HOMENAJE AL PROFESOR SIXTO RIOS
TRABAJOS DE ESTADISTICA Y DE INVESTIGACION OPERATIVA
Vol. 36, Núm. 3, 1985, pp. 259 a 268

ESTIMACION DE REGISTROS DESCONOCIDOS EN SERIES DE DATOS

Miguel Sanchez García — *Dpto. de Estadística e I.O. Facultad de Matemáticas. Universidad de La Laguna*

Luis Javier López Martin — *Dpto. de Estadística y Econometría. Facultad de C. Econòmicas. Universidad de La Laguna*

Se dispone de dos ó más series de datos, de las cuales, al menos una, no se conoce completamente. Se supone que las series se pueden modelizar con la hipótesis lineal; así como, que existe alguna estructura de correlación entre ellas. Se desarrollan dos modelos para estimar los valores desconocidos de la(s) serie(s) de datos.

Palabras Clave: Modelo Lineal; Función Objetivo; Mínimos Cuadrados; Optimación.

Clasificación AMS (1980): Primaria, 62J99; Secundaria, 62J15.

Estimation missing values in data series

The paper deals with the problem of (at least) one of the data series, given two or more them, is not completely known. We asume that the series can be modelled according to the linear hypothesis, and that there some correlation among them. In this paper we develop two models for estimation of the unknown values of that data series.

Key words: Lineal Model; Objective function; Least Square; Optimitation.

AMS Classification (1980): Primary, 62J99; Secondary, 62J15.

1. INTRODUCCION

La capacidad ingente de almacenamiento de datos por los computadores, unido al hecho de que los hombres quieran disponer de más y mejor información, lleva consigo la tarea de análisis de los datos disponibles con el fin de que sea más comprensible su información. A veces sucede que no tenemos la información completa sobre un determinado fenómeno observado a lo largo del tiempo; pues, ó bien algunos registros no se han podido observar, ó bien se observaron pero se

extraviaron con posterioridad, etc. El poder estimar los datos que faltan, para completar las series, es uno de los problemas actuales más interesantes. Estas estimaciones se suelen realizar bajo la hipótesis de que los datos observados siguen un determinado modelo. Entre los más usuales se encuentran el modelo lineal y los modelos de series temporales AR, MA, ARMA, etc.

Tradicionalmente, la estimación de los valores que no se conocen se realizaba estimando los parámetros del modelo al que suponemos se ajustan los datos; calculando con este modelo, posteriormente, los datos desconocidos; ó bien, en el caso de que los datos sean series temporales, estimando la función de densidad espectral del proceso correspondiente. En esta línea podemos citar los trabajos de Jones (1962,1971); Parzen (1963); Dunsmuir & Robinson (1981) y Bloomfield (1970). En el dominio del tiempo, citaremos a Andel (1976).

Recientemente Brown & Kadiyala (1983), han estimado los valores desconocidos de una serie de datos fundamentándose, tanto en el modelo que siguen estos datos, como en la información, desde el punto de vista de la correlación, que sobre estos contiene otra serie totalmenete conocida; probando, sobre varias series, que se obtienen mejores resultados que con el método tradicional.

En el presente artículo exponemos dos métodos para estimar los valores desconocidos, siendo el segundo una generalización del de Brown-Kadiyala.

2. TERMINOLOGIA

Representamos por:

Y: afectada ó no de subíndice, una serie de datos tal que tiene algún registro desconocido.

X: denotará la matriz de datos del modelo lineal adecuado.

Z: representa series de datos totalmente conocidas.

E: salvo que se diga lo contrario, es un vector columna con todos sus elementos iguales a 1.

σ: representa un peso para construir la función objetivo, cuyo valor teórico es un coeficiente de correlación.

β ,γ: son parámetros de modelos lineales.

