Institute of Economic Development and Research

SCHOOL OF ECONOMICS
University of the Philippines System

Discussion Paper No. 74-3

April 19, 1974

ON MEASURING THE RELATIVE CONTRIBUTIONS OF SEVERAL CLASSIFICATORY VARIABLES IN THE EXPLANATION OF A DEPENDENT VARIABLE

by

José Encarnación, Jr.

Note: IEDR Discussion Papers are preliminary versions circulated privately to elicit critical comment. References in publications to Discussion Papers should be cleared with the author.

ON MEASURING THE RELATIVE CONTRIBUTIONS OF SEVERAL CLASSIFICATORY VARIABLES IN THE EXPLANATION OF A DEPENDENT VARIABLE

by J. Encarnación

Consider the explanation of a variable y (the logarithm of income for example) in terms of two classificatory variables A and B. A could be, say, an industrial sector classification with different categories such as manufacturing, agriculture, etc. Each classification consists of categories which are exhaustive and mutually exclusive, and we wish to explain an individual's y as a result of his belonging to some category of A and some category of B. We also want to estimate the relative contributions of those variables to the explanation of y variation. The purpose of this note is to make more explicit the treatment of these matters by Suits (1957), Morgan et al. (1962) and Oey (1972). It will be apparent how the discussion would proceed if there were three or more classificatory variables.

Let the categories of A be indexed by k ($k = 0, 1, \ldots, K$), those of B by j ($j = 0, 1, \ldots, J$); and let y_{kji} be the ith observation in the (k, j)-cell, $i = 1, \ldots, n_{kj}$. There are n observations, so

$$n = \sum_{k=0}^{K} \sum_{j=0}^{J} n_{kj} .$$

Write

$$n_{k.} = \sum_{j=0}^{J} n_{kj}$$

for the number of observations in the kth category of A; similarly,

$$n_{.j} = \sum_{k=0}^{K} n_{kj}$$

The mean y in category k of A is given by

$$\bar{y}_{k.} = \sum_{j=0}^{J} \sum_{i=1}^{n_{kj}} y_{kji}/n_{k.}$$

and for the mean y we write $\bar{y}$.

Let

$$x_k = \begin{cases} 1 & \text{if an observation belongs to } k \text{ of } A \\ 0 & \text{otherwise} \end{cases}$$

and define z_j similarly. With these dummy variables we can calculate a regression equation

$$(1) \quad y' = c + a'_1 x_1 + \ldots + a'_K x_K + b'_1 z_1 + \ldots + b'_J z_J$$

Note that the variables x_0 and z_0 are omitted from the regression in order to get determinate estimates of the coefficients; see Suits (1957). (Cf. the fact that a sex classification has two categories - male and female - but only one dummy variable would be included in a regression equation, i.e. one dummy variable is omitted.) For a given individual, at most one of the x_k, $k = 1, \ldots, K$, and at most one of the z_j, $j = 1, \ldots, J$, can be nonzero. Eq. (1) thus shows y', the predicted y, to depend on an individual's membership in the different categories of A and of B.

We wish to write (1) in the form

$$y' = \bar{y} + a_k + b_j \tag{2}$$

where $k = 0, 1, \ldots, K$ and $j = 0, 1, \ldots, J$. Then a_k would measure the effect on y resulting from belonging to category k of A, which effect is measured from $\bar{y}$. Similarly for b_j in regard to j of B.

In order to get the a_k, from least squares regression properties we know that in (1),

$$c = \bar{y} - \sum_1^K a_k' \bar{x}_k - \sum_1^J b_j' \bar{z}_j$$

$$= \bar{y} - \sum_1^K a_k' n_{k.}/n - \sum_1^J b_j' n_{.j}/n$$

But c is the predicted y when one belongs to category 0 of A (in which case $x_1 = \ldots = x_K = 0$) and 0 of B. From this it follows that

$$a_0 = -\sum_1^K a_k' n_{k.}/n \tag{3}$$

$$b_0 = -\sum_1^J b_j' n_{.j}/n$$

On the other hand, if one belongs to category k ($k = 1, \ldots, K$) of A and 0 of B, the predicted y is $c + a_k'$. Accordingly we have

$$a_k = a_0 + a_k' \qquad (k = 1, \ldots, K) \tag{4}$$

Eqs. (3) and (4) thus determine the a_k, and a similar procedure gives the b_j.

