Relative Contributions of Mixed Variables to the Variation of a Regressand

By José Encarnación, Ph. D., Academician

Consider a regression equation whose regressors 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 $\Sigma_{k=0}^{K} x_k = 1$. More precisely, for any given individual i, $x_{ki} = 1$ if i is in category k, 0 otherwise, and $\Sigma_{k=0}^{K} x_{ki} = 1$. 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

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

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

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

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

$$y_i' = \bar{y} + a_k + b_j + p(\mu_i - \bar{\mu}) + q(v_i - \bar{v}).$$

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

From least squares properties, using (1),

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

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

But c is also the predicted y for an individual satisfying $x_0 = 1$, $z_0 = 1$, $\mu = \bar{\mu}$ 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 satisifies $x_k = 1 (k \neq 0)$, $z_0 = 1$, $\mu = \bar{\mu}$, $v = \bar{v}$, the predicted y is $c + a_k^*$. Since we already know from (3) — (5) that

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

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

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

The b_j are similarly determined.

Substituting (6) in (1),

(8)
$$y' = \bar{y} + a_0 + b_0 + \sum_1^K a_k^* x_k + \sum_1^J b_j^* z_j + p(\mu - \bar{\mu}) + q(v - \bar{v})$$

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

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

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

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

But $1 - \Sigma_1^K x_k = x_0$ and $1 - \Sigma_1^J z_j = z_0$; 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) $$\sum_{h=1}^{n} \sum_{K=0}^{K} a_k x_{kh}/n = \sum_0^K a_k n_{k.}/n = \sum_0^K a_k \bar{x}_k = 0.$$

i.e. the mean $\Sigma_0^K a_k x_k = 0.$ (in the same way that the mean $p(\mu - \bar{\mu})$, 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 directlv 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

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

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

$$(11) \quad \frac{y' - \bar{y}}{s_y} = \frac{\Sigma_0^K a_k x_k}{s_y} + \frac{\Sigma_0^J b_j z_j}{s_y} + \frac{p(\mu - \bar{\mu})}{s_y} + \frac{q(v - \bar{v})}{s_y}$$

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

$$(12) \quad \beta_u = p\ s_u/s_y$$

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

$$(13) \quad \beta_x = \frac{(\Sigma_0^K a_k^2 n_{k.}/(n-1))^{1/2}}{s_y}$$

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 $\beta^2_u > \beta^2_v$, 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 $(\mu - \bar{\mu})/s_\mu$ standardizes u, so that all the beta coefficients are then directly comparable.

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

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

from which

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

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

$$(16) \qquad \frac{(\mu_i - \bar{\mu})^2}{s_u^2} = \frac{p^2 \, (\mu_i - \bar{\mu})^2}{\Sigma_{h=1}^{n} \; p^2 \, (\mu_n - \bar{\mu})^2/(n-1)}$$

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 $\Sigma_0^K \, a_k^2$ in (15), which pertains to a vector. The key observation is that x being a classificatory variable, $\Sigma_{k=0}^{K} \, a_k \, x_{ki}$ is the analogue of $p(\mu_i - \bar{\mu})$ 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 variable to the explanation of y variation.

References

Encarnación, J (1975), "Income Distribution in the Philippines: The Employed and the Self-Employed," in *Income Distribution, Employment and Economic Development in Southeast and East Asia*, Tokyo: Japan Economic Research Center, 742-775

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

Morgan, J. N. *et al.* (1962), *Income and Welfare in the United States*, New York: McGraw-Hill. Appendix E.

Suits, D.B. (1957), "Use of Dummy Variables in Regression Equations," *Journal of the American Statistical Association*, 52, 548 - 551.

RELATIVE CONTRIBUTIONS OF MIXED VARIABLE TO THE VARIATION OF A REGRESSAND

Cristina A. Parel, Ph.D.