3. MODELO I

En este modelo, para estimar los registros desconocidos de la serie de datos, representados por la variable Y, tendremos en cuenta, tanto la información suministrada por la hipótesis de que Y se adapta al modelo lineal $Y= X\beta + \varepsilon$ como la plausible dependencia entre Y y las series de datos, que supondremos conocidas, $Z_1,Z_2,...,Z_p$.

Es obvio que, si Y es independiente de $Z_1,Z_2,...,Z_p$, entonces la información válida disponible es la suministrada por el modelo lineal, y como consecuencia, los valores desconocidos de Y se deben hallar por el método tradicional. Por el contrario, si Y fuera linealmente dependiente de $Z_1,Z_2,...,Z_p$; entonces esta información sería idónea para estimar los valores desconocidos de Y. Estas ideas nos conducen a considerar una función objetivo, que tenga en cuenta las facetas previamente expuestas.

Si llamamos Y_1 a los valores conocidos de Y, e Y_2 a los desconocidos, la función objetivo, a la que anteriormente hemos hecho referencia, admite la siguiente expresión:

$$\text{Min}\{ (1 - \sigma) [| Y_1 - X_1\beta |^2 + | Y_2 - X_2\beta |^2] +$$

$$+ \sigma [| Y_1 E_1\alpha - \Sigma_i Z_i^1 \gamma_i |^2 + | Y_2 E_2\alpha - \Sigma_i Z_i^2 \gamma_i |^2] \}$$

Un valor plausible para σ, sería el coeficiente de correlación entre Y y $\alpha + \Sigma_i \gamma_i Z_i$, que en la práctica suele ser desconocido, aunque se puede estimar con la subserie relativa a Y_1. Consideramos, más conveniente, estimar distintos valores para Y_2, en función de los valores que se obtienen al variar σ entre cero y el coeficiente de correlación estimado con la subserie Y_1; y tomar la decisión, a posteriori, sobre qué σ es el más interesante. Una decisión válida sería tomar como valor de σ, aquel que minimice la función objetivo previa.

Derivando la función objetivo, respecto de cada una de las incógnitas, obtenemos:

$$X_1^T(Y_1 - X_1\beta) + X_2^T(Y_2 - X_2\beta) = 0$$

$$E_1^T(Y_1 - E_1\alpha - Z^1\gamma) + E_2^T(Y_2 - E_2\alpha - Z^2\gamma) = 0$$

$$Z^{1T}(Y_1 - E_1\alpha - Z^1\gamma) + Z^{2T}(Y_2 - E_2\alpha - Z^2\gamma) = 0$$

$$(1 - \sigma)(Y_2 - X_2\beta) + \sigma(Y_2 - E_2\alpha - Z^2\gamma) = 0 \qquad (I)$$

Despejando β de la primera ecuación se obtiene

$$\beta = (X^T X)^{-1} X^T Y = (X^T X)^{-1} X^T (Y_1^T Y_2^T)^T$$

Sustituyendo β por su igual en la 4a. ecuación, y sacando factor común Y_2 tenemos:

$$Y_2 = (I_2 + (1 - \sigma)D_2)^{-1} ((1 - \sigma)D_1 Y_1 + \sigma(E_2\sigma + Z^2\gamma))$$

siendo

$$D_2 = X_2(X^T X)^{-1} X_2^T \quad y \quad D_1 = X_2(X^T X)^{-1} X_1^T$$

Llamando $H = (I_2 + (1 - \sigma)D_2)^{-1}$ y sustituyendo el valor de Y_2 en la ecuación 2, se obtiene para α el siguiente valor:

$$\alpha = (N - \sigma E_2^T H E_2)^{-1}((E_1^T + (1 - \sigma)E_2^T H D_1) Y_1 - (E_1^T Z^1 + E_2^T(Z^2 - \sigma H Z^2)\gamma) \leftrightarrow$$

$$\alpha = H_1 Y_1 - H_2\gamma; \text{ siendo } H_1 = (N - \sigma E_2^T H E_2)^{-1}(E_1^T + (1 - \sigma)E_2^T H D_1) \text{ y}$$

