University of the Philippines
SCHOOL OF ECONOMICS

Discussion Paper 8101 February 1981

RELATIVE CONTRIBUTIONS OF MIXED VARIABLES TO THE VARIATION OF A REGRESSAND

by

José Encarnación, Jr.

Abstract

This note shows the direct comparability of the beta coefficients of ordinary scalar variables and of classificatory vector variables. Accordingly, even when both kinds of variables appear in a regression equation, their relative contributions to the variation of the regressand can be ranked by their beta squares.

Relative Contributions of Mixed Variables to the Variation of a Regressand

J. Encarnación

Consider a regression equation whose regressions include classificatory as well as ordinary scalar variables. A classificatory variable is essentially a vector that has as many components as there are different (mutually exclusive and exhaustive) categories in the classification. For example, one might estimate a regression equation that explains employees' salaries in terms of length of service (a scalar), occupation (a classificatory variable), etc. One might then want to estimate the relative contributions of the explanatory variables to the variation of the dependent variable. Handling this problem by beta coefficients is well known when the explanatory variables are all of one kind, either all scalar or all classificatory. There seems, however, to be no convenient reference that discusses this matter when the explanatory variables are mixed, i.e. when they include both kinds. This expository note might therefore be of some use.

I

Let $x = (x_0, x_1, \ldots, x_K)$ where $x_k = 1$ for an individual (or observation) if it belongs to category k $(k = 0, 1, \ldots, K)$ of classification x, $x_k = 0$ otherwise, and $\sum_{k=0}^{K} x_k = 1$. More precisely, for any given individual i, $x_{ki} = 1$ if i is in category k, 0 otherwise, and $\sum_{k=0}^{K} x_{ki} = 1$. To each i thus corresponds $x_i =$ $(x_{0i}, x_{1i}, \ldots, x_{Ki})$.

Suppose it is appropriate to explain y in terms of x, z, u and v by means of a regression equation, where z is another classificatory variable $(z_0, z_1, \ldots, z_J)$ while u and v are real variables. (Discussion of more than two variables of either kind would be straightforward.) We calculate

$$(1) \qquad y' = c + \sum_{1}^{K} a_k^* x_k + \sum_{1}^{J} b_j^* z_j + p(u - \bar{u}) + q(v - \bar{v})$$

where the a_k^*, b_j^*, p and q are the regression coefficients and y' is the predicted y. As usual, overbars denote means. Note that x_0 and z_0 are omitted in (1) in order to have determinate coefficients (Suits 1957).

We want to express (1) in the form

$$(2) \qquad y' = \bar{y} + \sum_{0}^{K} a_k x_k + \sum_{0}^{J} b_j z_j + p(u - \bar{u}) + q(v - \bar{v})$$

where x_0 and z_0 are included, and the a_k and b_j measure the effects on an individual's y resulting from its belonging to k of x and to j of z, respectively. It is to be noted that the a_k and b_j, which might be called category effects (Encarnación 1975), are measured from $\bar{y}$. For suppose that for an individual i, $x_{ki} = 1$ for a particular k and $z_{ji} = 1$ for a particular j. Then

$$y_i' = \bar{y} + a_k + b_j + p(u_i - \bar{u}) + q(v_i - \bar{v})$$

so that a_k and b_j are simply added on to $\bar{y}$.

From least squares properties, using (1),

$$(3)\qquad c = \bar{y} - \sum_{1}^{K} a_k^* \bar{x}_k - \sum_{1}^{J} b_j^* \bar{z}_j - p(\bar{u} - \bar{u}) - q(\bar{v} - \bar{v})$$

$$= \bar{y} - \sum_{1}^{K} a_k^* \bar{x}_k - \sum_{1}^{J} b_j^* \bar{z}_j .$$

But c is also the predicted y for an individual satisfying $x_0 = 1$, $z_0 = 1$, $u = \bar{u}$ and $v = \bar{v}$. Therefore

$$(4)\qquad a_0 = - \sum_{1}^{K} a_k^* \bar{x}_k$$

$$(5)\qquad b_0 = - \sum_{1}^{J} b_j^* \bar{z}_j .$$

Further, if an individual satisfies $x_k = 1$ $(k \neq 0)$, $z_0 = 1$, $u = \bar{u}$, $v = \bar{v}$, the predicted y is $c + a_k^*$. Since we already know from (3)-(5) that

$$(6)\qquad c = \bar{y} + a_0 + b_0$$

we have $c + a_k^* = \bar{y} + (a_0 + a_k^*) + b_0$ so that

$$(7)\qquad a_k = a_0 + a_k^* \qquad\qquad k = 1, \ldots, K.$$

The b_j are similarly determined.

Substituting (6) in (1),

$$(8)\qquad y' = \bar{y} + a_0 + b_0 + \sum_{1}^{K} a_k^* x_k + \sum_{1}^{J} b_j^* z_j + p(u - \bar{u}) + q(v - \bar{v})$$

$$= \bar{y} + a_0 + b_0 + \sum_{1}^{K} (a_k - a_0)\, x_k + \sum_{1}^{J} (b_j - b_0)\, z_j + p(u - \bar{u})$$

$$+ q(v - \bar{v})$$

$$= \bar{y} + a_0 (1 - \sum_1^K x_k) + \sum_1^K a_k x_k + b_0 (1 - \sum_1^J z_j) + \sum_1^J b_j z_j$$

$$+ p(u - \bar{u}) + q(v - \bar{v}).$$

But $1 - \sum_1^K x_k = x_0$ and $1 - \sum_1^J z_j = z_0$; hence (2).