Discussant

1. The use of dummy variables in regression equations has not always been regarded favorably by some statisticians. But in application, "dummy" variables are getting to be indispensable because of the nature of some factors. These factors may have only two or more mutually exclusive levels in which case one cannot set up a continuous scale for the variables. However, the inclusion of dummy variables renders the resulting normal equations "unsolvable" in view of the singularity of the matrix of coefficients. To remedy the situation; that is, to be able to estimate the regression coefficients, some additional linear constraints involving the coefficients of the "dummy" variables need to be introduced. For example, if there are r sets of "dummy" variables (or, classifications) used in the regression equation, there would be r constraints needed to have the regression coefficients estimable. Two alternative methods are commonly used: (1) the sum of the coefficients of the "dummy" variables is equated to zero; and (2) one specified coefficient of each set of "dummy" variables is equated to zero. Dr. Encarnación used the second method. Using either of these methods, however, the resulting normal equations (obtained by the least squares method) can be solved directly with the use of an electronic computer because after using the constraints, the matrix of coefficients of the reduced normal equations will no longer be singular.

2. To determine the relative importance of the independent variables on the dependent (or, response) variable, any of the following three measures may be used.

i) the *partial correlation coefficient*, $r_{yj\,.\,kl\,..}$, given by:

$$r_{yj\,.\,kl..} = b_j \frac{S_j}{S_y} \frac{\sqrt{1-R^2_{jj}}}{\sqrt{1-R^2_{y\hat{y}}}}$$

where b_j = the regression coefficient corresponding to the independent variable x_j ;

$$R^2_{j\hat{j}} = 1 - \frac{\Sigma(X_j - \hat{X}_j)^2}{\Sigma(X_j - \bar{X})^2} \quad ;$$

where

$\hat{X}_j$ = the regressed valued of the independent variable X_j on the remaining independent variables;

and $\bar{X}$ = the mean of the X_j values;

and S_j = the standard deviation of the X_j values

S_y = the standard deviation of the Y-values.

ii) the *beta coefficient* given by:

$$b_j^* = b_j \frac{S_j}{S_y}$$

iii) the coefficient of "part" correlation, given by:

$$r_{yj(kl\ldots)} = \frac{b_j S_j \sqrt{1 - R^2_{j\hat{j}}}}{S_y}$$

where b_j, S_j and S_y are as defined above. It is to be noted that the beta coefficient is the easiest measure, among the three, to compute. However, the beta coefficient involves the unadjusted standard deviations of the variables involved. Obviously, the three measures have different values. However, usually, the ranking in terms of importance of the independent variables on the dependent variable will be the same, although this will not always be the case.

3. Some general remarks may be pertinent at this point. The beta coefficients can be highly influenced by purposeful selection of sample values of one or more of the independent variables. That is, if the values of one or more of the independent variables are specified by the researcher, as in this case of "dummy" variables, the beta coefficients will have "sampling significance only with respect to a special universe in which the standard deviation of each of the independent variables is held constant for all possible samples." (Ezekiel & Fox, 1959). Thus, one should be judicious in using beta coefficients unless correlation models involving random sampling from a normally distributed "natural" universe are used.

REMARKS ON RELATIVE CONTRIBUTIONS OF MIXED EXPLANATORY VARIABLES TO THE VARIATION OF A REGRESSAND

By Tito A. Mijares, Ph.D., Academician

(The following prepared remarks were distributed to participants at the conference. Dr. Mijares restated the problem of "mixed" explanatory variables - discrete and continuous - in a general linear model, then proceeded to derive some tests on the regression coefficients to effect some comparison among them. By examining the correlation matrix of the "mixed" set of explanatory variables, Dr. Mijares arrived at an interesting result which offers a direct interpretation of coefficients of discrete independent variables in regression problems. The correlation coefficient between continuous and discrete variables measures the degree of inequality of a particular characteristic among the different attributes in the population; e.g. "income inequality").

