ON RELATIVE CONTRIBUTIONS OF MIXED EXPLANATORY VARIABLES TO THE VARIATION OF A REGRESSAND[1]

by

Tito A. Mijares, Ph. D., Academician

We have a general linear model in matric form

$$(1) \qquad \underline{Y} = \underline{X}\,\underline{\beta} + \underline{\mu}$$

where $\underline{Y}' \doteq (Y_1, \ldots, Y_n)$; $\underline{X} = (X_{ij})$, $i = 1, \ldots, n$, $j = 0, \ldots, k$ with the first column of $X's$ each equal to unity: $\underline{\beta}' = (\beta_0, \beta_1, \ldots, \beta_k)$ and $\underline{\mu}' = (\mu_1, \ldots, \mu_n)$. $\underline{\beta}$ is a vector of unknown parameters and $\underline{\mu}$ is a vector of random values. The usual assumptions are: (a) the expected value $E(\underline{\mu}) = 0$, (b) $E(\underline{\mu}\,\underline{\mu}') = \sigma_\mu^2 I_n$, where I_n is a unit matrix of order n and $\sigma_\mu^2 < \infty$ is the common variance of the $\mu's$, (c) $\underline{X}$ is a set of fixed real numbers with rank $k < n$. The vector of parameters $\underline{\beta}$ are to be estimated, usually by least squares.

Without loss of generality the model may be restated by expressing the dependent vector $\underline{Y}$ and the explanatory variables X_{ij} as deviates from their respective means and eliminating β_0. Thus equation (1) may be written

$$(2) \qquad \underline{Y} = \underline{x}\,\underline{\beta} + \underline{\epsilon}$$

where $\underline{y}' = (y_1, \ldots, y_n)$, $y_i = Y_i - \bar{Y}$, $\bar{Y} = \sum_{i=1}^{n} Y_i/n$

[1]Comments on "Relative Contributions of Mixed Explanatory Variables to the Variation of a Regressand", by J. Encarnacion, presented at the Third Annual Scientific Meeting of the Natural Academy of Science and Technology, July 9, 1981, PICC, Manila

$$\underline{x} = (x_{ij}),\ i = 1, \ldots, n,\ j = 1, \ldots, k,$$

$$x_{ij} = X_{ij} - \overline{X}_j,\ \overline{X}_j = \sum_{i=1}^{n} X_{ij}/n$$

$$\underline{\beta}' = (\beta_1, \ldots, \beta_k) \text{ and } \underline{\epsilon}' = (\epsilon_1, \ldots, \epsilon_n)$$

If $\hat{\underline{\beta}}' = (\hat{\beta}_1, \ldots, \hat{\beta}_k)$ is the vector of least squares estimates of $\underline{\beta}$ equation (2) may be written equivalently as

$$(3) \qquad \underline{y} = \underline{x}\,\hat{\underline{\beta}} + \underline{e}$$

where e is a vector of n residuals $y - \underline{x}\,\hat{\underline{\beta}}$. It can be established that $\hat{\underline{\beta}} = (\underline{x}^T \underline{x})^{-1}\ \underline{x}'\ \underline{y}$. The mean and variance of $\hat{\underline{\beta}}$ are respectively $\underline{\beta}$ and $\sigma_\epsilon^2\ (\underline{x}'\ \underline{x})^{-1}$. Equation (3) may be expressed by

$$(4) \qquad \underline{y} = \hat{\underline{y}} + e$$

where

$$(5) \qquad \hat{\underline{y}} = \underline{x}\,\hat{\underline{\beta}}$$

In terms of Dr. Encarnacion's formulation (cf. eq. (1) $\hat{\underline{y}}$ is the "predictor" of $\underline{y}$ Thus, the vector $\underline{y}$ consists of the vector of *explained* and *unexplained* parts, $\underline{e}$ being the latter portion. The total number of regression coefficients in his paper is $K + J + 4$ which is equal to dimension k in this note, if his p and q are denoted as $\hat{\beta}_{k-1}$ and $\hat{\beta}_k$, respectively. For a given element $\hat{y}$ of $\hat{y}$ in this note

$$\hat{y} = \overline{y} + a_0 + b_0$$

of that paper (cf. eg. (2), Encarnacion's paper). The coefficients $\hat{\beta}_1, \ldots, \hat{\beta}_{k-2}$ here are the same as the coefficients of the discrete explanatory variables in that same paper.

