Selección del mejor modelo de regresión usando una combinación convexa del método de los Mínimos Cuadráticos y el de la Mínima Desviación Absoluta Media

Carlos N. Bouza Herrera, Sira M. Allende Alonso y Gemayqzel Bouza Allende
Universidad de La Habana

1. Introducción

En muchos problemas económicos el ajuste de un modelo de regresión es requerido pero hay presentes observaciones que afectan el buen ajuste usando el métodos de los Mínimos Cuadráticos [MC]. Estas observaciones son llamadas extremales [outliers] y una solución es utilizar otro método de ajuste como el la Mínima Desviación Absoluta Media [Mindam]. Este es menos sensible a las observaciones extremales pero no es una solución totalmente efectiva. El uso de una combinación convexa de ellos fue propuesta por Dodge (1984) y extendido al caso en que no se conoce la proporción de ellas por Allende-Bouza (1993).
Uno de los problemas a resolver para hacer uso efectivo de las alternativas propuestas es el de poder seleccionar modelos como comúnmente se hace con al usar los MC. En este trabajo se sugiere solucionar esto a partir de heurísticas basadas en el Recocido Simulado. Se desarrollan algoritmos y se hace una ilustración a partir del ajuste de una ecuación de regresión para predecir la valoración de un enclave recreacional en Veracruz, México.

2. El problema de optimización

El modelo de regresión ;lineal usual está dado por

$$Y_i = x_i\beta + \varepsilon_i$$

donde

$$\beta \in \Re^{p+1}$$

$$x_i = \left(1, x_{i1}, \ldots x_{ip}\right)$$

y

$$E(\varepsilon_i) = 0, Var(\varepsilon_i) = \sigma^2$$

Generalmente se supone que la distribución de los errores ε_j ‘s es una normal y ellos son independientes. Este es considerado como un modelo central *F**. Bajo este el método de los

MC es óptimo. Cuando observamos valores poco probables de los errores dudamos de que F^* los genere. Tal es el caso en muchos problemas económicos en los que factores externos afectan las respuestas de los agentes y los sistemas. Decimos que un error es extremal si $\varepsilon \notin [-3\sigma, 3\sigma]$. Entonces se considera que en realidad estas variables tienen otra distribución y que el modelo conque trabajamos es un distribución contaminada de F^* con otra F'.

Entonces se trabaja con distribuciones de una vecindad λ-contaminada

$\wp = \{F | F = \lambda F^* + (1-\lambda)F', \lambda \in \{0,1]\}$

La Estadística Robusta sugiere el uso de un M-estimador del tipo un-paso (one-step M-Estimator) , Bickel (1984). Estos son funciones del tipo:

$$B(n, Y_1, ..., Y_n) = \beta^*(Y_1, ..., Y_n) + \sigma_0 \sum \Psi\left(\frac{Y_i - x_i^T \beta^*(Y_1, ..., Y_n)}{\sigma_0} \right)$$

donde $\beta^*(Y_1, ..., Y_n)$ es un estimador inicial de β, generalmente obtenido mediante el método de los MC , y σ_0 es un estimador de la escala. Las condiciones para su normalidad asintótica requiere de una serie de hipótesis sobre la regularidad de la función de distribución, ver Müller (1994).

Dodge (1984) propuso el siguiente problema de programación paramétrica para resolver el ajuste de una $F \in \wp$:

$$P1: Min_\beta \left(\lambda \sum_{i+1}^{n} \delta^2{}_i + (1-\lambda) \sum_{i+1}^{n} |\delta_i| \right)$$

$$\begin{aligned}
&sujeto \quad a\\
&X\beta - \delta \leq Y\\
&-X\beta - \delta \leq -Y\\
&\lambda \in [0,1] \quad fijo\\
&\delta = (\delta_1, ..., \delta_n)^T, \delta_i = Y_i - x_i^T \beta\\
&X = (x_1, ..., x_n)^T\\
&Y = (Y_1, ..., Y_n)^T
\end{aligned}$$

Allende -Bouza (1993) consideraron el caso en que λ era desconocido. Esto hace necesario determinar el parámetro de contaminación dentro del programa. El nuevo programa es:

$$P2: Min_{(\beta,\lambda)}\left(\sum_{i+1}^{n}\delta^2{}_i+\theta\sum_{i+1}^{n}|\beta_i|\right)$$

$$\begin{aligned}
&sujeto \quad a\\
&X\beta-\delta\leq Y\\
&-X\beta-\delta\leq -Y\\
&\theta=\frac{\lambda}{1-\lambda}, \qquad \lambda\in[0,1[\\
&\delta=(\delta_1,\ldots,\delta_n)^T, \delta_i=Y_i-x_i^T\beta\\
&X=(x_1,\ldots,x_n)^T\\
&Y=(Y_1,\ldots,Y_n)^T
\end{aligned}$$

En ese trabajo se abordó el estudio del problema de Complementariedad Lineal asociado y se propuso un algoritmo para resolverle. Este fue mejorado en Suárez Fariñas et. al (1997). Este reportó la posibilidad de resolver el problema brindando una partición del espacio parámetrico

$\cup_{k=1}^{K}\Lambda_k=[0,1]$

junto con el vector de coeficientes óptimos de la regresión $\beta(\lambda|\lambda\in\Lambda_k)$ para cada subintervalo. Queda en manos del decisor el seleccionar el mejor modelo. Es lógico que este use el modelo con λ^* que hace mínima la función objetivo.