We have a general linear model in matric form

$$Y = X\beta + \mu \tag{1}$$

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

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

$$y = x\beta + \epsilon \tag{2}$$

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

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

$x_{ij} = X_{ij} - \bar{X}_j$, $\bar{X}_j = \sum X_{ij}$ [illegible]

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

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

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

where e is a vector of n residuals $y - x\hat{\beta}$. It can be established that $\hat{\beta} = (x'x)^{-1}x'y$. The mean and variance of $\hat{\beta}$ are respectively β and $\sigma^2_\epsilon \ (x'x)^{-1}$. Equation (3) may be expressed by

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

where

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

In terms of Dr. Encamación's formulation (cf. eq. (1)) y is the "predictor" of $\hat{y}$. Thus, the vector y consists of the vector of *explained* and *unexplained* parts, 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 by $\hat{\beta}_{k-1}$ and $\hat{\beta}_k$, respectively. For a given element of $\hat{y}$ in this note

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

of that paper (cf. eq. (2), Encarnación'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. Encarnación as extensions 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

$$Y = \alpha_1 + \beta X + \epsilon \quad \text{(before the war)}$$
$$Y = \alpha_2 + \beta X + \epsilon \quad \text{(after the war)}$$

The two equations may be combined into a single equation

$$(6) \qquad Y = \alpha_0 + \delta_0 Z + \beta X + \epsilon$$

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

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

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

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., Encarnaciōn's paper). Hence, we may treat the problem as an ordinary linear regression problem, unrestricted case in the sense that no restrictions as 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^2_ϵ. The L.S. estimate of β is

$$\begin{aligned} \hat{\beta} &= (x'x)^{-1}\, x'y \\ &= \beta + (x'x)^{-1}\, x'\epsilon \end{aligned} \tag{7}$$

Then

$$\begin{aligned} E(\hat{\beta}) &= \beta \\ var\ (\hat{\beta}) &= E\ [(\hat{\beta}-\beta)\ (\hat{\beta}-\beta)'] \\ &= E\ [(x'x)^{-1}\ x'\epsilon\ \epsilon' x(x'x)^{-1}\ ] \\ &= \sigma^2_\epsilon\ (x'x)^{-1} \end{aligned} \tag{8}$$

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

$$t = \frac{c\,\hat{\beta} - c'\beta}{s_\epsilon \sqrt{c'\,(x'x)^{-1}\ c}} \tag{9}$$

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

We can now compare coefficients of classificatory variables (e.g. the coefficient of the i^{th} income group of one region against coefficient of the j^{th} income group of another region). By choosing

c appropriate to our hypotheses on the β's, we can make the tests on the coefficients. Let $c' = (0, \ldots, 0, 1, 0, \ldots, 0, -1, 0, \ldots, 0)$, the i^{th} element is 1 and the j^{th} element is -1 and zeros in other places. This is equivalent to testing $H_0: \beta_i - \beta_j = 0$ or β_j against $H_1: \beta_i \neq \beta_j$. The probability is α that $|t| > t_{\alpha/2, n-k}$, where $t_{\alpha/2, n-k}$ 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. Encarnación 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 $R = (r_{ij})$ where r_{0j} $(j = 1, \ldots, k)$ denotes the correlation between Y and X_j and $r_{ii} = 1 (i = 0, \ldots, k)$. Then the least squares regression $y = \beta_1 x_1 + \ldots + \beta_x x_k$ · where y, $x_1, \ldots, x_k$ are deviates of variables $Y, X_1, \ldots, X_k$ from their respective means would have coefficients

$$(11) \qquad \hat{\beta}_j = - \frac{s_0 R_{0j}}{s_j R_{00}}$$

where R_{0j} and R_{00} denote the co-factors of r_{0j} and r_{00} in the matrix R, respectively, and s_0, and s_j are the respective standard deviations of Y and X_j. An alternative expression for the least squares regression is

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

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

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

where $|R|$ is the determinant of matrix R. Since

$\sum y^2 = n s_0^2$, equation (10) becomes

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

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)

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

Take characteristic group h of classificatory variable X. Assume that the first v 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 values							Sums	Means
X:	1	1	$\ldots 1$	0	$\ldots 0$		v	$v/n = p$
W:	w_1	w_2	$\ldots w_v$	w_{v+1}	$\ldots w_n$		$\sum_1^n w_i$	$\frac{1}{n}\sum_1^n w_i$

Deviates

x:	$1-p$	$1-p$	$\ldots 1-p$	$-p$	$\ldots -p$	
w:	$(w_1-\bar{w})$	$(w_2-\bar{w})$	$\ldots (w_v-\bar{w})$	$(w_{v+1}-\bar{w})$	$\ldots (w_n-\bar{w})$	

