

Comparación de subconjuntos de regresiones cuando algunas de las variables explicativas provenientes de bases de datos biológicos están observadas parcialmente.

RECIBIDO: 30/05/2008

ACEPTADO: 19/03/2009

Badler, C. E. • 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) 480 2793 int. 151.

Email: cpuigsu@fcecon.unr.edu.ar

RESUMEN: El coeficiente de correlación múltiple es usualmente utilizado para comparar conjuntos de variables explicativas con respecto a cuán bien ellas predicen los valores futuros de la variable respuesta. Si la base de datos contiene variables observadas parcialmente, Donald Rubin propuso una adaptación del coeficiente de correlación múltiple con el fin de no descartar información. Para el cálculo de dicho coeficiente se requieren las estimaciones máximo verosímiles de ciertos parámetros, las cuales se realizan a través del operador "Sweep". La metodología es aplicada a un conjunto de datos provenientes de las historias clínicas perinatales de niños nacidos en el Hospital Roque Sáenz Peña de la ciudad de Rosario durante el año 2002.

PALABRAS CLAVE: Información faltante, Selección de variables, Coeficiente de correlación múltiple, Operador "Sweep".

SUMMARY: *Comparison of regression subsets when some predictor variables of biological data are partially observed.*

The multiple correlation coefficient is often used to compare sets of independent variables regarding how well they predict the future values of a dependent variable. If the data set contains partially observed variables, Donald Rubin proposed an adaptation of the multiple correlation coefficient in order not to discard information. For the calculation of this coefficient the maximum likelihood estimates of certain parameters are required, which are carried out through the "Sweep" operator. The methodology is applied to a data set taken from perinatal clinical histories of children born in the Roque Sáenz Peña Hospital of Rosario city, during the year 2002.

KEYWORDS: Missing data, Selection of variables, Multiple correlation coefficient, "Sweep" operator.