Multiplying (4) by $n_{k.}$, summing both sides and then adding $n_{0.}a_0$ to the results, one gets

$$\sum_0^K n_{k.}a_k = na_0 + \sum_1^K n_{k.}a'_k$$

which, in view of (3), implies that the weighted sum of the a_k is zero:

$$\sum_0^K a_k n_{k.}/n = 0 \tag{5}$$

In order now to get (2) from (1), write

$$\begin{aligned} y' &= \bar{y} + a_0 + b_0 + \sum_1^K a'_k x_k + \sum_1^J b'_j z_j \\ &= \bar{y} + a_0 + b_0 + \sum_1^K (a_k - a_0)x_k + \sum_1^J (b_j - b_0)z_j \\ &= \bar{y} + a_0(1 - \sum_1^K x_k) + \sum_1^K a_k x_k + b_0(1 - \sum_1^J z_j) + \sum_1^J b_j z_j \end{aligned}$$

But $1 - \sum_1^K x_k = x_0$ and similarly for the other set of dummy variables, and we obtain (2). Hence (1) and (2) yield identical predicted values (and therefore identical error terms).

Define a new variable

$$y^*_{kji} = y_{kji} - a_k - b_j$$

by removing from each observation the effects of the classificatory variables as calculated. Then

$$y^* - \bar{y} = y - y'$$

becomes a measure of the unexplained error.

The beta coefficients discussed by Morgan et al. serve to measure the relative contributions of the classificatory variables to the explanation of y variation:

$$\beta_A = \frac{\left(\sum_{k=0}^{K} n_{k.} a_k^2/(n-1)\right)^{1/2}}{s_y} \tag{6}$$

where s_y is the standard deviation of y. Similarly

$$\beta_B = \frac{\left(\sum_{j=0}^{J} n_{.j} b_j^2/(n-1)\right)^{1/2}}{s_y}$$

If, say, $\beta_A > \beta_B$, that would indicate that more of the variation in y is due to the classificatory variable A. As Morgan et al. point out, these coefficients are analogous to the partial beta coefficients of standard multiple regression.

Following Oey (1972), an approximate F-test could be used to test the significance of a classificatory variable in explaining y. For from (2),

$$e_{kji} = y_{kji} - \bar{y} - a_k - b_j = y^*_{kji} - \bar{y} \tag{7}$$

where e is the error term, and under certain assumptions, $(n - 1)s_e^2/\sigma_y^2$ would be a chi-square variable with $n - K - J - 1$ degrees of freedom. Accordingly in the case of A,

$$F_A = \frac{\sum_{k=0}^{K} n_{k.} a_k^2/K}{\sum_{k=0}^{K} \sum_{j=0}^{J} \sum_{i=1}^{n_{kj}} (y^*_{kji} - \bar{y})^2/(n - K - J - 1)} \tag{8}$$

If the sum of squares in the numerator is small relative to that of the denominator, that would indicate that the use of A does not add much to the explanation of y.

Finally, we consider the interesting (but false) conjecture that

$$R^2 = \beta_A^2 + \beta_B^2$$

where $R^2 = 1 - s_e^2/s_y^2$ is the coefficient of determination given by (1). One might argue in the following way. We have

$$y_{kji} - \bar{y} = a_k + b_j + (\bar{y}_{kj} - \bar{y} - a_k - b_j) + (y_{kji} - \bar{y}_{kj}) \tag{9}$$

where $\bar{y}_{kj}$ is the mean y in the (k, j)-cell, so that referring to (7),

(10) $$e_{kji} = (\bar{y}_{kj} - \bar{y} - a_k - b_j) + (y_{kji} - \bar{y}_{kj})$$

i.e. the error term is the sum of two differences: between the cell mean and the predicted value given by (1) or (2), and between the actual y and the cell mean. From (9),

