Método de Regresión Lineal de Covarianza Simbólica para datos de intervalo

Carrizo Jorgelina

LXXI Reunión Anual de Comunicaciones Científicas. Neuquén.
Septiembre de 2022

Introducción

La regresión simbólica es parte del análisis simbólico de datos,se presentan adaptaciones del modelo de regresión lineal clásica para variables de intervalo,con los parámetros de regresión estimados por el conocido procedimiento de mínimos cuadrados usando,promedios, varianzas y covarianzas simbólicas. La metodología empleada forma parte de lo que se denomina Regresión simbólica de datos, área de investigación activa desde que Billard y Diday [2] introdujeron el primer enfoque del modelo de regresión a datos con valores de intervalo, trabajando con un modelo de regresión clásica sobre los centros de los intervalos. Posteriormente propusieron estimaciones a partir de modelos independientes para los límites inferior y superior de los intervalos. Lima Neto y de Carvalho [7], siguiendo el enfoque de Billard y Diday, emplearon los centros y rangos de los intervalos para construir dos regresiones independientes.

Introducción

Sin embargo, la predicción del límite inferior podría exceder el límite superior previsto. Para evitar esta situación, Lima Neto y de Carvalho [8] consideraron un modelo restringido que obliga a todos los parámetros del modelo de los rangos a ser positivos.
El pensamiento de reducir los intervalos hasta los puntos centrales y rangos, para establecer un modelo; puede no reflejar la variación interna de los datos de la mejor manera desde un punto de vista simbólico.
Interesados en encontrar una solución que pueda usar directamente datos de valor intervalo, Xu [10] y Billard [11], propusieron un método de covarianza simbólico (llamado método SCM), para realizar regresión de datos con valores de intervalo. La idea principal es reconstruir el estimador de mínimos cuadrados de un modelo de regresión de una manera que utiliza la covarianza muestral simbólica.
Si bien existen otros métodos de Regresión Simbólica en la literatura, se mencionan sólo estos por ser los que se aplican en el presente trabajo.

Introducción

Los métodos presentados en este trabajo se aplican a varios conjuntos de datos. Para una evaluación de los métodos se utilizan el error cuadrático medio y el coeficiente de correlación.

Variable Simbólica

Una variable simbólica es una función que asigna un valor simbólico a cada elemento del espacio muestral.

Variable simbólica de intervalo

Una variable simbólica de intervalo es aquella que para cada individuo w_u, $u = 1, \ldots, m$; toma como posibles valores uno o más valores de su dominio $\mathcal{Y}$, i.e., $Y(w_u) = \xi_u = [a_u, b_u] \subset \mathbb{R}$, donde $a_u \leq b_u$, y $a_u, b_u \in \mathbb{R}$. Una forma alternativa de representar a los intervalos es como una bola de centro $c_u = \dfrac{b_u + a_u}{2}$ y radio $r_u = \dfrac{b_u - a_u}{2}$, i.e., $Y(w_u) = B(c_u, r_u) =< c_u, r_u >$

Estadísticos descriptivos univariantes

Dada una variable de intervalo $Y_j \equiv Z$, la realización de Z en el individuo $u \in E$ es el valor del intervalo $Z(u) = [a_u, b_u]$. Se asume que los vectores de descripción individuales $x \in vir(d_u)$ se distribuyen de manera uniforme sobre Z(u).

Función de distribución empírica:

$$F_Z(\xi) = \frac{1}{m}\left(g_Z(E) + \sum_{\xi \in Z(u)} \left(\frac{\xi - a_u}{b_u - a_u} \right) \right)$$

donde $g_Z(\xi)$ es el número de intervalos de la variable Z para los cuales ξ es mayor o igual que su extremo superior.

Función de densidad empírica

$$f_Z(\xi) = \frac{1}{m} \sum_{u \in E} \left(\frac{I_u(\xi)}{||Z(u)||} \right) \ , con\, \xi \in I\!R$$

donde $I_u(\cdot)$ es la función que indica si ξ está o no en el intervalo $Z(u)$ y donde $||Z(u)||$ es la longitud del intervalo. Nótese que el sumatorio se realiza sólo sobre aquellos individuos u para los que $\xi \in Z(u)$.