Then $\qquad \sum_1^n x_i w_i = \sum_1^v (1-p)\,(w_i-\bar{w}) - \sum_{v+1}^n p(w_i-\bar{w})$

$$= (1-p)\ \left[\sum_1^v w_i - \frac{v}{n}\sum^n w_i\right] - p\sum_{v+1}^n w_i + p(n-v)\bar{w}$$

$$= \sum^v w_i - p\sum^v w_i - p\sum^v w_i + p^2 \sum^n w_i - p\ \left[\sum^n w_i - \sum^v w_i\right] + p(n-v)\bar{w}$$

This easily simplifies to

(16) $$\sum_1^n x_i w_i = \sum_1^v w_i - p\left(\sum_1^n w_i\right)$$

since

$$p^2 \sum_1^n w_i = pv\bar{w} \text{ and } p \sum_1^n w_i = pn\bar{w}$$

The simple correlation between x and w is

(17) $$r_{xw} = \frac{\sum_1^v w_i - p\sum_1^n w_i}{\sqrt{p_q s_w}}$$

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

Reference

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

RELATIVE CONTRIBUTIONS OF MIXED VARIABLE TO THE VARIATION OF A REGRESSAND

Burton T. Oñate, Ph. D.
Discussant

Being the last discussant, I assume that Drs. Parel and Mijares would be able to cover perhaps 90 per cent of what should be said. But least squares and regression is a broad field and my paper will deal on their theoretical foundations. The four methods of estimation in a general linear form are (i) ordinary least squares, i (ii) generalized least squares, (iii) maximum likelihood and (iv) best linear unbiased estimator (blue). Their equivalents are indicated depending upon the assumptions made.

Two well known points are worth mentioning, namely; (i) least squares estimation does not pre-suppose any distributional properties of the *e's* other than finite means and finite variances; (ii) maximum likelihood estimation under normality assumptions lead to the same estimator, b, as generalized least squares; and this reduces to the ordinary least squares estimator b when $V=d^2I$ Therefore, one could see that the estimation procedures will require the use of some transformations which essentially was applied by Dr. Mijares to derive the estimators, and the variance and co-variance matrices. These results of Dr. Mijares could be compared with those given in the paper under discussion. Existing computer programs should be tested for "integrity" and using the ramifications indicated in Mijares' discussion paper.

A survey on the "Method of Least Squares" has been conducted by S. L. Harter which appeared in several issues of the International Statistical Review of 1974-75. Harter divided this era into four parts, (I) The Pre-Least Squares, (II) The Awakening, (III) The Modern Era I and (IV) The Modern Era II. A subject index to the references arranged in alphabetical order of the Code Letters was used to classify more than 5,000 papers/authors. The paper under review could fall in II, III and IV.

The uses of code and dummy (0, 1) variables are illustrated in the Philippines by the National Census Statistics Office indicators on income (salary). One would see that the code used would be called classificatory variable as the level and the category inside as the factors and inside the factor as level. In occupation, they have developed for example codes 1, 2, 3, 4, 5, 6. One criticism is that one cannot use the values because no relationship exists in terms of occupational status. And to get away from this problem, so called dummy variables are used. Another example is education as a factor (page 5) and there are many levels under education (factor). Here, there is some kind of order but even then this order is in terms of educational status. Again, dummy variables would be useful.

(2) *Generalized least squares*

On assuming that the variance-covariance matrix of e is $var(e) = V$, this procedure involves minimizing $(y = Xb)' V^{-1} (y = Xb)$ with respect to b which leads to

$$\hat{\hat{b}} = (X'V^{-1} X)^{-1} X'V^{-1} y.$$

When $V = \sigma^2 I$, the generalized and the ordinary least squares estimators are the same: $\hat{\hat{b}} = \hat{b}$.

