AGENCY FOR INTERNATIONAL DEVELOPMENT
WASHINGTON, D. C. 20523

BIBLIOGRAPHIC INPUT SHEET

FOR AID USE ONLY
Batch 52

1. SUBJECT CLASSIFICATION	A. PRIMARY	TEMPORARY
	B. SECONDARY	

2. TITLE AND SUBTITLE
Measuring effects of demographic variables on health services utilization

3. AUTHOR(S)
Reinke,W.A.; Baker,T.D.

4. DOCUMENT DATE	5. NUMBER OF PAGES	6. ARC NUMBER
1967	15p.	ARC

7. REFERENCE ORGANIZATION NAME AND ADDRESS
Johns Hopkins

8. SUPPLEMENTARY NOTES (Sponsoring Organization, Publishers, Availability)
(In Health services research,spring 1967,p.61-75)

9. ABSTRACT
(HEALTH R & D)

10. CONTROL NUMBER PN-AAD-265	11. PRICE OF DOCUMENT
12. DESCRIPTORS	13. PROJECT NUMBER
	14. CONTRACT NUMBER CSD-832 Res.
	15. TYPE OF DOCUMENT

AID 590-1 (4-74)

ead 832 - Rcs
PN-AAD-265

Measuring Effects of Demographic Variables on Health Services Utilization

by William A. Reinke and Timothy D. Baker

To predict demands for health services, planners must measure the impact of demographic change. This article presents a new method, the multi-sort technique, to improve analysis of the effects of multiple demographic variables. The technique is demonstrated in an analysis of the changing demand for physicians' services in Taiwan.

Multiple regression techniques, including those that utilize dummy variables, have been used to analyze effects of demographic variables, but interactions may be overlooked entirely or inadequately identified. Analysis of variance has proved useful in handling interactions, but uneven distribution of observations among cells creates orthogonality problems.

The multi-sort technique is an approximation procedure that simplifies computations while maintaining the analysis of variance approach. The procedure assigns cell weights in all evaluations according to the rules for evaluating main effects; thus the assessment of interactions is approximate but not tedious.

Comparison of computer results obtained by the new technique with those from certain multiple regression procedures indicates that the multi-sort technique offers more promise in the evaluation of demographic data than previous applications of analysis of variance would indicate. The study also gives practical suggestions for combining various techniques in a total analysis.

The Taiwanese data in this paper were obtained from a survey supported by The Agency for International Development.

Introduction

To predict future demands for health services in a dynamic society, health planners must be able to measure the impact of demographic change. Illness patterns and hospital and physician utilization rates usually vary according to age, economic status, education, and other factors. As these demographic factors change, so will the total demand for medical care. A method for measurement of the effects of demographic variables on the use of health services is presented.

Presentation of Data

A major barrier to understanding the impacts of multiple variables is the difficulty in presenting information clearly without over-simplification. One-way tabulations are frequently employed in an effort to depict the impact of each individual demographic variable in turn. These tabulations fail to recognize such factors as different levels of education that may prevail in different

socio-economic groups. In the absence of correlations and interactions, however, multi-dimensional tables present unnecessary detail and difficulty in interpretation. Our method of analysis offers opportunity for clarification without over-simplification in presentation of data.

Projections

An understanding of basic determinants is vital for health planning in a dynamic society. Knowledge of the overall population growth rate is not sufficient to predict trends in the annual number of physician visits. One must also identify the demographic components important in determining major differences in physician utilization rates. As the important components of the total population change in size, so will the overall rates of physician utilization and the projection of total services demanded. This method should be useful, for example, in projecting the effects of Medicare that will shift population components of known age, sex, and education composition into an economic level where financial barriers are virtually removed.

Correlations and Interactions

Unfortunately, associations among variables may hamper evaluation. Such associations take two forms, usually described as correlations and interactions. Correlations arise from unevenness in the distribution of observations. If, for example, urban areas tend to have relatively large numbers of high income residents, it would not be immediately clear whether observed rates of doctor usage resulted from conditions of urban living or from income differential. Indeed the two forces could act in opposite directions and obscure the real influence of each. Interactions of variables produce effects greater or less than anticipated from examination of each variable separately (analogous to the phenomenon of antagonism and synergism). To illustrate, the combination of high income and urban residence might produce a deviation from mean physician usage rates greater than simple addition of separate effects.

Objective

The objective here is to present a new approach to the solution of the above problems. A practical application of the techniques used in projecting health manpower needs in Taiwan provides evidence of its usefulness.