Media muestral simbólica

$$\overline{Z} = \frac{1}{2m} \sum_{u \in E} (b_u + a_u)$$

Varianza muestral simbólica

$$S^2 = \frac{1}{3m} \sum_{u \in E} (b_u^2 + b_u a_u + a_u^2) - \frac{1}{4m^2} \left[\sum_{u \in E} (b_u + a_u) \right]^2$$

Estadísticos descriptivos bivariantes

Consideremos dos variables simbólicas de intervalo $\mathbf{Y} = \{Y_{j1}, Y_{j2}\}$ valoradas para cada individuo $w_u \in E$ con $u = 1, \ldots, m$. Cada una de estas variables Y_j tomarí valores en un subconjunto $\mathcal{Y}_j$ de la recta real $\mathbb{R}$, $\mathcal{Y}_j \subseteq \mathbb{R}$, $\mathcal{Y}_j = \{[\alpha, \beta], -\infty < \alpha, \beta < \infty\}$. Renombramos a las variables de la siguiente manera $Y_{j1} \equiv Z_1$ y $Y_{j2} \equiv Z_2$. Dichas variables toman como valor para cada individuo $w_u \in E$ un rectángulo $\mathbf{Z} = Z_1(u) \times Z_2(u) = ([a_{u1}, b_{u1}], [a_{u2}, b_{u2}])$. Asumimos también que los vectores descripción individuales se distribuyen uniformemente sobre los intervalos $Z_1(u)$ y $Z_2(u)$.

Función de distribución conjunta empírica:

$$F_Z(\xi_1,\xi_2)=\frac{1}{m}\left(g_Z(\xi_1,\xi_2)+\sum_{\xi_1<b_{u1}o\xi_2<b_{u2}}\left(\frac{\xi_1-a_{u1}}{b_{u1}-a_{u1}}\right)\left(\frac{\xi_2-a_{u2}}{b_{u2}-a_{u2}}\right)\right)$$

Función de densidad conjunta empírica

$$f_Z(\xi_1,\xi_2)=\frac{1}{m}\sum_{u\in E}\frac{I_u(\xi_1,\xi_2)}{||Z(u)||}$$

Función de covarianza empírica

$$Cov(Z_1, Z_2) = \frac{1}{6n} \sum_{u \in E} [2G_1 + Q_1 + Q_2 + 2G_2)]$$

donde,

$$G_1 = (a_{uj} - \overline{Z}_1)(c_{uj} - \overline{Z}_2)$$

$$Q_1 = (a_{uj} - \overline{Z}_1)(d_{uj} - \overline{Z}_2)$$

$$G_2 = (b_{uj} - \overline{Z}_1)(c_{uj} - \overline{Z}_2)$$

$$Q_2 = (b_{uj} - \overline{Z}_1)(d_{ij} - \overline{Z}_2))$$

Coeficiente de correlación

$$r(Z_1, Z_2) = \frac{Cov(Z_1, Z_2)}{S_{Z_1} S_{Z_2}}$$

Material y Métodos

Los modelos de regresión simbólica de intervalos que se aplican en este análisis son:
Método del Centro: los coeficientes del modelo se estiman aplicando el modelo clásico al punto medio de los intervalos.
Método Mínimo Máximo: ajusta dos modelos de regresión lineal, para los límites inferior y superior del intervalo.
Método del Centro y del Rango: ajusta dos modelos de regresión lineal, para el punto medio y el rango del intervalo.
Método de Centros y Rangos Restringido: coeficientes del modelo de centros se estiman como en el modelo anterior y coeficientes del modelo de rango se estiman utilizando el algoritmo de Lawson y Hanson.

Método de Covarianza Simbólico : Emplea el modelo de regresión lineal centralizada (modelo con variables centradas). Los coeficientes de regresión se estiman mediante el método de mínimos cuadrados, utilizando la matriz de covarianza simbólica. Debido a que el proceso de cálculo de la covarianza muestral simbólica utiliza los límites inferior y superior de cada variable, el método refleja la variabilidad del intervalo, como la dependencia entre todas las variables de una manera integral.

Material y Métodos

Los métodos presentados se aplican a la tabla simbólica de datos,generada utilizando el paquete RSDA dentro de la plataforma R construida.
Rendimiento de los modelos. (Lima Neto y De Carvalho [8]).
La evaluación del desempeño de los modelos de regresión lineal descriptos se basan en el error cuadrático medio para límites inferiores y superiores, que miden las diferencias entre los valores predichos y los valores observados:

$$RMSE_L = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(Y_{(iL)} - \widehat{Y}_{(iL)}\right)^2} \quad RMSE_U = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(Y_{(iU)} - \widehat{Y}_{(iU)}\right)^2}$$

Resultados

Para comparar los distintos métodos se ha utilizado, dentro de la plataforma R, el paquete “iRegression”. Este paquete tiene como autores a Lima Neto E., De Souza Filho C. y Marinho P. Contiene algunos métodos de regresión para variables simbólicas de intervalo.

