Revista Investigación Educativa - Vol. 14, n.º 1, págs. 93-96

SOBRE UNA DESIGUALDAD QUE VERIFICA EL COEFICIENTE DE DETERMINACIÓN EN REGRESIÓN MÚLTIPLE

por
Pedro Sánchez Algarra[1]
Departamento de Estadística
Universidad de Barcelona

RESUMEN

Dada la gran relevancia de la regresión múltiple en ciencias de la educación y, en general, en ciencias sociales y del comportamiento, se demuestra a través de tres casos que en ciertas circunstancias el cuadrado de la correlación múltiple es mayor que la suma de los cuadrados de las respectivas correlaciones simples.

Descriptores: *Coeficiente de determinación, regresión múltiple.*

ABSTRACT

The multiple regression has a great relevance in educational research, and, in general, in social and behavioral sciences. In this paper it is demonstrated that the square of multiple correlation coefficient, thorough some circunstances, is greater than the sum of squares of simple correlations coefficients. It is presented in three cases.

Key words: *Determination coefficient, multiple regression.*

1 Dirección de contacto: Departamento de Estadística. Facultad de Biología. Diagonal, 645. 08028 Barcelona. Tel. (93)4021562.

En la regresión múltiple de una variable dependiente Y sobre dos variables predictoras X_1, X_2, la variabilidad explicada por la regresión simultánea de Y sobre X_1, X_2 puede ser mucho mayor que la variabilidad explicada por cada variable por separado. Esta interesante y en cierto modo sorprendente característica de la regresión múltiple ha sido estudiada por Hamilton (1987), y sus implicaciones en ciencias de la educación, psicológicas, médicas, sociales y económicas han sido discutidas por Bertrand y Holder (1988), quienes señalan la necesidad de tener en cuenta el efecto de todas las variables, las cuales deben ser tenidas en cuenta simultáneamente.

Como la variabilidad explicada viene medida por el coeficiente de determinación (correlación al cuadrado), se trata de verificar que en ciertas circunstancias

$$P^2>\rho^2_1+\rho^2_2 \tag{1}$$

donde P es la correlación múltiple de Y sobre X_1, X_2, y $\rho_i=\rho(Y,X_i)$, i=1,2, son las correlaciones simples. Por ejemplo, para los siguientes datos

X_1	X_2	Y
1	0	1
2	2	4
3	0	3
4	-2	2

se obtiene P=1, $\rho_1=0.2$, $\rho_2=\sqrt{0.4}$, por lo que

$$P^2=1>\rho^2_1+\rho^2_2=0.44$$

En este trabajo se pretende llevar a cabo una demostración general de la propiedad (1). Sea P la correlación múltiple de Y sobre $X_1,\ldots, x_n$. Indiquemos $r=r(Y,X_i)$, i=1,...,n, y sea R la matriz de correlaciones entre las variables X que suponemos no singular. También podemos suponer que todas las variables son centradas y de varianza 1. La mejor predicción de Y, en el sentido de los mínimos cuadrados, obtenida como combinación lineal de las variables X, viene dada por

$$\hat{Y}=\hat{\beta}_1X_1+\ldots+\hat{\beta}_nX_n$$

donde $\hat{\beta}=(\hat{\beta}_1,\ldots, \hat{\beta}_n)'$ verifica $\beta=R^{-1}\rho$, siendo $\rho=(\rho_1,\ldots, \rho_n)'$.

El coeficiente de determinación es entonces

$$P^2=\sum_{i=1}^{n}\hat{\beta}_i\rho_i=\hat{\beta}'\rho=\rho'R^{-1}\rho \tag{2}$$

Por otra parte, la suma de variabilidades explicadas por cada variable predictora por separado es

$$P^2_*=\sum_{i=1}^{n}\rho^2_i=\rho'\rho \tag{3}$$

Se trata de estudiar en qué condiciones (2) es mayor que (3). A tal fin analicemos el cociente P^2/P^2_*. Sean $\mu_1\leq...\leq\mu_n$ los valores propios de R_{-1}. Utilizando una conocida desigualdad (Magnus y Neudecker, 1988), se cumple que

$$\mu_i\leq \frac{\rho'R^{-1}\rho}{\rho'\rho} \leq\mu_n \tag{4}$$

Obsérvese que si $\lambda_1\geq...\geq\lambda_n$ son los valores propios de R, entonces como $Rv_i=\lambda_1 v_i\Rightarrow R^{-1}v_i=v_i/\lambda_i$, siendo v_i el correspondiente vector propio, resulta que $\mu_i=\lambda/\lambda_i$. Como además $\sum\lambda_i=traza\ (R)=n$, para algún entero k se verificará

$$\lambda_1\geq...\geq\lambda_k\geq1\geq\lambda_{k+1}\geq...\geq\lambda_n$$

y por lo tanto

$$\mu_1\leq...\leq\mu_k\leq1\leq...\leq\mu_n$$

Veamos los tres casos siguientes:

Caso 1. Supongamos R=I. Entonces $\mu_1=...=\mu_n=1$, por lo que

$$P^2=P^2_*$$

y la variabilidad de Y explicada por la regresión múltiple sobre las variables X es igual a la suma de variabilidades explicadas por cada una de las variables X.

Caso 2. Sea v_i un vector propio de R de valor propio λ_i tal que $v'_iv_i=1$. Entonces v_i define una componente principal Y_i que es combinación lineal de las variables X. Tomemos $Y=Y_i$. Como $\sigma^2(Y_i)=v'_iRv_i=\lambda_i$ tenemos que $\rho=Rv_i/(\lambda_i)^{1/2}=(\lambda_i)^{1/2}v_i$.

Además es obvio que $P^2=1$, luego

$$P^2=1>P^2_*=\rho'\rho=\lambda_i \qquad para \qquad i\geq k$$

Caso 3. Sea ρ un vector «próximo» a un vector propio v_i. Como $v'_iR^{-1}v_i/v'_iv_i=\mu_i$, por razones de continuidad podemos afirmar que

$$\mu_i< \frac{\rho'R^{-1}\rho}{\rho'\rho} <\mu_{i+1} \qquad para \qquad i\geq k \tag{5}$$

Así, si tomamos por ejemplo $Y=Y_i+e$, donde e es un término de error intercorrelacionado con las variables X, el vector de correlaciones r será «próximo» a v_i, y como consecuencia de (4) resultará que

$$1>P^2=\rho' R^{-1}\rho>\mu_i\rho'\rho<P^2_* \qquad (6)$$

puesto que $\mu_i>1$, siempre y cuando $R\neq I$. Nótese que la desigualdad (6) se invierte si tomamos μ «próximo» a v_i, $i<k$.

REFERENCIAS

BERTRAND, P.V. & HOLDER, R.L. (1988). A quirk in multiple regression: The whole regression can be greater than the sum of its parts. *The Statistician, 37,* 371-374.

HAMILTON, D. (1987). Sometimes $R^2>r^2_{yx1} + r^2_{yx2}$, correlated variables are not always redundant. *American Statistician, 41,* 129-132.

MAGNUS, J.R. & NEUDECKER, H. (1988). *Matrix Differential Calculus.* New York: Wiley.