AN AUXILIARY MODEL FOR QUANTIFYING THE SOCIOECONOMIC IMPACT OF A DEVELOPMENT PROJECT

José Encarnación, Jr. *

1. Introduction

A development project is typically designed to advance only one or a few of the objectives that concern development planners. Different projects specialize, so to speak, in the pursuit of different objectives, and a project may even have a negative impact on another area of concern. (A rural road might, for example, increase productivity in the area but also increase urban unemployment by facilitating rural-urban migration.) With sufficient data and correct model formulation, one could calculate the impact of each project on all the areas of concern. For this purpose it would be useful to distinguish between: (a) relationships among variables that are specific to projects, and (b) relationships that are common to all projects. With a model of (b) in hand, impact analysis of a project could focus on (a) and then make use of the results already available from (b). A model of (b) is then auxiliary to (a).

This paper gives a partial specification of (b) using the 1973 National Demographic Survey (NDS), which data are incomplete for the purposes of comprehensive project impact estimates.

2. Data and Notation

The data are from the 1973 NDS of over 8,000 households. Our sample size of 3,196 was obtained by selecting households satisfying

*University of the Philippines School of Economics. The author is indebted to Ms. Elizabeth Jacinto who did the computations for this paper.

the following criteria: the family is nuclear or extended vertically to the younger generation only; household head is male, working and his noncash income (if any) is less than ₱1,000 annually; the wife married only once and age is between 15-44 years; and information is provided on all the variables listed below.

AGn = 1 if wife is in age-group *n*, 0 otherwise,
where *n* = 4 if age is 15-19 years
5 if age is 20-24 years
6 if age is 25-29 years
7 if age is 30-34 years
8 if age is 35-39 years
9 if age is 40-44 years

AM = age of marriage of wife, in years

CEB = number of children born live

CMR = *CND* ÷ *CEB*

CND = number of children born live and now dead

DCM = 1 if *CND* > 0, 0 otherwise

DLR = 1 if rural residence, 0 otherwise

DMW = 1 if wife is a migrant whose place of residence is the same as in 1965 and different from place of birth, 0 otherwise

DRC = 1 if wife is Roman Catholic, 0 otherwise,

DWP = 1 if wife is working, 0 otherwise

EWm = 1 if wife has educational level *m*, 0 otherwise, where
m = 0 for no schooling
1 for one to four years of school
2 for five to seven years of school
3 for one to three years of high school
4 for high school graduate
5 for one to three years of college
6 for college graduate

MWk = 1 if wife is in category *k*, 0 otherwise,
where *k* = 0 for *DMW* = 0
1 for DMW = 1 and agricultural residence
2 for DMW = 1 and nonagricultural residence

*PWR*j = 1 if wife is in category *j*, 0 otherwise,
where *j* = 0 for *DWP* = 0
1 for *DWP* = 1 and place of work is at home
2 for *DWP* = 1 and place of work is away from home

YHi = 1 if husband's annual income is in category i, 0 otherwise,
where i = 1C for cash income less than ₱1000 and noncash income (if any) less than ₱1000 2C for ₱1000-2999 cash income and noncash income (if any) less than ₱1000
3 for ₱3000-4999 cash income
4 for ₱5000-6999
5 for ₱7000-9999
6 for ₱10000 and above

YWi defined the same way as *YHi* but with respect to wife's income

YW= 1 if *YW1C* = 1
2 if *YW2C* = 1
4 if *YW3* = 1
6 if *YW4* = 1
8 if YW5 = 1
11 if *YW6* = 1, in thousand pesos.

Individual income data in the 1973 NDS are reported only in brackets and family income as such is not given. We therefore do not use a family income variable as this would have involved too many categories or else a summing of individual incomes by taking the midpoints of categories as estimates of individual incomes. While the latter procedure is of course possible (cf. Canlas and Encarnación, 1977), it makes income data appear more precise than may be warranted.

The means of the variables in the sample are given in Table 1.