We note for later reference that $\bar{x}_k = n_{k.}/n$, where $n_{k.}$ is the number of individuals for which $x_{ki} = 1$ and n is the total number of individuals. Also, as one might expect,

$$(9) \quad \sum_{h=1}^{n} \sum_{k=0}^{K} a_k x_{kh}/n = \sum_0^K a_k n_{k.}/n = \sum_0^K a_k \bar{x}_k = 0$$

i.e. the mean $\sum_0^K a_k x_k = 0$ (in the same way that the mean $p(u - \bar{u})$, say, is zero). For, multiplying (7) by $n_{k.}$, summing both sides and then adding $n_{0.}\, a_0$ to the results,

$$\sum_0^K n_{k.}\, a_k = n\, a_0 + \sum_1^K n_{k.}\, a_k^*$$

which, in view of (4), gives (9).

II

The motivation for calculating the partial beta coefficients of standard multiple regression is to be able to compare the relative contributions of the explanatory (scalar) variables to the variation of the dependent variable (see, e.g., Ezekiel and Fox 1959, p. 196). Accordingly, the variables are standardized to zero means and unit variances, so that their beta coefficients become directly comparable. Similarly, the beta coefficients discussed by Morgan _et al._ (1962) perform the same function in the case of classificatory variables. Our problem is to see whether all the beta coefficients in a regression with mixed variables are directly comparable.

Write

$$(10)\qquad \frac{y' - \bar{y}}{s_y} = \beta_x f(x) + \beta_z g(z) + \beta_u \frac{u - \bar{u}}{s_u} + \beta_v \frac{v - \bar{v}}{s_v}$$

which is to be equivalent to (cf. (2))

$$(11)\qquad \frac{y' - \bar{y}}{s_y} = \frac{\sum_0^K a_k x_k}{s_y} + \frac{\sum_0^J b_j z_j}{s_y} + \frac{p(u - \bar{u})}{s_y} + \frac{q(v - \bar{v})}{s_y}$$

where s_y is the standard direction of y, etc.,

$$(12)\qquad \beta_u = p\, s_u / s_y$$

which is the textbook definition of a partial beta coefficient, similarly for β_v,

$$(13)\qquad \beta_x = \frac{\left(\sum_0^K a_k^2 n_{k.} / (n - 1)\right)^{1/2}}{s_y}$$

from Morgan et al. (1962), and the functions $f(x)$ and $g(z)$ are implicitly defined by the equivalence of (10) and (11) and the definitions of the β's. It is clear that if $\beta_u^2 > \beta_v^2$, u contributes more than does v to the explanation of y variation. Our object is to show that $f(x)$, say, standardizes x essentially in the same way that $(u - \bar{u})/s_u$ standardizes u, so that all the beta coefficients are then directly comparable.

From (10), (11) and (13), for individual i,

$$(14)\qquad f(x_i) = \frac{\sum_{k=0}^{K} a_k x_{ki}}{(\sum_{k=0}^{K} a_k^2 n_{k.}/(n-1))^{1/2}}$$

from which

$$(15)\qquad f(x_i)^2 = \frac{\sum_{k=0}^{K} a_k^2 x_{ki}^2}{\sum_{h=1}^{n} \sum_{k=0}^{K} a_k^2 x_{kh}^2/(n-1)}$$

since cross-product terms vanish and $x_{ki} = x_{ki}^2$ (because $x_{ki} = 0$ or 1 and $\sum_{k=0}^{K} x_{ki} = 1$). But

$$(16)\qquad \frac{(u_i - \bar{u})^2}{s_u^2} = \frac{p^2(u_i - \bar{u})^2}{\sum_{h=1}^{n} p^2(u_h - \bar{u})^2/(n-1)}$$

corresponds precisely to (15), the only difference being that while one can factor out p^2 in (16), which of course does not affect the ratio, it is not possible to factor out $\sum_0^K a_k^2$ in (15), which pertains to a vector. The key observation is that x being a classificatory variable, $\sum_{k=0}^{K} a_k x_{ki}$ is the analogue of $p(u_i - \bar{u})$ and both have zero means.

This completes our task, and all the beta squares may then be ranked to indicate the relative contributions of their corresponding variables to the explanation of y variation.

References

Encarnación, J. "Income Distribution in the Philippines: The Employed and the Self-Employed," in _Income Distribution, Employment and Economic Development in Southeast and East Asia_ (Tokyo: JERC and Manila: CAMS), 1975, pp. 742-775.

Ezekiel, M. and Fox, K.A. _Methods of Correlation and Regression Analysis_, 3rd ed. (New York: Wiley), 1959.

Morgan, J.N. _et al._ _Income and Welfare in the United States_ (New York: McGraw-Hill), 1962. Appendix E.

Suits, D.B. "Use of Dummy Variables in Regression Equations," _Journal of the American Statistical Association_, December 1957, pp. 548-551.