

Jorgelina Carrizo

Universidad Nacional de San Juan, Argentina

jorgelinav_carrizo@yahoo.com.ar

El análisis de datos simbólicos es un campo relativamente nuevo de estadística, que amplía el análisis de datos clásico teniendo en cuenta la variabilidad y estructura intrínseca de los datos. La adaptación de la regresión clásica a datos simbólicos, especialmente a datos de intervalo, se ha transformado en un tema de creciente investigación desde que Billard y Diday (2000) presentaron el primer enfoque para ajustar un modelo de regresión lineal a conjuntos de datos simbólicos de intervalos desde una perspectiva del SDA. Este trabajo presenta el método de regresión lineal, denominado Método de Covarianza Simbólica, que permite reconstruir el estimador de mínimos cuadrados de un modelo de regresión lineal clásica, en base a las definiciones de varianza y covarianza muestral simbólica de Billard (2007, 2008). Estas definiciones tienen en cuenta tanto las variaciones internas como externas de los datos de intervalo, como así también la dependencia entre todas las variables de manera integral. Por lo tanto, a diferencia de otros métodos que ajustan dos modelos de regresión lineal, uno para los puntos medios del intervalo y el otro para los rangos del mismo, este nuevo enfoque construye un modelo directamente para los datos de intervalo. El método presentado se aplica a datos reales y se compara su desempeño y facilidad de interpretación, respecto al Método del Centro y al Método del Centro y Rango. Para esta evaluación se utilizan el error cuadrático medio y el coeficiente de correlación.

Palabras Claves: Análisis de datos simbólicos; Variables simbólicas de intervalo; Regresión simbólica. Covarianza simbólica.

Referencias

- [1] Billard and Diday (2000). Regression analysis for interval-valued data. In: Proc. of IFCS00, Belgium, pp. 369-374, Springer.
- [2] Billard, L. (2007). Dependencies and Variation Components of Symbolic Interval-Valued Data.
- [3] Billard, L. (2008). Sample Covariance Functions for Complex Quantitative Data. Processing, World Conferences International Association of Statistical Computing 2008, Yokohama, Japan.
- [4] Billard, L. and Diday, E. (2000). Regression Analysis for Interval-Valued Data. Data analysis, Classification, and Related Methods (eds. H.A.L. Kiers, J.-P.
- [5] Billard, L. and Diday, E. (2003). From the Statistics of Data to the Statistics of Knowledge: Symbolic Data Analysis.
- [6] Billard, L. and Diday, E. (2007). Symbolic Data Analysis: Conceptual Statistics and Data Mining. Wiley, Chichester.
- [7] Douzal-Chouakria, A., Billard, L. and Diday E. (2009). Principal Component Analysis for Interval-valued Observations. Submitted manuscript.
- [8] Lima Neto, E.A and de Carvalho F.A.T. (2010). Constrained Linear Regression Models for Symbolic Interval-valued Variables.
- [9] Lima Neto, E.A., de Carvalho F.A.T. and Freire, E.S. (2005). Applying Constrained Linear Aggression Models to Predict Interval-Valued Data.
- [10] Xu, W. (2010), Symbolic data analysis: regression of data with interval values.