Methodology

Multiple Regression

Perhaps the most popular method for overcoming the problem of correlations among the independent variables is the multiple regression analysis. The

regression model must be expressed in detail—a requirement often incompatible with the uncertainty concerning the nature of interactions and nonlinearities in the effects.

As a result, some regression analyses omit important elements while others include any that have even the slightest possibility of being significant. With regard to the latter, various methods have been proposed for selecting a limited number of terms from among a multitude[1-6]. It is not uncommon for different methods to produce different combinations of variables[7]. In fact, one given method often produces different results from two independent samples of data.

Analysis of Variance

At the other end of the spectrum, the analysis of variance technique is extremely productive in uncovering effects and interactions whenever the correlation problem does not exist, i.e. whenever each combination of independent variables contains the same number of observations. Although this condition is seldom encountered except in controlled experiments, the principle of block-sorting employed by this technique deserves consideration under more generalized circumstances.

Suppose, for example, that three degrees of urbanization—cities, towns, and villages—are defined. The associated two degrees of freedom would be evaluated as a block. If in turn four economic levels were identified, the six $\{(3-1)(4-1)=6\}$ degrees of freedom associated with the residence-economic interaction would be investigated as a block. This would obviate the need for decision as to whether the interaction should be expressed mathematically as $(\text{residence})^2(\text{economic level})$, $(\text{residence})(\text{economic level})^3$, or in some other way.

The block-sorting principle is especially useful when the independent variables are classificatory rather than quantitative. This is typical of demographic information; even potentially continuous variables such as age are often expressed in terms of age groups. As a consequence, so-called dummy-variable techniques have been introduced into regression analyses to effect block-sorting within the general regression context[8].

Branching

More recently Sonquist and Morgan have proposed a branching technique for the segregation of dissimilar population segments[9]. The approach is such that interaction effects are identified in a relatively simple and direct manner. The sequential nature of the technique, however, causes later identification to be dependent upon the particular population classifications initially

produced. Without independent evaluations of the several factors and interactions, one must question whether the unique importance of each has been revealed.

Multi-sort Methodology

We present another approach to block-sorting for those cases where there is some information on nearly all of the possible segments, or cells, of the population. (Estimated responses can be inserted into empty cells. The estimates would be given minimal weight and, therefore, could be crude without distorting the conclusions.) This approach, the multi-sort technique, follows analysis of variance principles but is computationally simpler than the pure analysis of variance complicated by inequalities in observations per cell[10].

The analysis begins with a set of cell mean responses, X_{ij} . . . each based upon a specified number of observations, N_{ij} . . . Table 1 provides an example that will help illustrate the method for the two-variable case (each cell mean has been reduced by a constant amount in order to simplify the computation).

Table 1. HYPOTHETICAL DATA

Mean Responses (X_{ij})

	C	
	−1	6
R	−1	−3
	2	−3

Observations (n_{ij})

	C		N_i
	6	4	10
R	4	2	6
	2	2	4
N_j	12	8	N = 20

The general formulations described below are summarized in Table 2. In the formulations the following definitions are employed:

$$N_i = \sum_{j=1}^{c} n_{ij} \tag{1}$$

$$N_j = \sum_{i=1}^{r} n_{ij} \tag{2}$$

$$N = \sum_{i,j} n_{ij}. \tag{3}$$

Step 1. The Overall Effect. Calculation of any effect involves an averaging

process in which consideration must be given to the weight assigned to each component of the average. Specifically, we shall define the grand mean as

$$\overline{X} = \frac{\sum_{i,j} W_{ij} X_{ij}}{\sum_{i,j} W_{ij}} \tag{4}$$

where the weights, W_{ij}, are assigned.

It seems natural to weight the means in accordance with the numbers of observations on which they are based so that

$$W_{ij} = n_{ij}. \tag{5}$$

At this point we need only note the resulting definition of $\overline{X}$ listed at the top of Table 2.

In the illustrative case $\overline{X} = 0.3$, this amount is subtracted from each cell mean to produce a series of X'_{ij} values before proceeding to the analysis of other effects. The numerical results of these modifications are shown in Table 3.

Step 2. The Main (Row and Column) Effects. Row and Column effects are likewise weighted averages and statistical methodology suggests two possible weighting schemes. As one, we might use equal weights. This, however, would tend to place too much emphasis on imprecise mean responses based upon little information. As an alternative, we could use a weighting scheme that conforms to the usual notion of statistical adjustment. Specifically, this means that the row effect,

$$R_i = \frac{\sum_{j} W_{ij} X'_{ij}}{\sum_{j} W_{ij}}, \tag{6}$$

would be stated as

$$R_i = \frac{1}{N} \sum_{j} N_j X'_{ij}. \tag{7}$$

The problem here is that the evaluation may be distorted by interactions not removed equally from the calculations of main effects.