(3) *Maximum likelihood*

With least square estimation no assumption is made about the form of the distribution of the random error terms, which are represented by e. With maximum likelihood estimation some assumption is made about this distribution (often that it is normal) and the likelihood of the sample of observations represented by the data is then maximized. On assuming that the e's are normally distributed with zero mean and variance-covariance matrix V, *i.e.*, $e \sim N(0, V)$, the likelihood is

$$L = (2\pi)^{1/2N} \, |V|^{-1/2} \exp \, [-1/2(y - Xb)' V^{-1} (y - Xb)].$$

Maximizing this with respect to b is equivalent to solving $\partial \, (\log_e L)/\partial b = 0$. The solution is the maximum likelihood estimator of b is

$$\hat{\hat{b}} = (X'V^{-1} X)^{-1} X'V^{-1} y,$$

the same as the generalized least squares estimator. As before, when $V = \sigma^2 I$, $\hat{\hat{b}}$ simplies to $\hat{b}$. The estimator $\hat{b}$ is the maximum likelihood estimator, if we assume that

$$e \sim N(0, \sigma^2 I).$$

Two well-known points are worth mentioning about these estimators. First, least squares estimation does not pre-suppose any distributional properties of the e's other than finite means and finite variances. Second maximum likelihood estimation under normality assumptions lead to the same estimator, $\hat{\hat{b}}$, as generalized least squares; and this reduces to the ordinary least squares estimator $\hat{b}$ when $V = \sigma^2 I$.

(4) *The best linear unbiased estimator (b.l.u.e.)*

For any row vector t' comformable with b the scalar $t'b$ is a linear function of the elements of the parameter vector b. A fourth estimation procedure derives a best, linear, unbiased estimator (b.l.u.e.) of $t'b$.

The b.l.u.e. of $t'b$ is $t'(X'V^{-1}X)^{-1}X'V^{-1}y$, and its variance is

$$v(\text{b.l.u.e. of } t'b) = t'(X'V^{-1}X)^{-1}t.$$

From among all estimators of $t'b$ that are both linear and unbiased the one having the smallest variance is $t'(X'V^{-1}X)^{-1}X'V^{-1}y$; and the value of this smallest variance is $t'(X'V^{-1}X)^{-1}t$.

3. In view of this equivalence, it may be worthwhile to use the results for the Ordinary Least Squares Method and apply the suggested transformation in reducing the original x, z, u, and v to $N(0,1)$ instead of $N(0,I\sigma^2)$. Another suggested theoretical framework is the Principal Component Method.

Survey on Method of Least Squares

4. H. Leon Harter (1974, 1975) wrote a series of articles entitled "The Method of Least Squares and Some Alternatives", in the International Statistical Review (ISR). These series of articles are summarized as follows:

Part I Introduction, Pre-Least Squares Era (1632-1804) and Eighty Years of Least Squares (1805 - 1884); ISR (1974) 42, pp. 147-174.

II The Awakening (1885 - 1945); ISR (1974) 42, pp. 235 - 264, 282.

III The Modern Era (I) (1946 - 1964); ISR (1975) 43, pp. 1-44.

IV The Modern Era (II) (1965-1974); ISR (1975) 43, pp. 125-190; ISR (1975) 43, pp. 273-278 (Addendum).

A Subject Index to the references arranged in alphabetical order of the Code Letters was also made available (see Appendix Table A). A total of 148 Code Letters was used to classify more than

5,000 authors/papers. The paper under review could be classified under one or more of the Code Letters presented. If not, we could add a new Code Letter.

Uses of Codes and Dummy (0.1) Variables

5. An alternative analysis known as regression on dummy (0.1) variables has certain advantages but it may introduce into the linear model the problem of not of full rank. The NCSO uses codes in the presentation of detailed data on labor force, income and expenditure characteristics of household sampled. The regression of income (salary), expenditure and investment of sampled families on dummy (0.1) variables[1] may include class of worker (occupation), education and other characteristics which are coded: Examples of these codes are as follows:

Level/Class of Worker (Occupation) - Factor

1 - Worked for private employer
2 - Worked for government/government corporation
3 - Self-employed without any paid employee as defined in "4"
4 - Employer in own family-operated farm/business (with one or more regular paid employees or one or more hired employees most of the weeks of the last quarter in operation.)
5 - With pay on own family-operated farm or business
6 - Without pay on own family-operated farm or business

Highest Grade Completed (Education) - Factor Level

00 — No grade completed

Elementary

11 — 1st grade
12 — 2nd grade
13 — 3rd grade
14 — 4th grade
15 — 5th grade
16 — 6th grade and 7th grade

High School

21 — 1st year
22 — 2nd year
23 — 3rd year
24 — 4th year

College Graduate Level

31 — 1st year
32 — 2nd year
33 — 3rd year
34 — 4th year
35 — 5th year or higher

For college graduates

Specify the Bachelor's or highest degree completed and field of study.