$$H_2 = (N - \sigma E_2^T H E_2)^{-1}(E_1^T Z^1 + E_2^T(Z^2 - \sigma H Z^2))$$

Sustituyendo los valore de Y_2 y α en la ecuación 3a. obtenemos:

$$Z^{1T}(Y_1 - E_1 H_1 + Y_1(E_1 H_2 - Z^2)\gamma) + Z^{2T}((1 - \sigma)H D_1 Y_1 + (\sigma E_2 - E_2)$$

$$(H_1 Y_1 - H_2\gamma) + (\sigma Z^2 - Z^2)\gamma) = 0$$

y despejando γ se obtiene:

$$\gamma = (Z^{1T}(E_1H^2 - Z^2) + Z^{2T}(- (\sigma E_2 - E_2)H + \sigma Z^2 - Z^2))^{-1} .$$

$$(Z^{1T}(Y_1 - E_1H_1Y_1) + Z^{2T}((1 - \sigma)HD_1Y_1 + (\sigma E_2 - E_2)H_1Y_1))$$

Aunque hemos dado una solución explicita para las ecuaciones (I), ésta es demasiado compleja, por lo que proponemos seguidamente una solución algorítmica.

Para ello transformamos el sistema (I) en los dos bloques siguientes:

$$Y_2 = (1- \sigma)X_2\beta + \sigma(E_2\alpha + Z^2\gamma) = (1- \sigma)X_2\beta + \sigma(E_2\mathbf{Y} + Z'^2Y) \qquad (II)$$

$$\beta = (X^T X)^{-1}X^T Y \quad \text{siendo } X= (X_1^T X_2^T)^T \quad e \quad Y= (Y_1^T Y_2^T)^T$$

$$\alpha = \mathbf{X} - \mathbf{Z}^T\gamma \quad \text{siendo } \mathbf{Y} = N^{-1}E^TY \quad y \quad \mathbf{Z}^T = N^{-1}E^TZ; \quad Z = (Z^{1T} Z^{2T})^T \qquad (III)$$

$$\gamma = (Z'^T Z')^{-1}(Z')^T Y' \quad \text{siendo } Y'= Y - E\mathbf{Y} \quad y \quad Z'= Z - E\mathbf{Z}^T$$

Si se conoce el valor de Y_2, es obvio que las ecuaciones (III) son evaluadas. Por tanto, sustituyendo en (II) los valores de β y γ tenemos:

$$Y_2 = (1 - \sigma) X_2(X^T X)^{-1}X^T Y + \sigma(E_2\mathbf{Y} + Z'^2(Z'^T Z')^{-1}(Z')^T Y') \qquad (IV)$$

Proponemos el el siguiente algorítmo iterativo para hallar Y_2. En él denotaremos por $B_1= X_2(X^T X)^{-1}X^T$ y $B_2= Z'^2(Z'^T Z')^{-1}(Z')^T$

3.1 ALGORITMO

Paso 0 Se eligen unos valores arbitrarios para Y_2, que llamaremos Y_2^0.

Paso 1 Se coloca I=1, evaluando $\mathbf{Y}^0$ e Y'^0.

Paso 2 Se calcula $Y_2^I= (1 - \sigma)B_1Y_2^{I-1} + \sigma(E_2\mathbf{Y}^{I-1} + B_2Y'^{I-1})$

Paso 3 Se evaluan $\mathbf{Y}^I= N^{-1}(E_1Y_1 + E_2Y_2^I)$ e $Y'^I=Y^I - \mathbf{Y}^I$

Paso 4 Si $| Y_2^{I-1} - Y_1^{I-1} | < \varepsilon$ se para; caso contrario, se hace I= I+1 y se vuelve al paso 2.