Otro problema es el de seleccionar las variables que deben entrar en el modelo. A partir de tener p variables $X_1,\ldots,X_p$ el decisor desea evaluar otro modelo con p' variables. Esto lo hace al definir un conjunto de transformaciones de las variables iniciales y evaluar si estas deben entrar a formar parte del modelo. Eventualmente incluirles puede conllevar la eliminación de la variable analizada. Tomemos $M(G_1,\ldots,G_m)$ como una familia de tales transformaciones. Se analizará para una variable X_j si la inclusión de $G_h(X_j)$ redunda en una mejor ecuación de regresión. Por ejemplo podemos usar $M(X_j^2, \log X_j, , 0)$.

Los criterios básicos utilizados en el estudio del ajuste de los modelos son el Ascendente y el Descendente, ver Johnson y Wichern (1998). En el primero se parte de un grupo exiguo de variables y se van incrementando su número . En el segundo se utiliza un modelo con muchas variables y se van eliminando las menos importantes en la predicción. Bajo las hipótesis para las cuales el método de los MC es óptimo la entrada o eliminación de variables se lleva a cabo a usando pruebas del tipo F. En nuestro caso esto no es posible dada la no normalidad envuelta generalmente al trabajar con distribuciones contaminadas.

3. Los algoritmos

El método del Recosido simulado es una poderosa herramienta de búsqueda estocástica que se ha hecho muy popular dado el amplio espectro de problemas que puede resolver. En particular en el área de la optimización combinatoria se ha hecho de un nicho investigativo. Este es describible al fijar un espacio de configuraciones o espacio de soluciones:

$S=\{z \mid z=(z_1,\ldots,z_m)\}$

m es llamado dimensión del espacio . Además se define una función de costo $C:S \longrightarrow \Re$ que le asigna un valor real a cada configuración. El objetivo es hallar la configuración óptima: $z^* \in S$ tal que $\forall z' \in S$ entonces $C(z^*) \leq C(z')$. La convergencia al óptimo global se obtiene a partir de la modelación del proceso del Recocido Simulado mediante Cadenas de Markov, ver Pflug (1996).

Para el criterio de selección Ascendente desarrollamos el algoritmo A1.

A1. Selección de modelos mediante el criterio Ascendente

Initialize $\{\beta(\lambda^*), t\ e, G_1,\ldots,G_H, \alpha\}$

While t>e do

Repeat for r=1 to R

Repeat for h=1 to H

Begin

Seleccionar aleatoriamente $u \in (0,1,\ldots,p\}$ y $s \in (G_1,\ldots, G_H\}$

Resuelva P2 con $x_i=(1,x_{i1},\ldots,x_{ip}, G_s[x_u])$, i=1,...,n.

If $L(\lambda_r^*) = \text{Mín}_{\lambda \in [0,1]}$ $$\frac{\sum_{i+1}^{n}\delta_i^2+\frac{\lambda_r^*}{1-\lambda_r^*}\sum_{i=1}^{n}|\delta_i| - \sum_{i+1}^{n}\delta_i^2+\frac{\lambda^*}{1-\lambda^*}\sum_{i=1}^{n}|\delta_i|}{\sum_{i+1}^{n}\delta_i^2+\frac{\lambda^*}{1-\lambda^*}\sum_{i=1}^{n}|\delta_i|} = \frac{L(\lambda_r^*)-L(\lambda^*)}{L(\lambda^*)} \leq e$$

Then $\beta(\lambda^*)=\beta(\lambda^*_r)$

Else if $\exp\left(\frac{L(\lambda^*) - L(\lambda^*_r)}{tL(\lambda^*)}\right) > rand[0,1]$

then $\beta(\lambda^*)=\beta(\lambda^*_r)$ and $L(\lambda^*)=L(\lambda^*_r)$

End

r=r+1

End

$t=\alpha t$

h=h+1

END

El decisor fija las funciones a utilizar , el parámetro de enfriamiento y una cota e que fija la aceptación de un modelo aunque no se obtenga una mejoría.

El algoritmo para el caso Descendente es similar A1 en el núcleo asociado al Recocido Simulado pero es más simple pues solo se elige la variable cuya posible eliminación es avaluada. Este es

A1. Selección de modelos mediante el criterio Ascendente

Initialize $\{\beta(\lambda^*), t\ e, G_1,...,G_H, \alpha\}$

While t>e do

Repeat for r=1 to R

Repeat for h=1 to H

Begin

Seleccionar aleatoriamente u∈(0,1,...,p}

Resuelva P2 con $x_i=(1,x_{i1},...,x_{iu-1}, x_{iu+1}, ..., x_{ip},)$, i=1,...,n.