The latter scheme has produced more reasonable results than the system of equal weights in applications thus far. Subsequent discussion, therefore, will be based upon this weighting principle. (A multi-sort computer program has

Table 2. SUMMARY OF 2-VARIABLE MULTI-SORT COMPUTATIONS WITH ILLUSTRATIVE EXAMPLE

	Row Effect	Column Effect	Interaction
Cell Values Analyzed	$X'_{ij}=X_{ij}-\overline{X}$ $\left[\overline{X}=\frac{1}{N}\sum_{i,j} n_{ij}X_{ij} =0.30\right]$ [Tables 1 and 3]	X'_{ij}	$X''_{ij}=X'_{ij}-R'_i-C'_j$ [Table 4]
Unadjusted Effect	$R_i=\frac{1}{N}\sum_j N_j X'_{ij}$	$C_j=\frac{1}{N}\sum_i N_i X'_{ij}$	$I_{ij}=\frac{1}{N}\sum_{k,\cdot} N_{k\cdot} X''_{ijk\cdot}$
	$R_1=\frac{1}{20}[(12)(-1.3)+(8)(5.7)]$ $=1.5$ $R_2=$ -2.1 $R_3=$ -0.3	$C_1=\frac{1}{20}[(10)(-1.3)+\ldots+(4)(1.7)]=-0.7$ $C_2=$ 1.2	$I_{ij}=X''_{ij}$ [Table 4]
Effect Multiplier	$M(R_i)=\frac{N^2}{\sum_j \frac{N_j^2}{n_{ij}}}$	$M(C_j)=\frac{N^2}{\sum_i \frac{N_i^2}{n_{ij}}}$	$M(I_{ij})=\frac{N^2}{\sum_{k,\cdot} \frac{N_{k\cdot}^2}{n_{ijk\cdot}}}$
	$M(R_1)=\frac{20^2}{144/6+64/4}$ $=10.00$ $M(R_2)=$ 5.88 $M(R_3)=$ 3.85	$M(C_1)=\frac{20^2}{100/6+36/4+16/2}$ $=11.88$ $M(C_2)=$ 7.84	$M(I_{ij})=n_{ij}$ [Table 1]
Unadjusted Index of Total Differences	$S(R)=\sum_i R_i^2 M(R_i)$	$S(C)=\sum_j C_j^2 M(C_j)$	$S(I)=\sum_{i,j} I_{ij}^2 M(I_{ij})$
	$=(1.5)^2(10.00)+\ldots+(-0.3)^2(3.85)$ $=$ 48.78	$=(-0.7)^2(11.88)+(1.2)^2(7.84)$ $=$ 17.11	$S(I)=\sum_{i,j} X''^2_{ij} n_{ij}$ $=$ 132.56

Table 2 (cont.). SUMMARY OF 2-VARIABLE MULTI-SORT COMPUTATIONS WITH ILLUSTRATIVE EXAMPLE

	Row Effect	Column Effect	Interaction
Effect Mean	$\overline{R}=\dfrac{\sum_i R_i M(R_i)}{\sum_i M(R_i)}$	$\overline{C}=\dfrac{\sum_j C_j M(C_j)}{\sum_j M(C_j)}$	$\overline{I}=\dfrac{\sum_{i,j} I_{ij} M(I_{ij})}{\sum_{i,j} M(I_{ij})}$
	$=\dfrac{(1.5)(10.00)+\ldots+(-0.3)(3.85)}{10.00+5.88+3.85}$	$=\dfrac{(-0.7)(11.88)+(1.2)(7.84)}{11.88+7.84}$	$\overline{I}=\dfrac{\sum_{i,j} X''_{ij} n_{ij}}{N}$
	$=\dfrac{1.497}{19.73}$ $= 0.08$	$\dfrac{1.092}{19.72}$ $= 0.06$	$=\dfrac{0.4}{20}$ $= 0.02$
Adjusted Index of Total Differences	$S'(R)=S(R)-\overline{R}^2\sum_i M(R_i)$	$S'(C)=S(C)-\overline{C}^2\sum_j M(C_j)$	$S'(I)=S(I)-\overline{I}^2\sum_{i,j} M(I_{ij})$
	$=48.78-(0.08)^2(19.73)=$ 48.66	$=17.11-(0.06)^2(19.72)$ $=17.05$	$S'(I)=S(I)-\overline{I}^2N=$ 132.55
Index of Mean Differences	$D(R)=\dfrac{S'(R)}{r-1}$	$D(C)=\dfrac{S'(C)}{c-1}$	$D(I)=\dfrac{S'(I)}{(r-1)(c-1)}$
	$=\dfrac{48.66}{2}$ $=24.33$	$=\dfrac{17.05}{1}$ $=17.05$	$=\dfrac{132.55}{(2)(1)}$ $=66.28$
Adjusted Effect	$R'_i=R_i-\overline{R}$	$C'_j=C_j-\overline{C}$	$I'_{ij}=I_{ij}-\overline{I}$
	$R'_1=1.50-0.08$ $= 1.42$ $R'_2=$ -2.18 $R'_3=$ -0.38	$C'_1=-0.70-0.06$ $=-0.76$ $C'_2=$ 1.14	