Table 1 – Means of Variables

[illegible] 0.0185	CND: 0.4143	EW3: 0.1126	PWR2: 0.1805
[illegible] 0.1302	DCM: 0.2663	EW4: 0.0726	YH1C: 0.3457
[illegible] 0.1990	DLR: 0.6815	EW5: 0.0379	YH2C: 0.4562
[illegible] 0.2362	DMW: 0.2700	EW6: 0.0660	YH3: 0.1270
[illegible] 0.2280	DRC: 0.8483	MWO: 0.7300	YH4: 0.0291
[illegible] 0.1881	DWP: 0.2447	MW1: 0.1549	YH5: 0.0217
[illegible] 19.666	EWO: 0.0685	MW2: 0.1151	YH6: 0.0203
[illegible] 4.8276	EW1: 0.2735	PWRO: 0.7553	
[illegible] 0.0671	EW2: 0.3689	PWR1: 0.0641	

3. The Model

This is based on an earlier paper (Encarnación, 1982) which presented a model of choice where wife's fertility and her labor force participation are determined simultaneously by her education level, husband's income, and other variables. Briefly, the model implies the existence of "threshold values" for wife's education level, husband's income and family income, such that the qualitative effect of a variable changes when it passes the thresholds. Figure 1 illustrates.

In the upper panel of Figure 1, number of children C is measured on the vertical axis while family income Y and wife's education level E – Y and E are assumed to be perfectly correlated for purposes of a simple diagram – are measured on the horizontal axis. Natural fertility or capacity to bear children C_k increases with Y and E for reasons of better health and nutrition; number of child deaths C_m decreases for the same reasons. The number of children desired C^o falls with E and Y for a variety of reasons. What would then be observed for the number of children born would be the curve abC_b, and the number of surviving children the curve cd. These two variables are thus nonmonotonic functions of E and Y whose qualitative effects change at the threshold value E_c*.

In the lower panel, the proportion of wife's time spent at market work t is measured on the vertical axis while E and husband's income Y_h are measured on the horizontal axis. The wife's wage rate depends on E and we assume that the curve $c'd'e't'$ indicates what is required of t if minimum consumption standards for the family are to be met. On the other hand, the curve $ee't^o$ indicates what t would be if the wife's choice were not required to satisfy consumption standards. With this requirement, the observed t would be the curve $c'd'e't^o$. The threshold value E_t* defined by the intersection point e' is such that $E_c^* \leqq E_t^*$ under relatively weak assumptions.[1]

We thus have a roughly V-shaped curve for wife's labor force participation rate and an inverted V-shaped curve for her fertility as functions of education and income variables. Estimates of these two relationships are given in eq. (1) below and eqs. (2)–(5) set out in Table 2 (t-values in parentheses underneath regression coefficients):

$$(1) \quad DWP = \underset{}{0.3949} - \underset{(-0.83)}{0.0400}\, EWO - \underset{(-3.97)}{0.1642}\, EW-$$

