decreases if the marginals are not uniformly distributed (see also the notion of *prevalence dependency* of chance-corrected agreement; Guggenmoos-Holzmann, 1995).

6. When the number of disagreements decreases and is smaller than the number of agreements, κ increases monotonically; when the number of disagreements increases and is greater than the number of agreements, κ does not decrease monotonically (von Eye, & Soerensen, 1991).

7. Multiplied by 100, κ indicates the percentage by which two raters' agreement exceeds the agreement that could be expected from chance.

4 A general log-linear model for rater agreement

The general model for rater agreement that we present in this article contains four components. The first component is the *base model*, that is, a model against which rater agreement is assessed. As for Cohen's κ we use the model of rater independence. The second component reflects the association present. It is useful if ratings are on an ordinal scale. The third component focuses on the cells in the main diagonal of the $I \times I$ cross-classification. These are the cells where two raters agree in their judgements. The fourth component allows researchers to consider covariates or special hypotheses. The general model is

$$\log f_{ij} = \lambda_0 + \lambda_i^A + \lambda_j^B + \beta u_i u_j + \delta(i,j) + \lambda^C x_{ij}, \tag{2}$$

where λ_0 is the constant, and λ_i^A and λ_j^B are the main effects for the two raters. These three terms constitute the base model. The next term constitutes the association component of the model. The $u_1 < \ldots < u_I$ are fixed or known scores of the rating categories, typically their ranks (natural numbers: $1, 2, 3, \ldots$), and β is a weight parameter. The next term focuses on the cells in the main diagonal. We set

$$\delta(i,j) = \begin{cases} \delta_i & \text{if } i = j \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

The δ_i-parameters can be constrained, for example, $\delta_1 = \delta_2 = \ldots = \delta_I$. The last term contains the covariates and their weights. λ^C is the parameter for the covariate x_{ij}. In the following section we present applications of the model given in (2). Each application uses a particular set of the terms beyond the base model. All rater agreement models share the δ terms in common.

5 Applications of the model for rater agreement

To illustrate the application of the model for rater agreement given in (2) we use a data set that was collected in a project on proverb interpretation (von Eye et al. 1990). $n = 129$ young adults provided written interpretations of a proverb. Two raters evaluated the concreteness/abstractness of these interpretations. The data in Table 1 describe the ratings that the two raters provided. The sentences and proverbs were rated as 1 = concrete, 2 = between concrete and abstract, and 3 = abstract. In the following analysis we ask

	Rater 1		
Rater 2	1	2	3
1	11	2	19
2	1	3	3
3	0	8	82

Table 2: Two Raters' Perception of Proverb Concreteness

whether (1) the two raters agree beyond chance, and (2) whether the agreement is statistically significantly greater than could be expected from chance.

For the data in Table 1 we calculate a Likelihood Ratio $X^2 = 39.03$ ($\mathrm{df} = 4; p < 0.01$) and $\kappa = 0.375$ ($\mathrm{se}_\kappa = 0.079; p < 0.01$). We thus conclude that (1) the assumption of independence between these two raters' perceptions can be rejected; (2) the agreement between the two raters is significant; and (3) the raters' agreement is 37.5% greater than was expected from chance. In the following sections we employ the model given in (2).

For the first application we set $\beta = 0$ and $\lambda = 0$, and $\delta = 1$. The resulting model is identical to the equal weight agreement model proposed by (Tanner & Young, 1985). Applied to the data in Table 1, this model, although significantly better than the independence model, provides relatively poor fit. Specifically, we calculate a likelihood $X^2 = 9.22$ which, for $\mathrm{df} = 3$ suggests that the discrepancy between expected and observed cell frequencies is significant ($p = 0.027$). Application of the same model yet with different weights for each rating category, that is, application of a weight-by-response-category agreement model yields a likelihood ratio $X^2 = 5.72$ ($\mathrm{df} = 1; p = 0.017$). This value indicates a significant improvement over the independence model ($\Delta X^2 = 33.31; \mathrm{df} = 3; p < 0.01$) thus suggesting that the differential weight agreement model allows one to explain the frequency distribution in the cross- tabulation better than the base model. However, as the equal-weight agreement model, the differential weight agreement model does not stand for itself and can thus not be retained. We thus conclude that focusing on the cells in the main diagonal does not allow one to capture the structure of the agreement between these two raters.