[1] Searle, S.R. Linear Models. John Wiley & Sons, Inc. N.Y. 1971

The occupational and educational codes may be collapsed into 3 or 4 categories. One question is, "How can Occupation be Measured". One possibility is to measure it by the code numbers 1, 2, 3, 4, 5, 6. An inherent difficulty, however, occurs with the definition of x as a code number to measure occupational status. Although the six (6) categories of occupation or class of worker represent different kinds of occupation, the allocation of the numbers 1 to 6 to these categories as measures of occupational status may not accurately correspond to the underlying measure of whatever is meant by occupational status. The allocation of the number codes is, therefore, quite arbitrary. By giving a self-employed person an x-value of 3, we are not really saying that he has three times as much status as worked for private employer ($x = 1$). But in terms of the model, what we are saying is that

E(investment (i) or Income (In) of private employer) $= b_\sigma + b_1$

E((i) or (In) of self-employed) $= b_\sigma + 3b_1$

Thus, allocating codes to the different categories is not entirely justified so far as the suggested model is concerned. Such category codes are also used in many characteristics of interest such as education, management level, malnutrition, source of raw material, treatment and plant location in an industrial process, etc. This problem on code number is avoided by using the technique of regression on dummy (0,1) variables. Estimation procedures as illustrated above will immediately imply that a sound and scientific sample is drawn from the universe and from this sample, estimates are made of the parameters in the linear model. Even if the sample is drawn on a sound and scientific manner, it would be extremely difficult to generate equal number of data or the so-called balanced data. More often than not, there would be unequal numbers of observations in each category or sub-class including perhaps some categories with no observations at all. This situation is called unequal numbers data, unbalanced data or "messy" data. Some difficulties will be met in the analysis.

6. In studying the effects of occupation, education or malnutrition, on investment or income behavior, we are interested in the extent to which each category of each variable is associated with investment. To acknowledge the measurability of the variable and the associated arbitraries or subjectivity in dealing on their categories, the concept of "factor" and "level" may be introduced. The word "factor" denotes the occupation, education, malnutrition which in turn are divided into "levels". Examples were given earlier. The "factor" cannot be measured precisely by a cardinal value while the word "variable" is reserved for that which can be measured. Thus, investment, income or salary are variables. Note that each person falls into one and only one

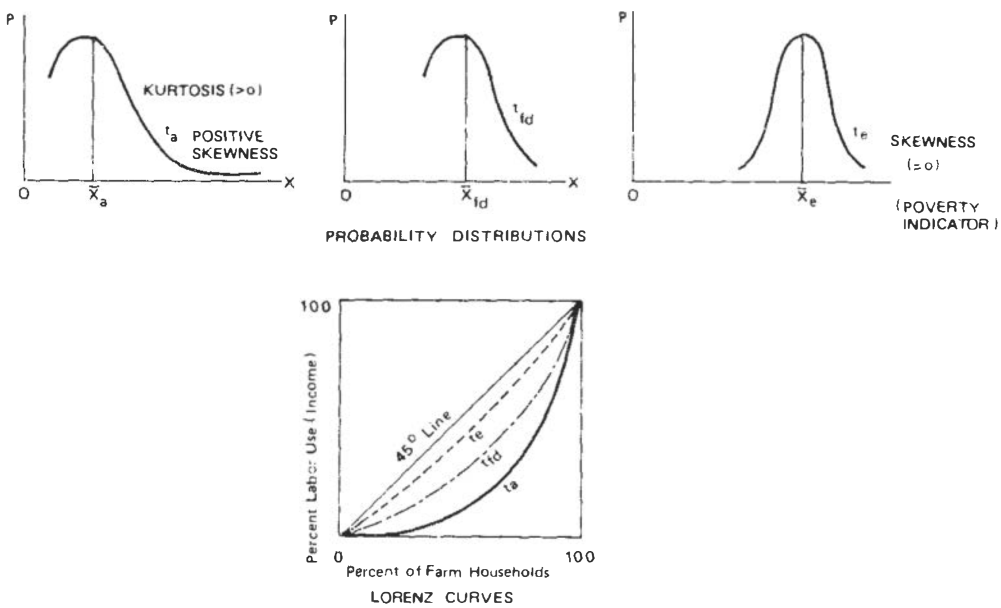
occupational or educational level to which he belongs to. Let the corresponding x take the value unity (1) and let all other x's for that person to have a value of zero (0). Note that in the model of the paper under review, there is a mixture of both dummy (0,1) and measurable variables similar to y. Care must be taken to insure that the resultant X matrix is of full rank.

