Métodos de Predicción en Situaciones Límite

Autor: Univ. José Alberto Flores Aguilar

1. Pasado, presente y futuro sobre las limitaciones para relacionar variables

El problema del ajuste de un conjunto de puntos representados en un sistema de ejes coordenados, por una recta o mas generalmente por una curva, era objeto de estudio desde mediados del siglo XVIII (Leonhard Euler, 1749; Johan Tobías Mayer, 1750). Sin embargo la primera mención al *método de mínimos cuadrados*, fue atribuida a Adrián-Marie Legendre (1805). En dicho estudio, se consideró este método como: "el más adecuado para relacionar variables de forma lineal" señalándose, además la conveniencia de la eliminación de individuos atípicos para optimizar el establecimiento de dichas interrelaciones.

Por último, merece la pena destacar la introducción de mínimos cuadrados, realizado por Robert Adrián (1808), quien aportó un punto de vista de gran interés, complementario al de los trabajos realizados por sus antecesoras. Sin embargo, dicho método no pudo ser justificado hasta la llegada de la ley de Laplace-Gauss, "bautizada" por Kart Pearson, a finales del XIX (1893), como "ley normal".

Se ha constatado en numerosas ocasiones que la presencia de la multicolinealidad conlleva a situaciones de "inestabilidad" de los coeficientes de regresión y que estos pueden ser "no significativos".

Cuando las variables explicativas están muy correlacionadas con la variable a explicar, produciendo dificultades de interpretación de la ecuación de regresión linéal a causa de signos erráticos en los coeficientes de regresión. Por ello, la aplicación del método de "mínimos cuadrados" conduce a resultados en ocasiones poco comprensibles para los investigadores que se dedican a las ciencias experimentales.

Es interesante no solo detectar al multicolinealidad (Belsley, Kuh y Welsh, 1980), sino también tomar medidas para atenuarla, sin embargo, la ecuación de predicción lineal bajo estas medidas sigue siendo desgraciadamente en ocasiones poco comprensible para el investigador:

Otras dos situaciones límite son:

1) Número menor de individuos que variables y
2) Datos ausentes.

En cuanto a la primera situación, que contempla menos individuos que variables, conlleva sistemáticamente a que el determinante de la matriz X^tX -que hay que resolver para la obtención de los coeficientes de regresión- "sea nulo" y, por tanto, no haya modo de encontrar tales coeficientes.

De todos estos resultados concluimos que: el método de mínimos cuadrados -intensamente para relacionar variables- no funciona bien en las situaciones límite tales como:

- La multicolinealidad,
- Menor número de individuos que variables y
- Datos ausentes.

Es aconsejable sustituirlas, en esas circunstancias, por el *método de mínimos cuadrados parciales* (PLS)

2. Una reflexión en cuanto a la normalización de los datos

Vamos a presentar dos tipos de normalización de datos que se encuentran con frecuencia en las referencias bibliografías y en especial en Audrain, Lesquoy-de-Turckeim, Miller y Tomassone, 1992, Pág. 179-180. El primero consiste en restar para cada una de las variables su media, y dividir por la raíz cuadrada de la suma de las desviaciones a su media.

$$y_i^{[1]} = \frac{y_i - \bar{y}}{\sqrt{\sum_{i\to 1}^{n}(y_i - \bar{y})^2}} \quad (i = 1,2,....,n)$$

$$x_{i,j}^{[1]} = \frac{x_{i,j} - \bar{x}_j}{\sqrt{\sum_{i\to 1}^{n}(x_{i,j} - \bar{x}_j)^2}} \quad (i = 1,2,...,n)\,(j = 1,2,..,p)$$

El segundo consiste en restar para cada una de las variables su media y, dividir por la raíz cuadrada de la suma de cuadrados de las desviaciones a su media por (n-1).

