Procedimiento de selección de factores de influencia en el análisis de regresión(•)

Method for selecting the influencing factors in regression analysis

J. Tošenovský(*)

En las aplicaciones del análisis de regresión ocurre, con frecuencia, que durante la elaboración del modelo, cuando se buscan los factores principales que influyen en el índice y, se incluyen en este prácticamente todos los factores x_1, x_2, ..., x_k que pueden influir, aunque sea en pequeña medida.

Al mantener todas esas variables puede suceder que el modelo resultante no sólo sea complicado, sino que, sobre todo, sea incorrecto.

Por esta razón, es deseable hacer una selección preliminar para obtener un modelo funcional que, al mismo tiempo, sea sencillo y correcto. En este artículo se trata de mostrar una posible selección.

ALGORITMO PARA LA SELECCIÓN DE LOS FACTORES SIGNIFICANTES

Con frecuencia, como modelo inicial se considera la función

$$y = \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + ... + \beta_k \cdot x_k \qquad [1]$$

Al decidir qué factores (variables) del conjunto $\{x_1, x_2, ..., x_k\}$ tienen que incluirse en el modelo [1], se puede proceder según los siguientes pasos:

a) Se forma el vector de las correlaciones entre y y cada una de las variables X_1, x_2, ..., x_k:

$$R_0^T = (r_{y,x_1}, r_{y,x_2}, ..., r_{y,x_k}) = (r_1, r_2, ..., r_k)$$

b) Se forma la matriz de las correlaciones entre los pares de las variables independientes x_1, ..., x_k:

$$R = \begin{matrix} r_{11} & r_{12} & \cdots & r_{1k} \\ r_{21} & r_{22} & \cdots & r_{2k} \\ \cdots & \cdots & \cdots & \cdots \\ r_{k1} & r_{k2} & \cdots & r_{kk} \end{matrix}$$

c) Se calcula el valor crítico del coeficiente de correlación r^*:

$$r^* = \frac{t_{n-2}(\alpha)}{\sqrt{t_{n-2}(\alpha) + n - 2}}$$

donde $t_{n-2}(\alpha)$ es el valor crítico de la distribución de Student y se busca en las tablas correspondientes; α es el nivel de significación.

Después de esto se procede a realizar la selección, que consta de dos etapas según la siguiente "filosofía": el coeficiente de correlación, r_i, que no es suficientemente grande, indica que y no depende notablemente de x_i y por eso se retira. El término "notablemente" significa mayor que r^*. De esta forma, se consigue otro punto del algoritmo:

d) Excluir esas x_i de [1] para las cuáles

$$|r_{y,x_i}| = |r_i| \le r^*$$

Entre el resto de r_i se busca el mayor coeficiente de correlación, por ejemplo, r_k, para el que vale

$$|r_k| = \max_i \{|r_i|\}$$

y se busca la correlación entre x_k y las demás variables.

Las variables que tienen con x_k una correlación mayor que r^* se retiran de [1], porque su influencia sobre y está originada por una gran correlación con el factor de influencia más significativa, x_k.

Por eso, el último paso del algoritmo es:

(•) Trabajo recibido el día 5 de junio de 1996.
(*) Universidad Técnica en Ostrava. República Checa.

(c) Consejo Superior de Investigaciones Científicas
Licencia Creative Commons 3.0 España (by-nc)
http://revistademetalurgia.revistas.csic.es

e) Retirar esas variables x_i de [1] para las cuales

$$r^* < |r_{k,i}|$$

Ejemplo

El uso del algoritmo citado se muestra en el siguiente problema, que trata sobre la relación entre la resistencia del coque y los factores elegidos.

Se supone, en primer lugar, que la resistencia del coque, y, depende de los siguientes factores:

$$y = f(x_1, x_2, ..., x_{10})$$

x_1 = ceniza en el coque
x_2 = resistencia del coque al desgaste por fricción en frío
x_3 = resistencia del coque en frío
x_4 = dilatación total
x_5 = dilatación del carbono
x_6 = contenido de la materia inflamable superficial en el carbón
x_7 = contenido de ceniza en el carbón
x_8 = temperatura última de coquefacción
x_9 = humedad del coque
x_{10} = reactividad del coque

a) Se construyen las matrices necesarias:

$$R_0^T = (0{,}38,\ -0{,}45,\ 0{,}46,\ 0{,}62,\ 0{,}61,\ -0{,}48,\ 0{,}50,\ -0{,}27,\ -0{,}28,\ -0{,}91)$$

b)