Dummy Variables

We may now view the problem addressed by Dr. Encarnacion as extension of a general linear model in certain aspects. In econometric work the introduction of discrete variables is generally meant the inclusion of "dummy" variables in the usual regression model. Suppose Y is income expressed by gross national product (GNP) and X is total investment. A linear model for two periods may be expressed

$$Y = \alpha_1 + \beta X + \epsilon \text{ (before the war)}$$

$$Y = \alpha_2 + \beta X + \epsilon \text{ (after the war)}$$

The two equations may be combined into a single equation

$$y = \alpha_0 + \beta_0 Z + \beta X + \epsilon \tag{6}$$

where $Z = 0$ *before* the war and $Z = 1$ *after* the war. Hence,

$$E(Y \mid Z = 0) = \alpha_0 + \beta X$$

$$E(Y \mid Z = 1) = (\alpha_0 + \beta_0) + \beta X$$

Note that α_1 is now equivalent to α_0 and $\alpha_2 = \alpha_0 + \beta_0$ (cf. lines 5 and 6 from the bottom, p. 2, Encarnacion's paper).Hence, we may treat the problem as an ordinary linear regression problem, unrestricted case in the sense that no restrictions are imposed on the coefficients.

Tests on the Coefficients

To make tests on the coefficients an additional assumption on the distribution of the residual term ϵ_i, $i = 1, \ldots, n$ in equation (2) is needed. Suppose the ϵ_i's are independently and identically normally distributed random variables with zero means and common variance σ_ϵ^2. The L. S. estimate of $\underline{\beta}$ is

$$\hat{\underline{\beta}} = (\underline{x}'\underline{x})^{-1} \underline{x}'\underline{y} \tag{7}$$

$$= \underline{\beta} + (\underline{x}'\underline{x})^{-1} \underline{x}'\underline{\epsilon}$$

Then

$$E(\underline{\hat{\beta}}) = \underline{\beta} \tag{8}$$

$$\begin{aligned} \text{var}(\underline{\hat{\beta}}) &= E[(\underline{\hat{\beta}} - \underline{\beta})(\underline{\hat{\beta}} - \underline{\beta})'] \\ &= E[(\underline{x}'\underline{x})^{-1}\underline{x}'\underline{\epsilon}\,\underline{\epsilon}'\underline{x}(\underline{x}'\underline{x})^{-1}] \\ &= \sigma_\epsilon^2 (\underline{x}'\underline{x})^{-1} \end{aligned}$$

One sees from (7) that $\underline{\hat{\beta}}$ has a multinormal distribution over a k-dimensional space with density $N_k(\underline{\beta}, \sigma_\epsilon^2 (\underline{x}'\underline{x})^{-1})$. Hence, a linear function $\underline{c}'\underline{\hat{\beta}}$ has a univariate normal distribution with density $N(\underline{c}'\underline{\beta}, \sigma_\epsilon^2\, \underline{c}'(\underline{x}'\underline{x})^{-1}\underline{c})$. The statistic

$$t = \frac{\underline{c}'\underline{\hat{\beta}} - \underline{c}'\underline{\beta}}{s_e\sqrt{\underline{c}'(\underline{x}'\underline{x})^{-1}\underline{c}}}$$

will be distributed as *Student's-t* with n-k degrees of freedom. where $s_e = \sqrt{\underline{e}'\underline{e}/(n-k)}$. $\underline{\hat{\beta}}$ and $\underline{e}$ are independently distributed.

We can now compare coefficients of classificatory variables (e.g. the coefficient of the ith income group of one region against the coefficient of the jth income group of another region.) By choosing $\underline{c}$ appropriate to our hypotheses on the β's, we can make the tests on the coefficients. Let $\underline{c}' = (0, \ldots, 0, 1, 0, \ldots, 0, -1, 0, \ldots, 0)$, the ith element is 1 and the jth element is -1 and zeros in other places. This is equivalent to testing $H_0: \beta_i - \beta_j = 0$ or $\beta_i = \beta_j$ against $H_1: \beta_i \neq \beta_j$. The probability is α that $|t| > t_{\alpha/2,\, n\text{-}k}$, where $t_{\alpha/2,\, n\text{-}k}$ is the tabulated value of t with n-k $d.f.$

Concluding Remarks

The formulation of the general linear model given in (1) includes an assumption that the domain of the explanatory variables are real numbers and results derived therefrom apply also to the mixed case which Dr. Encarnacion deals with in his paper.