Si restamos de $Y^{I+1} - Y^I$, obtenemos:

$$(Y_2^{I+1} - Y_2^I) = (1- \sigma)B_1(Y^I - Y^{I-1}) + \sigma(E_2(\mathbf{Y}^I - \mathbf{Y}^{I-1}) + B_2(Y'^I - Y'^{I-1}))$$

3.2 CONVERGENCIA DEL ALGORITMO

Para demostrar la convergencia, transformamos la ecuación (IV) en:

$$Y_2 = ((1 - \sigma) X_2(X^T X)^{-1}X^T + \sigma Z'^2(Z'^T Z')^{-1}Z'^T)Y$$

$$+ \sigma(E_2 - Z'^2(Z'^T Z')^{-1}Z'^T E) \mathbf{Y}$$

Llamando:

$$F = (1 - \sigma) X_2(X^T X)^{-1}X^T + \sigma Z'^2(Z'^T Z')Z'^T$$

$$G = \sigma(E_2 - Z'^2(Z'^T Z')^{-1}Z'^T E)$$

La ecuación previa se nos transforma en;

$$Y_2 = FY + G\mathbf{Y} = F_2Y_2 + F_1Y_1 + G\mathbf{Y}$$

siendo

$$F_i = (1 - \sigma) X_2(X^T X)^{-1}X_i^T + \sigma Z'^2(Z'^T Z')^{-1}Z'^{iT} \; ; \; i=1,2$$

El algoritmo computa sucesivos valores aproximados de Y_2 mediante la fórmula siguiente:

$$Y_2^{n+1} = F_2Y_2^n + F_1Y_1 + G\mathbf{Y}_n = F_2^{n+1}Y_2^0 + \Sigma_k (F_2^kF_1Y_1 + F_2^kG\mathbf{Y}^{n-k})$$

Una condición suficiente para la convergencia de Y_2^n; es que la matriz F_2 tenga norma menor que 1; es decir $|F_2| < 1$.

Ahora bien, los operadores $X(X^T X)^{-1}X^T$ y $Z'(Z'^T Z')^{-1}Z'^T$ son proyectores, y tanto de norma 1.

$$|F_2| \leq |(1 - \sigma)X(X^T X)^{-1}X^T + \sigma Z'(Z'^T Z')^{-1}Z'^T|$$

$$\leq (1 - \sigma)|X(X^T X)^{-1}X^T| + \sigma|Z'(Z'^T Z')^{-1}Z'^T| = 1 - \sigma + \sigma = 1$$

Para que $|F_2| < 1$, es necesario y suficiente que una de las dos desigualdades pprevias sea estricta. La segunda de ellas es estricta, salvo que los dos operadores sean iguales. Sean, en otro caso, ambos operadores iguales, es decir:

$$X(X^T X)^{-1}X^T = Z'(Z'^T Z')^{-1}Z'^T ; \quad \text{y sea}$$

$$X(X^T X)^{-1}X^T = \begin{vmatrix} F_0 & F_1^T \\ F_1 & F_2 \end{vmatrix}$$

Es obvio que si $F_1 = 0$, entonces $|F_2| = \begin{vmatrix} F_0 & F_1^T \\ F_1 & F_2 \end{vmatrix} = 1.$

Además, también es evidente que, si en en todas las columnas de F_1 existe algún elemento no nulo, $|F_2| < 1$. Como $F_1F_1^T + F_2F_2 = F_2$; salvo que $F_1 = 0$, se tiene que $|F_2| < 1$.

4. MODELO II

Para este modelo disponenos de k+1 series de datos; cada una de las cuales se puede representar adecuadamente por un modelo lineal, en los siguientes términos: $Z_1 = X_1\beta_1 + \varepsilon_1 \quad Z_2 = X_2\beta_2 + \varepsilon_2 \ldots Z_k = X_k\beta_k + \varepsilon_k \quad Y_0 = X_0\beta_0 + \varepsilon_0$

Supondremos que disponemos de N obsevaciones; todas ellas conocidas para $Z_1, Z_2, \ldots, Z_k$; mientras que de Y_0 solo conocemos N_1 puntos, que, sin pérdida de generalidad, serán los N_1 primeros.

