Encuentro de Jóvenes Investigadores

MÍNIMOS CUADRADOS PARCIALES

Basa, Jerónimo[1]

[1] *Facultad de Ingeniería Química*

Director/a: Forzani, Liliana
Codirector/a: Marcos, Miguel Andres

Área: Ciencias Exactas

Palabras claves: Regresión, PLS, Dimensión

INTRODUCCIÓN

Los métodos de regresión son una pieza fundamental en el análisis estadístico, ya que permiten explicar la relación entre una o varias variables (respuesta) y otro conjunto de variables, que llamaremos predictoras. La versión más sencilla que conocemos es la regresión lineal, en el que un conjunto de n datos $(\boldsymbol{X}_i, Y_i)$ se relacionan siguiendo el siguiente modelo

$$Y_i = \beta^T \boldsymbol{X}_i + \varepsilon_i, \quad i = 1, \dots, n \quad Y \in \mathbb{R}, \boldsymbol{X} \in \mathbb{R}^p \tag{1}$$

donde los datos son muestras independientes, idénticamente distribuidas del par $(\boldsymbol{X}, Y)$, los errores ε_i son independientes con una distribución normal de media 0 y varianza τ^2, y $\beta \in \mathbb{R}^p$ es un parámetro desconocido a estimar. La ecuación (1) es lo que se llama *versión muestral* del modelo poblacional

$$Y = \beta^T \boldsymbol{X} + \varepsilon$$

pero, ¿cómo estima uno el vector de parámetros desconocido β? Haciendo algunas cuentas clásicas de estadística podemos razonar del siguiente modo. Llamemos $\sigma = E(\boldsymbol{X}Y)$ y $\boldsymbol{\Sigma} = E(\boldsymbol{X}\boldsymbol{X}^T)$.

Título del proyecto: Big data para quimiometría: estadística detrás de Partial Least Squares
Año convocatoria: 2020
Organismo financiador: CONICET
Director/a: Forzani, Liliana

Federación Universitaria del Litoral

UNL UNIVERSIDAD NACIONAL DEL LITORAL

Entonces

$$\begin{aligned} Y = \beta^T \boldsymbol{X} + \varepsilon &\Longrightarrow \boldsymbol{X}Y = \boldsymbol{X}\boldsymbol{X}^T\beta + \boldsymbol{X}\varepsilon \\ &\Longrightarrow E(\boldsymbol{X}Y) = E(\boldsymbol{X}\boldsymbol{X}^T)\beta \\ &\Longrightarrow \sigma = \boldsymbol{\Sigma}\beta \\ &\Longrightarrow \beta = \boldsymbol{\Sigma}^{-1}\sigma \end{aligned}$$

Esto nos da una fórmula funcional para β, que es $\beta = \boldsymbol{\Sigma}^{-1}\sigma$. Por lo tanto, para el caso de la versión muestral podemos *insertar* (lo que en estadística clásica se conoce como pluggin) las versiones muestrales; $\widehat{\beta} = \widehat{\boldsymbol{\Sigma}}^{-1}\widehat{\sigma}$, donde usamos el *hat* $\widehat{\ }$ para denotar que es un estimador. En general, uno aquí considera los estimadores clásicos

$$\widehat{\sigma} = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})(\boldsymbol{X}_i - \bar{\boldsymbol{X}}) \quad \text{y} \quad \widehat{\boldsymbol{\Sigma}} = \frac{1}{n-1}\sum_{i=1}^{n}(\boldsymbol{X}_i - \bar{\boldsymbol{X}})(\boldsymbol{X}_i - \bar{\boldsymbol{X}})^T,$$

donde $\bar{Y}, \bar{\boldsymbol{X}}$ son los respectivos promedios.

Esta sencilla forma que se deriva del modelo depende fuertemente de la existencia de $\widehat{\boldsymbol{\Sigma}}^{-1}$, lo cual es algo que tenemos asegurado en el caso en que $p < n$, es decir, la cantidad de observaciones es mayor a la cantidad de información que tengo sobre cada observación. Sin embargo, en los últimos años han surgido problemas que parecen no respetar esta relación; por el contrario, dichos problemas cumplen la relación $p > n$. Esto representa un problema ya que $\widehat{\boldsymbol{\Sigma}}^{-1}$ no existe. El contexto de $p \gg n$ es lo que hoy en día se conoce como *alta dimensión* (o *big data*). Uno de los casos más destacados en este reino es el de la quimiometría. Una situación concreta de ésta puede ser cuando la variable respuesta Y es una concentración química (por ejemplo de una proteína), y predictoras $\boldsymbol{X} = (X_1, \dots, X_p)$ son absorciones de luz a distintas longitudes de onda o medidas que obtenemos (p de ellas) usando algún tipo de instrumento espectroscópico.