Sampling Variation and Resultant Distribution

7. Selected indicators will illustrate the level of and distributional property of poverty indicators though the major periods in the project cycle. i.e.,

t_a = prior to or at appraisal time,
t_{pe} = at completion time or at post-evaluation
t_{fd} = at full development, and
t_e = at end of project life.

Chart A. Probability Distributions and Lorenz Curves of Indicators from ARD Projects.



Some of the indicators are production oriented. They are, however, related to poverty indicators such as ownership, size of land and yield, employment and labor inputs, etc. All of the indicators are skewed to the right showing extreme inequalities at the beginning of the project life (t_a or t_{pe}) except perhaps the data on price or value of paddy. Chart A shows empirically how the Project Benefit Monitoring and Evaluation System (PBMES) will be able to measure and illustrate the level and distribution of each poverty indicator which is relevant to the project site.[1] These distributions could serve as framework in the sampling procedures and to the levels of variation in the V matrix on a time series.

[1] Oñate, B.T. Benefit Monitoring and Evaluation System as a Component of ARD Project Design. ADB. 1981

Appendix Table A

METHODS OF LEAST SQUARES AND SOME ALTERNATIVES
(H. LEON HARTER)

Glossary of Code Letters

AC	Arley's criterion (for rejection of outliers)
AD	Average (absolute) deviation
AE	adaptive estimators
AI	Adichie's estimators (of regression coefficients)
AM	arithmetic mean
AR	Anscombe's rules (for rejection of outliers)
AS	average slope (of regression lines)
AT	Andrew's tests (for rejection of outliers)
AV	average (all types)
BC	Bertrand's criterion (for rejection of outliers)
BF	Bartlett's (method of) fitting (straight lines)
BM	Brown-Mood estimators (of regression parameters)
BT	best two (out of three)
CC	Chauvenet's criterion (for rejection of outliers)
CD	censored data
CH	cliff hangers
CM	Cauchy's method (of interpolation)
CT	(Bliss)-Cochran-Tukey criterion (for rejection of outliers)
CU	Cucconi's criterion (for rejection of outliers)
DA	discard averages (trimmed means)
DC	Dixon's criterion (for rejection of outliers)
DH	differences at half range