Metodología

Muestra seleccionada para aplicar los modelos.

Conjunto de datos Empresas

Para este estudio se consideraron un conjunto de empresas argentinas que cotizan públicamente sus acciones, entre los años 2004 a 2012, excluyendo las que pertenecen al sector financiero . A partir de la información disponible en el sitio web de la Bolsa de Comercio de Buenos Aires (http://www.bolsar.com/net/principal/contenido.aspx),la Dra:Maria Inés Stímolo (UNC) construyó una base de datos con los balances esquematizados anuales entre los años 2004 y 2012.

Estas empresas pertenecen a distintos sectores económicos con una estructura de costos muy diferenciada. La Dra:Maria Inés Stímolo en su tesís para acceder al grado de doctor(2014), trabajó con la misma , a fín de analizar el comportamiento de los costos según el modelo empírico propuesto por Anderson(2003)

Modelo empírico

El modelo empírico ABJ propuesto Anderson (2003)

$$\ln\left[\frac{C_{i,t}}{C_{i,t-1}}\right] = \beta_0 + \beta_1 \times \ln\left[\frac{V_{i,t}}{V_{i,t-1}}\right] + \beta_2 \times D_{i,t} \times \ln\left[\frac{V_{i,t}}{V_{i,t-1}}\right] + \varepsilon_{i,t}$$

donde

$C_{i,t}$ es el costo (según la medición que se realice) de la empresa i en el año t

$V_{i,t}$ es el ingreso o nivel de actividad de la empresa i en el año t

$D_{i,t}$ es una variable dummy que asume el valor 1 cuando el ingreso o nivel de actividad aumenta en la empresa i en el año t.

Tabla Simbólica

Comportamiento de Costos en Empresas Argentinas

Empresas	X1	X2	Y
AERO	[-0.01,0.31]	[-0.01,0]	[-0.05,0.08]
AGRO	[-0.38,0.4]	[-0.38,0]	[-0.36,0.43]
ALPA	[0.17,0.26]	[0,0]	[0.09,0.3]
ALTO	[-0.1,0.44]	[-0.1,0]	[-0.04,0.24]
ALUA	[-0.11,0.22]	[-0.11,0]	[-0.05,0.47]
APSA	[-0.18,0.37]	[-0.18,0]	[-0.53,0.47]
ARCO	[0.02,0.23]	[0,0]	[0.01,0.22]
BOLT	[-0.03,0.58]	[-0.03,0]	[0.05,0.67]
CAPU	[-0.4,1.08]	[-0.4,0]	[-1.12,1.68]
...	...	...	...
...	...	...	...
YPFD	[-0.12,0.12]	[-0.12,0]	[-0.14,0.24]

Exploración de algunos modelos propuestos

Modelos	*Expresiones que permiten predecir los intervalos*
CM	$\widehat{c}_{Lnc\_total}(j) = 0.0138 + 0.8822 C_{ventas}(j) - 0.0542 C_{DECvtas}(j)$
MinMax	$\widehat{\underline{I}}_{Lnc\_total}(j) = -0.020 + 1.0918 \underline{I}_{ventas}(j) - 0.2826 \underline{I}_{DECvtas}(j)$ $\widehat{\overline{I}}_{Lnc\_total}(j) = 0.0049 + 0.9966 \overline{I}_{ventas}(j) + 0.7529 \overline{I}_{DECvtas}(j)$
CRM	$\widehat{c}_{Lnc\_total}(j) = 0.0138 + 0.8822 C_{ventas}(j) - 0.0542 C_{DECvtas}(j)$ $\widehat{r}_{Lnc\_total}(j) = -0.0042 + 1.170 r_{ventas}(j) - 0.5404 r_{DECvtas}(j)$
CCRM	$\widehat{c}_{Lnc\_total}(j) = 0.0138 + 0.8822 C_{ventas}(j) - 0.0542 C_{DECvtas}(j)$ $\widehat{r}_{Lnc\_total}(j) = 0.0000 + 0.9988 r_{ventas}(j) + 0.0000 r_{DECvtas}(j)$
SCM	$\widehat{Y}_{Lnc\_total}(j) = 0.0022 + 0.9749 \overline{I}_{ventas}(j) - 0.0885 \overline{I}_{DECvtas}(j)$