Esto abrió un nuevo mundo de posibilidades donde se intenta contestar la sencilla pregunta, ¿cómo estimamos los parámetros de un modelo en un contexto de alta dimensión? Durante un tiempo se creyó que la mejor respuesta era considerar lo que se conoce como *sparsity*. El modelo *sparse* asume que en realidad sólo algunas de las variables $X_1, \dots, X_p$ aportan información relevante al modelo. Si este es el caso, entonces tiene sentido eliminar algunas de ellas, lo que conduce a reducir la dimensión del problema; ahora sólo tenemos $\tilde{p}$ variables predictoras con $\tilde{p} \ll p$. Sin embargo, esta eliminación de variables suele ser peligrosa, ya que elimina información, hace que la interpretación sea engañosa y aumenta el riesgo de obtener modelos falsos.

Es por ello que ha surgido la necesidad de buscar alternativas mejores para el problema de estimar parámetros en alta dimensión. Uno de ellos es el objeto de nuestro trabajo, conocido como *Partial Least Squares* (PLS). Este interesante método fue propuesto por Herman Wold

como un algoritmo llamado NIPALS. Éste nace con aplicaciones a la econometría y luego a las ciencias sociales, y daba muy buenos resultados. Pero al tratarse sólo de un algoritmo no se conocían ninguna de sus propiedades estadísticas, ya que no había desarrollado un fundamento teórico detrás de él. Nuestro estudio se concentra justamente en el modelo PLS, que además lo consideraremos en un caso particular que corresponde a $u = 1$, lo que significa realizar un paso del algoritmo PLS.

OBJETIVOS

Se plantearon los siguientes objetivos

- Estudiar el modelo poblacional correspondiente al método conocido como PLS, utilizado ampliamente en la literatura de quimiometría.
- Derivar propiedades del estimador para $\beta^T \boldsymbol{X}_N$ como su distribución asintótica, para construir intervalos de confianza a partir de ella. Aquí $\boldsymbol{X}_N$ representa una nueva observación independiente de $\boldsymbol{X}$.
- Realizar simulaciones para sustentar las ideas teóricas desarrolladas y poder aplicarlas a bases de datos reales.
- Investigar qué ocurre para los casos en que n y p crecen, teniendo en cuenta que en las aplicaciones reales uno tiene cantidades fijas, en la mayoría donde $p \gg n$.

METODOLOGÍA

En la bibliografía presentada aquí se mencionan los resultados de consistencia para el estimador del parámetro en el modelo de PLS, para comportamientos particulares de n y p creciendo. Sin embargo, esto dejó abierta la pregunta de cómo construir intervalos de confianza (o plantear test de hipótesis) para este estimador. A partir de aquí comenzó nuestra búsqueda de dar con un resultado asintótico que pueda dar con las respuestas. Una vez probada la convergencia asintótica que buscábamos, se realizaron una serie de simulaciones en el software *R* que confirmaron los resultados teóricos encontrados.

CONCLUSIONES

Se encontró una demostración (aún bajo revisión) sobre el comportamiento de la predicción de $\beta^T \boldsymbol{X}_N$ para una nueva observación de la variable $\boldsymbol{X}$, bajo el modelo PLS. Los gráficos construidos por las simulaciones realizadas han mostrado el comportamiento esperado. Además, observamos que las hipótesis a considerar para este modelo pueden producir (o no) la existencia de un sesgo. Muchas de las conclusiones que se pueden hallar dependen fuertemente de la relación entre n y p.

Encuentro de Jóvenes Investigadores

BIBLIOGRAFÍA BÁSICA

Cook, R.D. and Forzani, L. (2018), *Big data and partial least?squares prediction*. Can. J. Statistics, 46: 62-78. doi:10.1002/cjs.11316

Cook, R.D., Helland, I.S., Z. Su. *Envelopes and Partial Least Squares Regression.* Journal of the Royal Statistical Society. Series B (Statistical Methodology), vol. 75, no. 5, 2013, pp. 851?877., www.jstor.org/stable/24772470.

Federación Universitaria del Litoral

UNL UNIVERSIDAD NACIONAL DEL LITORAL