1. In the earlier paper cited above, it was shown that $E_c^* = E_t^*$ under somewhat stronger assumptions.

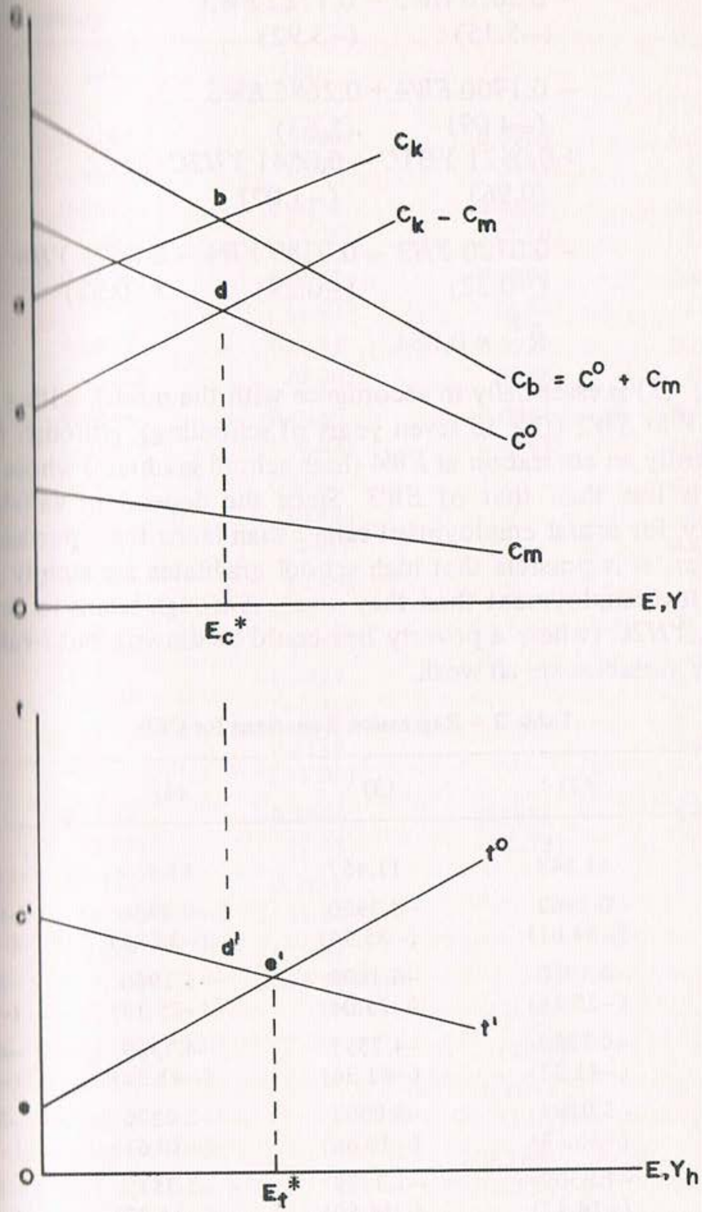

FIGURE 1

$$
\begin{aligned}
&- 0.2078\,EW2 - 0.1722\,EW3 \\
&\quad (-5.15) \qquad\quad (-3.92) \\
&- 0.1900\,EW4 + 0.2685\,EW6 \\
&\quad (-4.09) \qquad\quad (5.63) \\
&+ 0.0521\,YH1C - 0.0541\,YH2C \\
&\quad (0.96) \qquad\qquad (-1.02) \\
&- 0.0120\,YH3 - 0.0189\,YH4 - 0.0675\,YH6 \\
&\quad (-0.22) \qquad\quad (-0.29) \qquad\quad (-0.95) \\
&\bar{R}^2 = 0.084
\end{aligned}
$$

Eq. (1) is essentially in accordance with the model, with a trough for *EW* at *EW*2 (five to seven years of schooling), although there is apparently an aberration at *EW*4 (high school graduate) whose coeffi-cient is less than that of *EW*3. Since the dependent variable is a dummy for actual employment rather than labor force participation, however, it is possible that high school graduates are simply getting much less employment than they want. A trough seems to occur for *YH* at *YH*2C (where a poverty line could be drawn), but *t*-values for the *YH* variables are all weak.

Table 2 – Regression Equations for CEB

	(2)	(3)	(4)	(5)
const.	11.543	11.457	11.539	11.4[illegible]
AM	−0.2962	−0.2960	−0.2956	−0.2[illegible]
	(−34.01)	(−33.94)	(−33.88)	(−3[illegible])
AG4	−6.1980	−6.1899	−6.1966	−6.1[illegible]
	(−25.14)	(−25.04)	(−25.13)	(−25.[illegible])
AG5	−4.7382	−4.7555	−4.7369	−4.7[illegible]
	(−41.33)	(−41.36)	(−41.32)	(−41.[illegible])
AG6	−3.0293	−3.0392	-3.0276	−3.0[illegible]
	(−30.63)	(−30.66)	(−30.61)	(−30.[illegible])
AG7	−1.3509	−1.3539	−1.3517	−1.3[illegible]
	(−14.47)	(−14.50)	(−14.47)	(−14.[illegible])
AG9	1.0110	1.0189	1.0105	1.01[illegible]
	(10.25)	(10.32)	(10.23)	(10.[illegible])
EWO	0.3835	0.3291	0.3688	0.32[illegible]
	(1.84)	(1.54)	(1.76)	(1.[illegible])

[illegible] (Continued)

	(2)	(3)	(4)	(5)
[illegible]	0.6199	0.5600	0.6055	0.5547
	(3.52)	(3.07)	(3.42)	(3.04)
[illegible]	0.4615	0.4012	0.4505	0.3976
	(2.67)	(2.26)	(2.60)	(2.34)
[illegible]	0.4606	0.3981	0.4591	0.4023
	(2.43)	(2.07)	(2.42)	(2.09)
[illegible]	0.0913	0.0533	0.0929	0.0587
	(0.45)	(0.26)	(0.46)	(0.29)
[illegible]	−0.0844	−0.0160	−0.0941	−0.0288
	(−0.41)	(−0.08)	(−0.45)	(−0.14)
[illegible]		0.1035		0.0901
		(0.44)		(0.38)
[illegible]		0.2065		0.2016
		(0.89)		(0.89)
[illegible]		0.1095		0.1134
		(0.46)		(0.48)
[illegible]		−0.1025		−0.1025
		(−0.36)		(−0.36)
[illegible]		−0.3176		−0.3046
		(−1.03)		(−0.98)
[illegible]	0.1939	0.1894	0.1936	0.1888
	(2.13)	(2.08)	(2.13)	(2.07)
[illegible]	0.1956	0.1898		
	(2.71)	(2.62)		
[illegible]	−0.2548	−0.2465		
	(−3.31)	(−3.19)		
[illegible]			0.2561	0.2461
			(2.87)	(2.76)
[illegible]			0.1094	0.1070
			(1.07)	(1.03)
[illegible]			−0.3104	−0.3061
			(−2.38)	(−2.34)
[illegible]			−0.2308	−0.2193
			(−2.62)	(−2.47)
[illegible]	0.563	0.563	0.562	0.563