Apart from the problem that units of measures in the variables are not easily interpretable when compared, working with correlations among variables are of frequent interest because the square of multiple correlation coefficient

$$(10) \qquad R^2_{0\,.\,1\,2,\ldots k} = 1 - \frac{\Sigma e^2}{\Sigma y^2}$$

explains directly the proportion of total variation in the dependent variable Y explained by variables $X_1, \ldots, X_k$. Occasionally also the available data we have on the problem are expressed in correlation coefficients. Alternatively, the β's in the linear regression model of equation (2) can be derived from correlations among the variables. We can compute the simple (zero-order) correlations between the variables $Y, X_1, \ldots, X_k$ and display them in matric form $\underline{R} = (r_{ij})$ where r_{oj} $(j = 1, \ldots, k)$ denotes the correlation between Y and X_j and $r_{ii} = 1$ $(i = 0, \ldots, k)$. Then the least squares regression $\hat{y} = \hat{\beta}_1 x_1 + \ldots + \hat{\beta}_k x_k$, where $y, x_1, \ldots, x_k$ are deviates of variables $Y, X_1, \ldots, X_k$ from their respective means would have coefficients

$$(11) \qquad \hat{\beta}_j = - \frac{s_0}{s_j} \; \frac{R_{oj}}{R_{oo}}$$

where R_{oj} and R_{oo} denote the co-factors of r_{oj} and r_{oo} in the matrix $\underline{R}$, respectively, and s_o and s_j are the respective standard deviations of Y and X_j. An alternative expression for the least squares regression is

$$(12) \qquad \frac{R_{00}}{s_0}\,\hat{y} + \frac{R_{01}}{s_1}\,x_1 + \frac{R_{02}}{s_2}\,x_2 + \ldots + \frac{R_{0k}}{s_k}\,x_k = 0$$

The residual sum of squares $\Sigma e^2 = \underline{e}'\underline{e}$ may be expressed as

$$(13) \qquad \Sigma e^2 = \frac{n s_0^2 \, |\underline{R}|}{R_{00}}$$

where $|\underline{R}|$ is the determinant of matrix $\underline{R}$.

Since $\Sigma y^2 = ns_0^2$, equation (10) becomes

$$R^2_{0.12\ldots k} = 1 - \frac{|R|}{R_{00}} \tag{14}$$

The only thing left to relate equations (11) and (12) to Dr. Encarnacion's model is to determine the standard deviations and correlations of the discrete variables. Note that the classificatory variable x_j has mean p_j, the proportion of individuals in the j^{th} class. Its variance is $p_j\,(1 - p_j)$. The correlation between X_i and X_j in the same class is (c.f. Cramer, p. 318)

$$r_{ij} = \sqrt{\frac{p_i\, p_j}{(1 - p_i)\,(1 - p_j)}} \tag{15}$$

Take characteristic group h of classificatory variable X. Assume that the first ν of n individuals in the sample belong to h. Let the sequence of values of the continuous variable w in the h group be denoted by $w_1, \ldots, w_n$. The pairs of values of X and W and their deviates are

Original value

							Sums	Means
X:	1	1	. . . 1	0	 0		ν	$\nu/n = p$
W:	w_1	w_2 . . .	w_ν	$w_{\nu+1}$	W_n	$\sum_i^n w_i/_n$	=	$\bar{w}$

Deviates

x:	$1 - p$	$1 - p$	. . . $1 - p$	$-p$	. . . . $-p$
w:	$(w_1 - \bar{w})$	$(w_2 - \bar{w})$ . . .	$(w_\nu - \bar{w})$	$(w_{\nu+1} - w)$ . .	$(w_n - \bar{w})$

Then $\sum_1^n x_i w_i = \sum_1^\nu (1-p)(w_i - \overline{w}) - \sum_{\nu+1}^n p(w_i - \overline{w})$

$$= (1-p)\left[\sum^\nu w_i - \frac{\nu}{n}\sum^n w_i\right] - p\sum_{\nu+1}^n w_i + p(n-\nu)\overline{w}$$

$$= \sum^\nu w_i - p\sum^\nu w_i - p\sum^\nu w_i + p^2 \sum^n w_i$$

$$- p\left[\sum^n w_i - \sum^\nu w_i\right] + p(n-\nu)\overline{w}$$

This easily simplifies to

$$\sum_1^n x_i w_i = \sum_1^\nu w_i - p\left(\sum_1^n w_i\right) \tag{16}$$

since $\qquad p^2 \sum_1^n w_i = p\nu\overline{w}$ *and* $p\sum_1^n w_i = pn\overline{w}$

The simple correlation between x and w is

$$r_{xw} = \frac{\sum_1^\nu w_i - p\sum_1^n w_i}{\sqrt{p_q}\; s_w} \tag{17}$$

where $\qquad s_w = \sqrt{\sum_1^n (w_i - w)^2 / (n-1)}$ *and* $q = 1 - p$

Reference

H. Cramer: "Mathematical Methods of Statistics", Princeton University Press, Princeton, N. J., 1946