$$y_i^{[2]} = \frac{y_i - \bar{y}}{\sqrt{\frac{\sum_{i\to 1}^{n}(y_i - \bar{y})^2}{n-1}}} \quad (i = 1,2,....,n)$$

$$x_{i,j}{}^{2} = \frac{x_{i,j} - \bar{x}_j}{\sqrt{\frac{\sum_{i\to 1}^{n}(x_{i,j} - \bar{x}_j)^2}{n-1}}} \quad (i = 1,2,...,n)\,(j = 1,2,....., p)$$

3. Cálculo de coeficientes de regresión para ambas normalizaciones

Aunque las operaciones intermedias que hay que realizar para llegar a los coeficientes de regresión difieran del tipo de normalización de los datos, los coeficientes de regresión asociados a las variables

$$x_1^{[1]}, x_2^{[1]}, ..., x_p^{[1]}$$

Como resultado de la primera normalización y a las variables.

$$x_1^{[2]}, x_2^{[2]}, ..., x_p^{[2]}$$

Como resultado de la segunda normalización, son los mismos.

- En la primera normalización de los datos las operaciones son las siguientes:

$$\sum_{i=1}^{n}\left(y_i^{[1]}\right)^2 = 1$$

$$\sum_{i=1}^{n}\left(x_{i,j}^{[1]}\right)^2 = 1 \qquad j = 1,2,...., p$$

$$\sum_{i=1}^{n}\left(y_i^{[1]} x_{i,j}^{[1]}\right) = r_{y,x_j} \qquad j = 1,2,....., p$$

$$\sum_{i=1}^{n}\left(x_{i,j}^{[1]} x_{i,j'}^{[1]}\right) = r_{x_j,x_{i,j'}} \quad (j = 1,2,....., p; j' \neq j)$$

Donde r_{y,x_j} representa la correlación muestral entre x_j e y.

- En la segunda normalización de los datos las operaciones son las siguientes:

$$\sum_{i=1}^{n}\left(y_i^{[2]}\right)^2 = n-1$$

$$\sum_{i=1}^{n}\left(x_{i,j}^{[2]}\right)^2 = n-1 \qquad j = 1,2,...., p$$

$$\sum_{i=1}^{n}\left(y_i^{[2]} x_{i,j}^{[2]}\right) = (n-1) r_{y,x_j} \qquad j = 1,2,....., p$$

$$\sum_{i=1}^{n}\left(x_{i,j}^{[2]} x_{i,j'}{}^{2}\right) = (n-1) r_{x_j,x_{i,j'}} \quad (j = 1,2,....., p; j' \neq j)$$

Tanto en la primera como en la segunda normalización de los datos los coeficientes de regresión afectados, tanto por unas como por otras variables, son los mismos.

Partimos de la expresión que nos permite calcular los coeficientes de regresión.

Para la primera normalización:

$$\hat{\beta}^{[1]} = (X^{[1]\prime} X^{[1]})^{-1} X^{[1]\prime} y^{[1]}$$

Dado que $X^{[1]\prime} X^{[1]} = R$, matraz de correlación de las variables explicativas.

$$X^{[1]\prime} Y^{[1]} = \begin{pmatrix} r_{y,x_1} \\ r_{y,x_2} \\ . \\ . \\ r_{y,x_p} \end{pmatrix}$$

La fórmula para l cálculo de los coeficientes de regresión es la que a continuación mostramos.

$$\hat{\beta}^{[1]} = R^{-1} \begin{pmatrix} r_{y,x_1} \\ r_{y,x_2} \\ . \\ . \\ r_{y,x_p} \end{pmatrix}$$

Para la segunda normalización:

$$\hat{\beta}^{[2]} = (X^{[2]\prime} X^{[2]})^{-1} X^{[2]\prime} y^{[2]}$$

Dado que,

$$X^{[2]\prime} X^{[2]} = (n-1)R \qquad \text{y}$$

$$X^{[2]\prime} Y^{[2]} = (n-1) \begin{pmatrix} r_{y,x_1} \\ r_{y,x_2} \\ . \\ . \\ r_{y,x_p} \end{pmatrix}$$

La fórmula para el cálculo de los coeficientes de regresión es la que a continuación mostramos.