In Table 2, eqs. (2)-(5) for number of children born live
out as columns; (2) omits the *YH, MW* and *PWR* variables
other three equations similarly omit some of the variables
the first column. In all four equations, a peak for *EW* is seen
(one to four years of schooling). A peak for *YH* is apparent
in (3) *a*nd (5) even though *t*-values are weak. (The model
family income here and not husband's income, but we use the
as a rough proxy.)

Of interest are the dummies for religion (*DRC*), migrant
(*DMW*) and current employment (*DWP*), which are all signi
Apparently, looking at (2) and (3), being a Catholic adds about
children to a couple, as also being a migrant, while being em
reduces family size by 0.25. A closer look at the migrant and
ployment variables shows finer detail.

Eqs. (4) and (5) use *MW*1, *MW*2, *PWR*1 and *PWR*2 in pla
the more crude *DMW* and *DWP*. Here it is migrants to agricu
areas (who are likely to have come from other agricultural
who have higher fertility, while other migrants exhibit a
increase not significantly different from zero. This would be
sistent with the model if one considers that agricultural mi
probably improve their livelihood relatively more than do
migrants. (Cf. Encarnación, 1977, for similar suggestive
in Southeast Asia.) As for the employment dummy varia
breakdown of this to *PWR*1 and *PWR*2 gives results that
to go against usual expectations. Here we find that ceteris pa
wives who work at home have apparently less children than
whose place of work is away from home. A possible explanation
be that wives working at home find it difficult to hold down
away from home because of poorer health; their working at h
and having less children would then be due to the same set
cumstances.[2]

As the *CEB* equations involve age-at-marriage *AM*, we have
following equation:

$$(6) \qquad AM = 21.575 - \underset{(-4.45)}{1.9811\, EWO} - \underset{(-5.15)}{1.9765\, EW1}$$

2. However, it should be noted that if the regression coeffic
PWR1 and PWR2 are treated as means in a standard test of the diff
between two means, we find that the difference between them is not
enough to reject the null hypothesis.

$$- 1.9149\,EW2 - 1.4952\,EW3$$
$$(-5.13) \qquad (-3.70)$$
$$+ 0.0831\,EW4 + 2.1397\,EW6 - 0.7416\,DLR \quad \bar{R}^2 = 0.101$$
$$(0.19) \qquad (4.91) \qquad (-4.63)$$

This is quite in conformity with usual expectations: *AM* is higher with higher *EW* and is lower for rural women.

Finally, the *NDS* data permit the estimation of several equations concerning child mortality. Eqs. (7) and (8) below have *CMR*, the ratio of child deaths to children ever born, as a function of educational level and, in the case of the latter equation, of husband's income (as proxy for family income) as well. Its relationship to these

$$(7) \quad CMR = 0.0395 + 0.0775\,EWO + 0.0423\,EW1$$
$$(5.04) \qquad (3.21)$$
$$+ 0.0281\,EW2 + 0.0158\,EW3$$
$$(2.17) \qquad (1.10)$$
$$- 0.0028\,EW4 - 0.0175\,EW6 \quad \bar{R}^2 = 0.023 \quad F = 13.76$$
$$(-1.18) \qquad (-1.13)$$

$$(8) \quad CMR = 0.0291 + 0.0692\,EWO + 0.0347\,EW1$$
$$(4.38) \qquad (2.54)$$
$$+ 0.0221\,EW2 + 0.0111\,EW3$$
$$(1.66) \qquad (0.77)$$
$$- 0.0062\,EW4 - 0.0155\,EW6 + 0.0251\,YH1C$$
$$(-0.41) \qquad (-0.99) \qquad (1.41)$$
$$+ 0.0115\,YH2C + 0.0167\,YH3$$
$$(0.65) \qquad (0.93)$$
$$- 0.0023\,YH4 - 0.0009\,YH6 \quad \bar{R}^2 = 0.025 \quad F = 8.34$$
$$(-0.11) \qquad (-0.04)$$

the variables is generally monotonic as one might expect. Perhaps more useful, however, are eqs. (9) and (10), where the dependent variable *DCM* is a dummy equal to one if a child has died. *DCM* could be interpreted as the probability (approximately) of a child death as a function of the variables on the right-hand side. It could also serve as a crude proxy for health.