Admitiremos que la estructura de covarianza de los errores, tiene la forma siguiente:

$$E\begin{vmatrix}\varepsilon_1\\ \varepsilon_2\\ .\\ .\\ .\\ \varepsilon_k\\ \varepsilon_0\end{vmatrix}(\varepsilon_1\,\varepsilon_2\,\ldots\,\varepsilon_k\,\varepsilon_0) = \begin{vmatrix}\sigma_{11} & \sigma_{12} & \ldots & \sigma_{1k} & \sigma_{10}\\ \sigma_{21} & \sigma_{22} & \ldots & \sigma_{2k} & \sigma_{20}\\ . & . & & & \\ . & . & & & \\ . & . & & & \\ \sigma_{k1} & \sigma_{k2} & \ldots & \sigma_{kk} & \sigma_{k0}\\ \sigma_{01} & \sigma_{02} & \ldots & \sigma_{0k} & \sigma_{00}\end{vmatrix} \mathbf{x}\ I = S\ \mathbf{x}\ I$$

Colocaremos la serie Y_0, particionada en dos cajas:

$$\begin{vmatrix}Y_0^1\\ \\ Y_0^2\end{vmatrix} = \begin{vmatrix}X_0^1\\ \\ X_0^2\end{vmatrix}\beta_0 + \varepsilon_0 \quad \leftrightarrow \quad \begin{vmatrix}Y_0^1\\ \\ 0\end{vmatrix} = \begin{vmatrix}X_0^1 & 0\\ \\ X_0^2 & -I\end{vmatrix}\begin{vmatrix}\beta_0\\ \\ Y_0^2\end{vmatrix} + \varepsilon_0$$

Llamaremos

$$Y = \begin{vmatrix}Y_0^1\\ \\ 0\end{vmatrix} \qquad \gamma = \begin{vmatrix}\beta_0\\ \\ Y_0^2\end{vmatrix} \qquad \text{y} \qquad U = \begin{vmatrix}X_0^1 & 0\\ \\ X_0^2 & -I\end{vmatrix}$$

La última serie, con esta nueva terminología, se puede expresar en los siguientes términos: $Y = U\gamma + \varepsilon_0$

Definiremos una función objetivo adecuada para estimar los parámetros. Parece lógico pensar que esta función objetivo, debe depender de la estructura de covarianza entre las series. A este respecto supondremos que S es no singular.

Denotando entonces por

$$S^{-1} = \begin{vmatrix}\sigma^{11} & \sigma^{12} & \ldots & \sigma^{1k} & \sigma^{10}\\ \sigma^{21} & \sigma^{22} & \ldots & \sigma^{2k} & \sigma^{20}\\ . & . & \ldots & . & .\\ . & . & \ldots & . & .\\ . & . & \ldots & . & .\\ \sigma^{11} & \sigma^{12} & \ldots & \sigma^{1k} & \sigma^{10}\\ \sigma^{11} & \sigma^{12} & \ldots & \sigma^{1k} & \sigma^{10}\end{vmatrix} ;\ \sigma^{ij} = (\sigma_{ij})^{-1}$$

y siguiendo a Brown-Kadiyala, planteamos la siguiente función objetivo:

$$\text{Min}\{\ \sigma^{00}\,|\,Y - U\gamma\,|^2 + \ 2\,\Sigma_i\,\sigma^{0i}(Y - U\gamma)^T(Z_i - X_i\beta_i)$$

$$+\ \Sigma_i\,\Sigma_j\,\sigma^{ij}\,(Z_i - X_i\beta_i)^T(Z_j - X_j\beta_j)\ )\}\ ;\ i,j=1,...,k$$