Tabla 2

Evaluación del rendimiento de los modelos

Modelos	$RMSE_L$	$RMSE_U$
CM	0.1354	0.1361
MinMax	0.1321	0.1305
CRM	0.1287	0.1302
CCRM	0.1312	0.1321
SCM	0.1406	$0,1564$

Tabla 3 Empresas

Evaluación del rendimiento de los modelos

Conjunto de datos de Murciélagos
Se presentan métodos de regresión simbólica aplicados a datos obtenidos por agregación a partir de la matriz de datos Murciélagos, murciélagos que viven en Europa se registran en términos de especie, longitud de la cabeza, longitud de la cola, longitud del antebrazo y peso. Información más detallada sobre el conjunto de datos se puede encontrar en Douzal-Chouakria et al. (2009).

Modelos	$RMSE_L$	$RMSE_U$	r
CM	$4,47$	$4,76$	$0,955$
CRM	$3,74$	$4,08$	$0,967$
SCM	$4,42$	$4,62$	$0,956$

Tabla 4 Murciélagos

Evaluación del rendimiento de los modelos

Conjunto de datos de hongos

El conjunto de datos de hongos, extraído del índice de especies de hongos de California, contiene mediciones de tres características que refieren a estipe (descripción del tallo del hongo) y capa de pileo (sombreo del hongo) de 100 especies de hongos, miembros del género Agaricies. Los métodos presentados se aplican a la tabla simbólica de datos, construida tomando como concepto cada especie de hongo,como variable respuesta el ancho de la capa del píleo y longitud del estipe como variable explicativa.

Modelos	$RMSE_L$	$RMSE_U$	r
CM	$1,89$	$5,02$	$0,629$
CRM	$1,58$	$4,91$	$0,642$
SCM	$1,82$	$4,99$	$0,629$

Tabla 5 Hongos

Conclusiones

En este trabajo se han presentado algunos modelos de regresión lineal simbólica para variables de intervalo, los cuáles han sido probados usando varios conjuntos de datos. Se ha mostrado las ventajas de trabajar con regresión simbólica en el caso de tener clases de individuos en lugar de entes individuales y por ende intervalos de valores tanto para las variables explicativas como de respuesta. La extensión de los supuestos probabilísticos, presentes en la teoría del modelo de regresión lineal clásica, al caso de datos de intervalo, sigue siendo un tema de análisis en la teoría de datos simbólicos que requiere más investigación. Esto motiva que a futuro se desarrollen metodologías de regresión simbólica, utilizando la aritmética de intervalos y respuestas de naturaleza aleatoria.

Referencias

[1] Billard and Diday (2000). Regression analysis for interval-valued data.In: Proc. of IFCS00, Belgium, pp. 369-374,Springer.
[2] Billard, L., Diday, E. (2002). Symbolic Regression Analysis. In: Proc.IFCS02, Poland, pp. 281-288, Springer.
[3] Billard, L. (2007). Dependencies and Variation Components of Symbolic Interval-Valued Data.
[4] Billard, L. (2008). Sample Covariance Functions for Complex Quantitative Data. Processing, World Conferences International Association of Statistical Computing 2008, Yokohama, Japan.
[5] Billard, L. and Diday, E. (2000). Regression Analysis for Interval-Valued Data. Data analysis, Classificaiton, and Related Methods (eds. H.A.L. Kiers, J.-P.
[6] Billard, L. and Diday, E. (2003). From the Statistics of Data to the Statistics of Knowledge: Symbolic Data Analysis.

Referencias

[7] De Carvalho, F.A.T., Neto, E.A.L., (2010). Centre and Range method for fitting a linear regression model to symbolic intervalar data.Computational Statistics, Data Analysis.
[8] Lima Neto, E.A., de Carvalho F.A.T. and Freire, E.S. (2005). Applying Constrained Linear Aggression Models to Predict Interval-Valued Data.
[9] Billard, L. and Diday, E. (2007). Symbolic Data Analysis: Conceptual Statistics and Data Mining. Wiley, Chichester.
[10] Xu, W. (2010), Symbolic data analysis: regression of data with interval values.
[11] Douzal-Chouakria, A., Billard, L. and Diday E. (2009). Principal Component Analysis for Interval-valued Observations. Submitted manuscript.
[12] Lima Neto, E.A and de Carvalho F.A.T. (2010). Constrained Linear Regression Models for Symbolic Interval-valued Variables.

MUCHAS GRACIAS!!