Información faltante en bases de datos biológicos. Estimación máximo verosímil a través del operador "sweep"*

RECIBIDO: 20/5/06
ACEPTADO: 26/10/06

Badler, C.E. • Alsina, S.M. • Puigsubirá, C.R. • Vitelleschi, M.S.

Instituto de Investigaciones Teóricas y Aplicadas de la Escuela de Estadística. Facultad de Ciencias Económicas y Estadística. Universidad Nacional de Rosario. Bvrd. Oroño 1261. (2000) Rosario, Santa Fe, Argentina.
Teléfono:0341-4802793-int.151. E-mail: salsina@fcecon.unr.edu.ar

RESUMEN: Cuando una base de datos presenta pérdidas parciales en ciertas variables y se desea estimar el vector de promedios y la matriz de covariancias, algunos de los procedimientos habituales para el tratamiento de datos faltantes se caracterizan por su simplicidad pero descartan información de las unidades incompletas.
Una alternativa es la de calcular dichas estimaciones a través del método de máxima verosimilitud, el cual utiliza la información tanto de las observaciones completas como de aquellas que tienen datos faltantes en alguna de las variables.
En situaciones en que el mecanismo de pérdida de la información es ignorable, el esquema de pérdida es monótono y se cumple el supuesto de distribución normal multivariada, las estimaciones pueden ser obtenidas a través de la maximización de una secuencia de predicciones por regresión, que pueden ser realizadas mediante la aplicación del operador "Sweep".
En este trabajo se aplica la metodología propuesta para calcular las estimaciones máximo verosímiles del vector de promedios y la matriz de covariancias de variables correspondientes a una base de datos de niños y adolescentes diabéticos atendidos en establecimientos asistenciales públicos y privados de la ciudad de Rosario en el año 2003.
PALABRAS CLAVE: Información faltante, Estimación máximo verosímil, Operador "Sweep"

SUMMARY: *Missing information in biological data sets. Maximum likelihood estimation using the sweep operator.*
When a database presents missing information for some units of certain variables and the purpose is to estimate the mean and convariance matrix, some of the classic methods of treatment are easy to apply but do not use the information of the variables not completely observed.
If those estimations are calculated by the likelihood method, all the available information can be incorporated for the estimations.
When the missing data mechanism is ignorable, the missing pattern is monotone and the variables follow a multivariate normal distribution, the maximum likelihood estimates can be obtained as the maximization of a sequence of regression predictions, which can be performed by the Sweep operator.
In this work, the methodology is applied to find maximum likelihood estimates of the mean and covariance matrix of the variables of a data set of diabetic children and young adults that have been assisted in public and private institutes of Rosario city, during the year 2003.
KEY WORDS: Missing information, Maximum Likelihood Estimation, Sweep Operator.