$$(9) \quad DCM = 0.7637 - 0.0233\,AM - 0.3851\,AG4$$
$$(-11.54) \qquad (-6.75)$$
$$- 0.3069\,AG5 - 0.2099\,AG6$$

$$\begin{aligned}
&\quad (-11.66) \qquad (-9.19) \\
&- 0.1060\,AG7 + 0.0734\,AG9 + 0.1608\,EW0 \\
&\quad (-4.90) \qquad (3.21) \qquad (3.38) \\
&+ 0.0985\,EW1 + 0.0629\,EW2 \\
&\quad (2.42) \qquad (1.58) \\
&+ 0.0322\,EW3 - 0.0162\,EW4 - 0.0485\,EW6 \quad \bar{R}^2 = 0.1 \\
&\quad (0.73) \qquad (-0.35) \qquad (-1.02)
\end{aligned}$$

$$\begin{aligned}
(10) \quad DCM &= 0.7530 - 0.0231\,AM - 0.3976\,AG4 \\
&\qquad\qquad (-11.44) \qquad (-9.96) \\
&- 0.3116\,AG5 - 0.2145\,AG6 \\
&\quad (-11.81) \qquad (-9.38) \\
&- 0.1076\,AG7 + 0.0747\,AG9 + 0.1288\,EW0 \\
&\quad (-4.98) \qquad (3.27) \qquad (2.64) \\
&+ 0.0684 + EW1 + 0.0383\,EW2 \\
&\quad (1.63) \qquad (0.94) \\
&+ 0.0117\,EW3 - 0.0313\,EW4 - 0.0371\,EW6 \\
&\quad (0.26) \qquad (-0.67) \qquad (-0.77) \\
&+ 0.0587\,YH1C + 0.0229\,YH2C \\
&\quad (1.08) \qquad (0.43) \\
&+ 0.0272\,YH3 - 0.0398\,YH4 - 0.0915\,YH6 \quad \bar{R}^2 = 0.1 \\
&\quad (0.49) \qquad (-0.60) \qquad (-1.28)
\end{aligned}$$

probability (approximately) of a child death as a function of the variables on the right-hand side. It could then serve as a crude pro[xy] for health.

4. Using the Model

With due caution, one can use the regression equations rep[orted] above[3] for purposes of estimating the impact of a develop[ment] project on some variables of concern: fertility, labor force par[ti]cipation, and health. Accepting the usual interpretation of cro[ss-]section regression results as long-term relationships among th[e]

3. These are ordinary least-squares estimates since the model can be taken as recursive.

variables, the procedure would simply be the following: Calculate the changes in the "independent" variables resulting from the project, then use the regression equations to estimate the changes in the dependent variables. The latter changes are then imputable to the project as its impact.

For example, suppose that one long-term effect of a project is to raise male family heads' incomes in the region from $YH1C$ to $YH2C$. From eqs. (1), (5) and (10), the coefficients of these two variables and the differences between them are given in Table 3. Accordingly, we obtain estimates of a reduction in wives' labor force participation, an increase in births and a decrease in child deaths by multiplying the last column in Table 3 by the number of families involved. Comparability among projects can then be had by expressing the estimates per peso of project costs.

Table 3 – Coefficients of YH1C and YH2C

Equation	YH1C	YH2C	Difference
(1) DWP	.0521	–.0541	–.1062
(5) CEB	.0901	.2016	.1115
(10) DCM	.0587	.0229	–.0358

If a project affects other "independent" variables, similar computations can be made and then added to get the total impact of the project, since the effects of the independent variables are additive in the regression equations.[4]

Several observations might be made regarding the estimates so obtained. First, they are not predictions of changes between the present and the future (after the installation of a project), since these changes will occur with or without the project. One is here merely estimating the difference that a project makes, ceteris paribus. Second, the estimates have to do with long-term, not short-term,

4. It is sometimes of interest to calculate the relative contributions of the independent variables to the variation of the dependent variable in a regression equation, especially when the independent variables are "mixed" as in eqs. (9) and (10); see the Appendix.

results. Finally, it is on the basis of some model which one siders correct that one justifies any particular interpretation statistical observations – one cannot discuss the latter in a retical vacuum. This last observation would not be worth tioning were it not that statistical data are sometimes erroneously thought to be capable of "establishing" a causal relationship.