	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	
	1	0,28	0,32	−0,49	0,43	0,05	−0,70	−0,58	0,24	−0,39	x_1
		1	0,37	−0,08	0,07	−0,20	−0,24	−0,14	0,20	−0,31	x_2
			1	0	−0,03	0,47	0	0,08	0,09	0,03	x_3
R =				1	−0,49	0,06	0,48	0,81	−0,53	0,48	x_4
					1	0,06	−0,44	−0,77	0,53	−0,46	x_5
						1	0,30	0,20	−0,19	0,08	x_6
							1	0,65	−0,26	0,52	x_7
								1	−0,61	0,56	x_8
									1	−0,37	x_9
										1	x_{10}

c)

$$r^* = \frac{t_{n-2}(\alpha)}{\sqrt{t_{n-2}(\alpha) + n - 2}} =$$

$$= \frac{t_{30-2}(0{,}05)}{\sqrt{t_{30-2}(0{,}05) + 30 - 2}} = 0{,}374$$

d) Menores que r^* son en R_0 los coeficientes de correlación r_8 y r_9; por eso se retiran las variables x_8 y x_9.

La correlación más fuerte con y la tiene, en R_0, el factor x_{10}.

e) Las variables que tienen una correlación con y mayor que r^* son: x_1, x_4, x_5 y x_7.
Al retirar estas variables del modelo [1] quedan

$$y = f(x_2, x_3, x_6, x_{10})$$

la que se considera, según [1] en la forma

$$y = \beta_2 \cdot x_2 + \beta_3 \cdot x_3 + \beta_6 \cdot x_6 + \beta_{10} \cdot x_{10}$$

Ahora, la estimación β_2, β_3, β_6 y β_{10} se realizará sin problemas utilizando la fórmula bien conocida

$$\hat{\beta} = (X^T X)^{-1} \cdot X^T Y \qquad [2]$$

al contrario de la situación, en la que se consideran todos los factores.

El problema antes mencionado se basa en la búsqueda de la matriz inversa en [2]: si en la matriz X hay dos columnas correlacionadas, o no es posible encontrar la matriz inversa o existe, pero la estimación de los coeficientes tiene una dispersión enorme. Esto puede suceder en los casos en que en el modelo se incluyen todos los factores, sin que se realice la selección antes mencionada.

Es verdad que hay situaciones en las que en la matriz X hay columnas correlacionadas, que según las condiciones de los mínimos cuadrados ordinarios no son admisibles, pero no es posible prescindir de ellas, ya que cada una de ellas es indispensable para el modelo. En esos casos, en [2] se utiliza, en lugar de la matriz inversa $(X^T \cdot X)^{-1}$, otro procedimiento.

$$(X^T \cdot X)^{-1} = \sum_{i=1}^{k} \lambda_i^{-1} P_i P_i^T$$

donde λ_i = núm. característico de la matriz $X^T X$
P_i = vector característico de la matriz $X^T X$

si

$$\left| \frac{\lambda_i}{\sum_{i=1}^{\lambda_i} \lambda_i} \right| \leq p$$

entonces se retira λ_i
p = núm. bastante pequeño, por ejemplo $p = 10^{-15}$.

BIBLIOGRAFÍA

NOWAK, E. Zarys metod ekonometrii. Wydawnictvo naukove. Varsovia (Polonia), 1990: 19-22.

(c) Consejo Superior de Investigaciones Científicas
Licencia Creative Commons 3.0 España (by-nc)
http://revistademetalurgia.revistas.csic.es

DRAPER, N. y SMITH, H. Applied Regression Analysis. John Wiley and Sons. Nueva York (EE.UU.), 1981.

MELOUN, M. y MILITKY, J. Statistické zpracování experimentálních dat. Plus. Praga (R. Checa), 1994.

GUTTMAN, I. Linear models-An Introcution. John Wiley and Sons. Nueva York (EE.UU.), 1982.

MARQUART, D.M. *Technometrics*, 12, 1970: 591.

ATKINSON, A.C. Plot, Transformation, Regression. Claredon Press. Oxford (R.U.), 1986.

ROUSSEUW, P.J. y LEROY A.M. Robust Regression and Outliers Detection. John Wiley and Sons. Nueva York (EE.UU.), 1987.

NASZODI, L.J. *Technometrics*, 20, 1978: 201.

COOK, R.D. y Weisberg, S. *Biometrika*, 71, 1983: 1.

MALOWS, C.L. *Technometrics*, 28, 1986: 313.

JOINER, B. *Amer. Statist.*, 35, 1981: 227.

RICE, J.A. Mathematical Statistics and Data Analysis. Wadsworth and Brooks. California (EE.UU.), 1988.

KLEINBAUM, D.G. Applied Regression Analysis and Other Multivariate Methods. PSW-KENT Publishing Comp. Boston (EE.UU.), 1988.

HIMMELBLAU, D. Process Analysis by Statistical Methods. John Wiley & Sons. Nueva York (EE.UU.), 1969.

(c) Consejo Superior de Investigaciones Científicas
Licencia Creative Commons 3.0 España (by-nc)
http://revistademetalurgia.revistas.csic.es