In the second application we set $\beta = 0$, $\delta = 1$, and employ one covariate thus specifying an equal-weight agreement model with a covariate. To illustrate this model we include the variable Wordiness (Verbosity) as a covariate in the analysis of rater agreement concerning

the concreteness/abstractness of proverb interpretation. The respondents' responses were also evaluated on a wordiness scale with 1 = wordy, 2 = intermediate, and 3 = not wordy. The rater agreement/disagreement frequencies in the nine cells of the 3 x 3 cross-classification of the two raters are $x_w = (17, 27, 3, 16, 45, 14, 1, 3, 3)'$.

The model with this covariate provided excellent fit (LR-$X^2 = 1.85; \mathrm{df} = 2; p = 0.40$) and is significantly better than both the main effect base model and the equal- weight agreement model without covariate. The improved fit is thus due solely to the covariate. The estimate of the λ^C for the covariate is -0.16 ($\mathrm{se}_{\lambda^C} = 0.07; z = -2.20; p = 0.036$). The estimate of the δ for the equal weight agreement coding variable is 3.65 ($\mathrm{se}_\delta = 1.13, z = 3.23; p = 0.022$). We thus conclude that the data in Table 1 can be explained from an equal-weight agreement hypothesis and knowledge of the raters' wordiness ratings. Because of the already very small X^2 can a differential weight solution not improve this solution.

For the third application we set $\lambda^C = 0$ but we consider the ordinal nature of the rating categories. For reasons of comparison, we first estimate a model without the term for the diagonal agreement, that is, a uniform association model. Second, we estimate a model that does include the rater agreement-specific term. The uniform association model yields the estimates LR-$X^2 = 13.17$ ($\mathrm{df} = 3; p = 0.004$). Thus, while significantly better than the null model, this model fails to sufficiently describe the two raters' proverb interpretations. Including the term for the main diagonal improves the model significantly (LR-$X^2 = 8.90; \mathrm{df} = 2; p = 0.012; \Delta X^2 = 4.2693; \Delta\mathrm{df} = 1; p = 0.039$). However, this improvement is still not strong enough for a sufficient data description. Knowledge of the ordinal nature of the rating categories of concreteness/abstractness does not improve the data description in this example.

6 Discussion

The general model for rater agreement can be extended to accommodate three or more raters, more than one rating object, or both. In addition, any combination of the three model terms in Equation 2 can be used, if enough degrees of freedom are available. Even rater-specific trends are conceivable. Trend- and other specific hypotheses can be cast in terms of covariates. Furthermore, one can combine agreement models with trend models to determine whether disagreement is random. Latent class agreement models have been

proposed as well as models where agreement is stratified, that is, modeled separately for the categories or strata of some external variable. All these and the present models allow one to describe the structure of rater agreement in far more detailed ways than single coefficients such as Cohen's κ

References

Cohen, J. (1960) A coefficient of agreement for nominal scales, *Educational and Psychological Measurement*, **20**, 37-46.

Guggenmoos-Holzman, I. (1995) Modelling covariate effects in observer agreement studies: the case of nominal scale agreement (letter to the editor), *Statistics in Medicine*, **14**, 2285-2286.

Tanner, M. A. and Young, M. A. (1985) Modeling agreement among raters, *Journal of the American Statistical Association*, **80**, 175-180.

von Eye, A., Jacobson, L. P. and Wills, S. D. (1990) Proverbs: Imagery, Interpretation, and Memory, *Paper presented at the 12th West Virginia University Conference on Life-Span Developmental Psychology.*

von Eye, A. and Soerensen, S. (1991) Models of chance when measuring interrater agreement with kappa, *Biometrical Journal*, **33**, 871-887.

Esta publicación consta de 410 ejemplares y se terminó de
imprimir en el mes de agosto del 2000 en los talleres gráficos del
Instituto Nacional de Estadística, Geografía e Informática
Av. Héroe de Nacozari Núm. 2301 Sur, Acceso 11, PB
Fracc. Jardines del Parque, CP 20270
Aguascalientes, Ags.
México