$$\hat{\beta}^{[2]} = R^{-1} \begin{pmatrix} r_{y,x_1} \\ r_{y,x_2} \\ . \\ . \\ r_{y,x_p} \end{pmatrix}$$

De lo que concluimos que los coeficientes de regresión son invariantes con respecto al tipo de normalización.

$$\hat{\beta}^{[1]} = \hat{\beta}^{[2]}$$

4. Ecuaciones de predicción lineal

La ecuación de predicción lineal en función de las variables normalizadas por primer caso:

$$y^{[1]^*} = \sum_{j=1}^{p} \hat{\beta}_j^{[1]^*} x_j^{[1]}$$

Deshaciendo el cambio, llegamos a la ecuación de predicción en función de las variables originales.

$$\frac{y^* - \bar{y}}{\sqrt{\sum_{i=1}^{n} (y_i - \bar{y})^2}} = \sum_{j=1}^{p} \hat{\beta}_j^{[1]^*} \left(\frac{(x_{ij} - \bar{x}_j)}{\sum_{i=1}^{n} (x_{ij} - \bar{x}_j)^2} \right) \Rightarrow$$

$$y^* = \left(\bar{y} - \sum_{j=1}^{p} \sqrt{\frac{SCD_y}{SCD_{x_j}}}\, \hat{\beta}_j^{[1]^*} \bar{x}_j \right) + \sum_{j=1}^{p} \sqrt{\frac{SCD_y}{SCD_{x_j}}}\, \hat{\beta}_j^{[1]^*} x_j$$

donde

$$SCD_y = \sum_{i=1}^{n} (y_i - \bar{y})^2 \qquad \text{y}$$

$$SCD_{x_j} = \sum_{i=1}^{n} (x_{ij} - \bar{x}_j)^2$$

La ecuación de predicción lineal en función de las variables normalizadas por el segundo caso:

$$y^{[1]^*} = \sum_{j=1}^{p} \hat{\beta}_j^{[1]^*} x_j^{[1]}$$

Deshaciendo el cambio, llegamos a la ecuación de predicción en función de las variables originales.

$$y^* = \left(\bar{y} - \sum_{j=1}^{p} \sqrt{\frac{SCD_y}{SCD_{x_j}}}\, \hat{\beta}_j^{[2]^*} \bar{x}_j \right) + \sum_{j=1}^{p} \sqrt{\frac{SCD_y}{SCD_{x_j}}}\, \hat{\beta}_j^{[2]^*} x_j$$

Dado que: $\hat{\beta}^{[1]^*} = \hat{\beta}^{[2]^*} \quad (j = 1,2,\ldots\ldots, p)$ se concluye que el tipo de normalización no influye en la ecuación de predicción lineal.

5. Bibliografía

Es bastante difícil encontrar bibliografía sobre este tema. De hecho no existe ninguna publicación en español; generalmente aparecen resultados dentro del campo de la química, por ser precisos en el campo de la cromatografía y espectrofotometría.

Incluyo los más fundamentales al ser realizados por los creadores del método.

[1] *Bry, X* (1996): Analices factorielles multiples, Paris, Ed.Economia.

En este libro no solo se ve la manera didáctica de la regresión PLS, sino también la relación que existe entre la regresión PLS y el análisis de componentes principales.

[2] *Hoskuldsson, A.* (1988): PLS regresión methods. Journal of chemometrics.

En este artículo se desarrolla la estructura matemática y estadistica de la regresión PLS. Es algo difícil de leer.

[3] *Tenenhaus, M.* (1998): La regresión PLS. Theorie et pratique, Paris, Editions Techip.

Este libro contiene una introducción a las técnicas que proporcionaron la regresión PLS como el análisis canónico utilizado por la regresión PLS, tanto para datos cuantitativos (PLS1 y PLS2) como para datos cualitativos.

[4] *R. Martínez Arias*: El análisis Multivariante en la investigación científica.

[5] *J. Etxeberria*: Regresión múltiple.

"Tengo mis resultados hace tiempo, pero no sé cómo llegar a ellos"

C. F. Gauss

Índice