(11) $$\sum_{kji}\sum\sum (y - \bar{y})^2 = \sum\sum\sum [a_k + b_j + (\bar{y}_{kj} - \bar{y} - a_k - b_j) + (y_{kji} - \bar{y}_{kj})]^2$$

Suppose cross-product terms on the right-hand side all vanish. Then

$$\sum_{kji}\sum\sum (y - \bar{y})^2 = \sum\sum\sum a_k^2 + \sum\sum\sum b_j^2 + \sum\sum\sum (\bar{y}_{kj} - \bar{y} - a_k - b_j)^2 + \sum\sum\sum (y - \bar{y}_{kj})^2$$

in which case

(*) $$\frac{(n-1)s_y^2}{\sigma_y^2} = \frac{(n-1)s_a^2}{\sigma_y^2} + \frac{(n-1)s_b^2}{\sigma_y^2} + \frac{(n-1)s_p^2}{\sigma_y^2} + \frac{(n-1)s_q^2}{\sigma_y^2}$$

where $s_a^2 = \sum_k n_{k.} a_k^2/(n-1)$, $s_p^2 = \sum\sum\sum (\bar{y}_{kj} - \bar{y} - a_k - b_j)^2/(n-1)$, etc. Assuming that y is normally distributed, it is known that $(n-1)s_y^2/\sigma_y^2$ is a chi-square variable with $n-1$ degrees of freedom, so that the terms on the right-hand side of (*) must then be independent chi-square variables whose degrees of freedom add up to $n-1$. That is,

(**) $$\chi^2_{y,n-1} = \chi^2_{a,K} + \chi^2_{b,J} + \chi^2_{p,KJ} + \chi^2_{q,n-KJ-K-J-1}$$

One could then get the F-statistic (8) by adding, in view of (10), the degrees of freedom in the last two terms of (**). In other words, e would also be chi-square with $n - K - J - 1$ degrees of freedom, and

(***) $$s_e^2 = s_p^2 + s_q^2$$

Then from (*) and (***) one has

$$1 = \frac{s_a^2}{s_y^2} + \frac{s_b^2}{s_y^2} + \frac{s_e^2}{s_y^2}$$

$$= \beta_A^2 + \beta_B^2 + \frac{s_e^2}{s_y^2}$$

which says that R^2 is the sum of the β^2s. This statement is wrong, however, as the cross-product terms on the right-hand side of (11) do not all vanish. For instance,

$$\sum\sum\sum a_k (\bar{y}_{kj} - \bar{y} - a_k - b_j) = \sum_k a_k n_{k.} \bar{y}_{k.} - \bar{y} \sum_k n_{k.} a_k$$

$$- \sum_k n_{k.} a_k^2 - \sum\sum\sum a_k b_j$$

The second and fourth terms on the right-hand side are easily seen to vanish, but

$$\sum_k a_k n_{k.} \bar{y}_{k.} - \sum_k n_{k.} a_k^2 = \sum_k a_k n_{k.} (\bar{y}_{k.} - a_k)$$

is zero only if $\bar{y}_{k.} - a_k$ is a constant, as would be the case if $a_k = \bar{y}_{k.} - \bar{y}$. The last clearly does not hold in general, however. If it did, the calculated a_k would not change as one adds more classificatory variables in the explanation of y, but in fact the a_k -which derive from the regression coefficients of (1)- would change as we add more variables in the specification.

The author is indebted to Professor Lydia H. Flores, of the University of the Philippines Department of Mathematics, for very helpful discussions on the subject of this note.

References

Hogg, R.V. and Craig, A.T. Introduction to Mathematical Statistics, 2nd ed., Macmillan, 1965.

Morgan, J.N. et al. Income and Welfare in the United States, McGraw-Hill, 1962. Appendix E.

Oey, M. "Income Distribution in Brazil," unpublished Ph.D. dissertation, University of California at Berkeley, 1972.

Scheffé, H. The Analysis of Variance, Wiley, 1959.

Suits, D.B. "Use of Dummy Variables in Regression Equations," J. Amer. Stat. Assn., December 1957, pp. 548-51.