If $L(\lambda_r{}^*)=$ Mín $_{\lambda\in[0,1]}$ $$\frac{\sum_{i+1}^{n}\delta_i^2 + \frac{\lambda_r{}^*}{1-\lambda_r{}^*}\sum_{i=1}^{n}|\delta_i| - \sum_{i+1}^{n}\delta_i^2 + \frac{\lambda^*}{1-\lambda^*}\sum_{i=1}^{n}|\delta_i|}{\sum_{i+1}^{n}\delta_i^2 + \frac{\lambda^*}{1-\lambda^*}\sum_{i=1}^{n}|\delta_i|} = \frac{L(\lambda_r{}^*) - L(\lambda^*)}{L(\lambda^*)} \le e$$

Then $\beta(\lambda^*)=\beta(\lambda^*_r)$

Else if $\exp\left(\frac{L(\lambda^*) - L(\lambda^*_r)}{tL(\lambda^*)}\right) > rand[0,1]$

then $\beta(\lambda^*)=\beta(\lambda^*_r)$ and $L(\lambda^*)=L(\lambda^*_r)$

End

r=r+1

End

$t=\alpha t$

h=h+1

END

Estos algoritmos están en estudio pero fueron aplicados en el estudio de las preferencias de los posibles usuarios de un nuevo parque recreacional a desarrollar en las playas de Veracruz, México, ver Castro (2000). La propuesta del Parque Acuático se analizó usando técnicas de Análisis Conjuntual (Conjoint Analysis). Al modelar la relación entre la valoración (utilidad de preferencia) del parque y la utilidad asignada a los niveles de los atributos (ofertas y sus niveles) se ajustaron diversas ecuaciones de regresión. Analizamos el comportamiento de P2 en el ajuste de los datos obtenidos en 77 entrevistados de los cuales 22 podían ser considerados como "observaciones extremales". Las variables fueron las utilidades de las siguientes ofertas:

1. Atracción Acuática

1.1. Toboganes Gigantes.

12. Alberca de Olas.

1.3 Río Salvaje

2. Espectáculo Adicional.

2.1 Canchas de Voly y Fútbol Playero.

2.2. Delfines y Ballenas.

2.3. Juegos Marinos

3. Gastronomía

3.1. Restaurantes.

3.2. Comida Rápida.

3.3. Áreas da Autopreparación.

4. Precio de Acceso por persona.

4.1. 45, 00 PM

4.2. 60,00 PM.

4.3. 90,00 PM.

El modeloaditivo fue

Preferencia= $\sum^{4}_{i=3} \sum^{3}_{j=1} \beta_{ij}$ utilidad$_{ij}$ +ε

Los resultados obtenidos en el ajuste aparecen en la tabla siguiente:

Tabla con el valor de $\frac{1}{n}\left[\lambda * \sum_{i=1}^{n} \delta_i^2 + (1-\lambda *)\sum_{i=1}^{n} |\beta_i|\right]$

θ=λ/(1-λ)	Ascendente	Descendente
0(MC)	63,682	70,338
1 (Mindam)	77,561	73,415
0,852 (Optimo en General)	43,630	47,325

Como se ve el método ascendente es el más recomendable en términos de la función objetivo . Sin embargo el tiempo de cómputo del descendente fue el 74,2% del usado por el Ascendente. Se requiere de una mayor experimentación computacional antes de valorar la efectividad de estas heurísticas. En particular la implementación del Paso-a-Paso parece requerir de algoritmos Genéticos para valorar los nuevos modelos como hijos de los previos.

Referencias

Allende S. and Bouza, C. (1993): A parametric programming approach to the estimation of the coefficient of a linear regression model. En “Approximation and Optimization”, P.Lang Verlag, Berlin. 9-21.

Castro López, C. R. (2000) Aportación Estadística al Conjoint Analysis. Tessi de MSC. Universidad de La Habana.

Dodge, Y. (1984): Robust estimation of the regression coefficients by minimising a convex combination of LS and LAD. Comp. Stat. Quart. J. 1, 139-153.

Johnson, R.A. y D.W. Wichern (1998): Applied Multivariate Analysis. Prentice Hall, N. Jersey.
Müller, Ch. H. (1994): Optimal designs for robust estimation in conditionally contaminated models. J. of Stat. Planning and Inference. 38, 125-140.
Pflug, G.Ch. (1996): Optimization of Stochstic Moedls: The interface between Simultaion and Optimization.. Kluwer Academic Publishers, Massachussetts.
Suárez Fariñas, M., S. A. Garcet Rodríguez, S. Allende Alonso y C. Bouza Herrera. (1997): El problema de complementariedad lineal: algoritmos de solución y aplicación al ajuste del modelo de regresión lineal multivariado con LS-LAD. Inv. Operacional, 18, 11-56.