Derivando respecto de $\gamma, \beta_1,\beta_2,...,\beta_k$; e igualando a cero las derivadas, obtenemos el siguiente sistema de ecuaciones:

$$\begin{vmatrix} \sigma^{01}\,X_1^TY + \Sigma_i\,\sigma^{ij}\,X_1^TY_j \\ \sigma^{02}\,X_2^TY + \Sigma_j\,\sigma^{2j}\,X_2^TY_j \\ \cdot \\ \cdot \\ \cdot \\ \sigma^{0k}\,X_k^TY + \Sigma_j\,\sigma^{kj}\,X_k^TZ_j \\ \sigma^{00}\,U^TY + \Sigma_j\,\sigma^{0j}\,U^TZ_j \end{vmatrix} = \begin{vmatrix} \sigma^{11}\,X_1^TX_1 & \sigma^{12}\,X_1^TX_2 & \sigma^{1k}\,X_1^TX_k & \sigma^{10}\,X_1^TU \\ \sigma^{21}\,X_2^TX_1 & \sigma^{22}\,X_2^TX_2 & \sigma^{2k}\,X_2^TX_k & \sigma^{20}\,X_2^TU \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ \sigma^{k1}\,X_k^TX_1 & \sigma^{k2}\,X_k^TX_2 & \sigma^{kk}\,X_k^TX_k & \sigma^{k0}\,X_k^TU \\ \sigma^{01}\,U^TX_1 & \sigma^{02}\,U^TX_2 & \sigma^{0k}\,U^TX_k & \sigma^{00}\,U^TU \end{vmatrix} \begin{vmatrix} \beta_1 \\ \beta_2 \\ \cdot \\ \cdot \\ \cdot \\ \beta_k \\ \gamma \end{vmatrix}$$

La resolución del anterior sistema de ecuaciones; da los estimadores de los parámetros del modelo.

El problema que se suele presentar es el del desconocimiento de S. En este caso es preciso estimarla por la matriz:

$$S^* = \begin{vmatrix} S_{11} & S_{12} & \cdots & S_{1k} & S_{10} \\ S_{21} & S_{22} & \cdots & S_{2k} & S_{20} \\ \cdot & \cdot & & \cdot & \cdot \\ \cdot & \cdot & & \cdot & \cdot \\ \cdot & \cdot & & \cdot & \cdot \\ S_{k1} & S_{k2} & \cdots & S_{kk} & S_{k0} \\ S_{01} & S_{02} & \cdots & S_{0k} & S_{00} \end{vmatrix}$$

siendo $S_{ij} = (\,S_{11}\,S_{jj}\,)^{1/2}\ \gamma_{ij}$ con $S_{ii} = \Sigma_j\,\varepsilon_{ij}^2 / N$; $j=1,...,n$, $i=1,2,...,k$ y $S_{00} = N^{-1}\,\Sigma_j\,\varepsilon_{0j}^2$; donde ε_{ij} es el error para la serie i-ésima, cuando se estiman los parámetros por el método de mínimos cuadrados ordinarios.

REFERENCIAS

ANDEL, J. (1976). *Multiple autoregressive Processes*. Charles University, Czechoslovakia.

BLOOMFIELD, P.(1970). Spectral Analysis with randomly missing observations. *J. R. Statist. Soc.* **B32**, 369-380.

BROWN, K.C. y KADIYALA, K.R. (1983). Construction of economic index numbers with an incomplete set of data. *The Review of Economics and Statistics*. **3**, 520-524.

DUNSMUIR, W. y ROBINSON, P.M. (1981). Parametric estimators for stationary time series with missing observations. *Adv. Appl. Prob.* **13**, 129-146.

JONES, R.H. (1962). Spectral Analysis with regularly missed observations. *Ann. Math. Statistic.* **33**, 455-461

JONES R.H. (1971). Spectrum Estimation with missing observations. *Ann . Inst. Statist. Math.* **23**, 387-398.

PARZEN E. (1963). On Spectral Analysis with missing observations and amplitude modulation. Sankhya, series A, **25**, 383-392.