Appendix

Relative Contributions of Mixed Variables to the Variation of a Regressand

Consider a regression equation whose regressors include ficatory as well as ordinary scalar variables. A classificatory variable is essentially a vector that has as many components as there different (mutually exclusive and exhaustive) categories in classification. For example, one might estimate a regression equation that explains employees' salaries in terms of length of service scalar), occupation (a classificatory variable), etc. One might want to estimate the relative contributions of the explanatory variables to the variation of the dependent variable. Handling 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 include both kinds. This expository note might therefore be of use.

I

Let $x = (x_0, x_1, \ldots, x_K)$ where $x_k = 1$ for an individual 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, any given individual i, $x_{ki} = 1$ if i is in category k, 0 otherwise, $\Sigma_{k=0}^{K} x_{ki} = 1$. 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 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) $$y' = c + \sum_{1}^{K} a_k^* \; x_k + \sum_{1}^{J} b_j^* \; z_j + p\,(u - \overline{u}) + q(v - \overline{v})$$

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

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 $\overline{y}$. For suppose that for an individual i, $x_{ki} = 1$ for a particular k and $z_{ji} = 1$ for a particular j. Then

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

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

From least squares properties, using (1),

(3) $$c = \overline{y} - \sum_{1}^{K} a_k^* \, \overline{x_k} - \sum_{1}^{J} b_j^* \, \overline{z_j} - p(\overline{u} - \overline{u}) - q(\overline{v} - \overline{v})$$

$$= \overline{y} - \sum_{1}^{K} a_k^* \, \overline{x_k} - \sum_{1}^{J} b_j^* \, \overline{z_j}.$$

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

(4) $$a_0 = -\sum_{1}^{K} a_k^* \, \overline{x_k}$$

(5) $$b_0 = -\sum_{1}^{J} b_j^* \, \overline{z_j}.$$

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

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

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

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

The b_j are similarly determined.

Substituting (6) in (1),

$$\begin{aligned}(8) \qquad y' &= \overline{y} + a_0 + b_0 + \sum_1^K a_k^* x_k + \sum_1^J b_j^* z_j + p(u - \overline{u}) + q(v - \overline{v}) \\ &= \overline{y} + a_0 + b_0 + \sum_1^K (a_k - a_0) x_k + \sum_1^J (b_j - b_0) z_j \\ &\quad + p(u - \overline{u}) + q(v - \overline{v}) \\ &= \overline{y} + a_0 \left(1 - \sum_1^K x_k\right) + \sum_1^K a_k x_k + b_0 \left(1 - \sum_1^J z_j\right) \\ &\quad + \sum_1^J b_j z_j + p(u - \overline{u}) + q(v - \overline{v}).\end{aligned}$$

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 $x_k = n_{k.}/n$, 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,

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

i.e., the mean $\Sigma_0^K a_k x_k = 0$ (in the same way that the mean $p(u - \overline{u})$, 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. The problem is to see whether all the beta coefficients in a regression with mixed variables are directly comparable.

Write

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

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

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

where s_y is the standard deviation 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, and

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

from Morgan et al. (1962). 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 $\beta_u > \beta_v$, 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 $(u - \bar{u})/s_u$ standardizes u, so that all the beta coefficients are then directly comparable.

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

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

from which

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

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

$$(16) \quad \frac{(u_i - \overline{u})^2}{s_u^2} = \frac{p^2 (u_i - \overline{u})^2}{\Sigma_{h=1}^{n} p^2 (u_h - \overline{u})^2/(n-1)}$$

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 $\Sigma_0^K a_k^2$ 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 $p(u_i - \overline{u})$ and both have zero mean.

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

Canlas, D. B. and Encarnacion, J. (1977), "Income, Education, Fertility and Employment: Philippines 1973," *Philippine Review of Business and Economics,* 14 (2): 1-27.

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, pp. 742-775.

---- (1977), "Population and Development in Southeast Asia: A Fertility Model," *Philippine Economic Journal,* 16: 319-340.

---- (1982), "Fertility Behavior and Labor Force Participation: A Model of Lexicographic Choice," in *Research in Population Economics,* ed. J. L. Simon and P. Lindert, Greenwich, Conn.: JAI Press.

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.