DI	dispersion (measures of)
DQ	Quesenberry-David criterion (for rejection of outliers)
EA	equal areas (under joint p.d. curve) (Laplace's "most advantageous method")
EB	empirical Bayes approach (to outliers)
EE	van Eeden estimators (of location parameters)
EM	Edgeworth's modification (of Stone's second criterion)
EX	extremes (largest and smallest values in sample)
FC	Ferguson's criterion (for rejection of outliers)
FM	folded medians
GA	Gastwirth estimators
GC	Glaisher's criterion (for rejection of outliers)
GD	Gini's mean difference
GE	geometric midrange
GG	geometric range
GM	geometric mean
GP	generalized Pitman estimators
GR	Goodwin's rule (for rejection of outliers)
GS	Grubbs' criterion (for rejection of outliers)
HA	Hodges' alternative (to Hodges-Lehmann estimator)
HC	Heydenreich's criterion (for rejected outliers)
HE	Harter's estimators (1972)
HG	Hogg's revised estimator (1972)
HL	Hodges-Lehmann estimator
HM	harmonic mean
HO	Hogg's estimator (1967)
HQ	Hogg's estimators based on Q statistic.
HS	Hulme-Symms alternative (to the rejection of outliers)
HU	Huber's estimator
HV	Harter's regression estimators with variable boundaries
IC	Irwin's criterion (for rejection of outliers)
IR	interquartile range
JA	Jeffrey's alternative (to the rejection of outliers)
JE	Jureckova's estimators (of regression coefficients)
JO	Jorgenson's estimators
KC	Kudo's criterion (for rejection of outliers)
KE	Kraft-van Eeden estimators (of location parameters)
KT	Kendall's tau estimator (Sen)
LA	Laurent's analogue (of Thompson's criterion)
LD	largest (absolute) deviation
LE	L-estimators (linear combinations of order statistics)
LF	least (sum of absolute) first (powers) (Laplace's "method of situation")
LN	least number of deviations (least sum of zero powers)
LP	least (sum of) *p*th (powers of absolute deviations)
LR	linear regression
LS	least squares
LW	linearly weighted means
MA	method of averages
MC	Merriman's criterion (for rejection of outliers)
MD	median
ME	M-estimator (maximum likelihood type)
MG	method of group averages
MH	Harter's modified estimators (1973)

MK	McKay's criterion (for rejection of outliers)
ML	maximum likelihood
MM	minimax method (minimize maximum residual)
MO	mode
MQ	median-quartile average
MR	midrange
MS	method of successive differences
MT	median and two other order statistics
MU	Murphy's criterion (for rejection of outliers)
MV	Moore's variable-bound estimators
MW	multivariate Wilks' criterion (for rejection of outliers)
MZ	Mazzuoli's criterion (for rejection of outliers)
M4	maximum (sum of) fourth (powers of p.d.f. of errors)
NC	Nair's criterion (for rejection of outliers)
ND	median deviation
NM	Newcomb's method (of treating outliers)
NR	nonlinear regression
NS	Nair-Shrivastava method (of curve fitting)
OM	Ogrodnikoff's method (of treating outliers)
OS	order statistics
PA	plus approximative methode (most approximative method)
PC	Peirce's criterion (for rejection of outliers)
PD	dispersion with norm p
PL	location with norm p
PM	power means
PS	Pearson—Chandra Sekar criterion (for rejection of outliers)
QA	quadratic average (mean)
QD	quartile deviation (semi-interquartile range)
QL	quasilinear estimators
QM	quasi-midrange (quasi-median)
QN	quantiles
QR	quasi-range
QT	quarter technique
RA	range
RC	Rohne's criterion (for rejection of outliers)
RE	R-estimators (based on rank tests)
RL	robust estimators of location
RM	range method
RR	robust estimators of regression
RS	robust estimators of scale
SA	stochastic approximation estimators
SB	semi-Bayesian approach (to outliers)
SC	Stone's (first) criterion (for rejection of outliers)
SD	standard deviation (or variance = SD^2)
SE	sine estimator
SH	shortest half estimators
SI	successive interval method
SK	skipped procedures
SM	Stewart's method (criterion) (for rejection of outliers)
SN	Schuster-Narvarte estimator
SP	(method of) selected points
SR	semirange
ST	Student's rule (for rejection of outliers)
SW	Switzer's estimator

S2	Stone's second criterion (for rejection of outliers)
TC	Tippett's criterion (for rejection of outliers)
TD	transformation of data (and choice of model)
TE	theory of errors
TF	Tukey's FUNOR-FUNOM procedure
TJ	Topsoe-Jensen criterion (for rejection of outliers)
TM	Thompson's method (criterion) (for rejection of outliers)
TO	treatment of outlying observations
TR	trimming
VC	Vallier's criterion (for rejection of outliers)
WA	weighted average
WC	Wright's criterion (for rejection of outliers)
WH	Wright-Hayford criterion (for rejection of outliers)
WI	Winsorization
WK	Walsh-Kelleher estimators
WM	Winsorized means
WR	Walsh's rule (criterion) (for rejection of outliers)
WV	Winsorized variances
YE	Yanagawa's estimator