been written to permit computations from both weighting schemes for purposes of comparison.) The values for R_i produced by the hypothetical data are shown in Table 2, as are analogous computations for C_j.

Step 3. Indices of Differences for Main Effects. The magnitude of differences among the R_i and the C_j is next measured along the lines of the mean square of analysis of variance. Once again the problem of weighting arises, for each of the effects is based upon different quantities of information. We have chosen to use the scheme employed in the analysis of variance for investigating the main effects with unequal numbers of observations per cell. Contrary to the analysis of variance method, we apply this scheme throughout the evaluation of interactions as well.

We then have weights (or multipliers) $M(R_i)$ and $M(C_j)$ as indicated in Table 2. In general, larger numerical results arise where estimates of effects are based upon large numbers of observations, especially if those observations are approximately uniformly distributed among cells.

Applying these weights to the unadjusted effects, R_i and C_j, we obtain crude indices of total differences $S(R)$ and $S(C)$ respectively. The crude indices must be adjusted for deviations from the means $\overline{R}$ and $\overline{C}$ to produce the final indices of total differences, $S'(R)$ and $S'(C)$.

Table 3. HYPOTHETICAL DATA ADJUSTED FOR OVERALL EFFECT

$(X'_{ij} = X_{ij} - \overline{X})$

	C	
	−1.3	5.7
R	−1.3	−3.3
	1.7	−3.3

Based upon the number of degrees of freedom available to each effect, we proceed from the indices of total differences to the indices of mean differences —much as the analysis of variance moves from sums of squares to mean squares. In the illustrative example we see some indication that the row effect is greater than the column effect.

As a final step prior to the investigation of interaction effects, we must adjust the main effects R_i and C_j for the means $\overline{R}$ and $\overline{C}$. The adjusted effects are then removed from the cells to produce a series of modifications like those of Table 4.

Table 4. HYPOTHETICAL DATA ADJUSTED FOR OVERALL AND MAIN EFFECTS

$$(X''_{ij} = X'_{ij} - R'_i - C'_j)$$

	C	
	−1.96	3.14
R	1.64	−2.26
	2.84	−4.06

Step 4. The Interaction Effects. Paralleling the preceding analysis, the interaction evaluation requires the application of appropriate multipliers to estimated effects to obtain an index of total differences. As before, this index is modified according to the number of degrees of freedom available to produce an index of mean differences—our ultimate concern.

In the two-variable case the unadjusted effects are simply the cell residuals X''_{ij} found in Table 4, and the effect multipliers are merely the numbers of observations n_{ij} recorded in Table 1.

In the interest of a more general understanding of the procedure, the two-variable interaction formulations of Table 2 are presented as if three or more variables were under consideration. The subscript k refers to the third variable. We introduce the notation N_k to symbolize the total number of observations obtained at the *ith* level of the first variable and the *jth* level of the second variable. This notation is a direct extension of our definition of N_i as a sum of all observations in the *ith* row regardless of the index j. Of course, with three or more variables the definition of N_i would become

$$N_i = \sum_{j,k} n_{ijk} \tag{8}$$

Step 5. Index of Differences for Interaction. As before, the unadjusted index of total differences is a sum of squared unadjusted effects appropriately weighted. In our example we simply square the entries of Table 4, multiply the result by the corresponding number of observations recorded in Table 1, and sum the products. After making the necessary adjustment for the mean, we apply the available degrees of freedom to obtain an index of mean differences. In this case the result, 66.28, indicates that the interaction effect is larger than either of the main effects.

If three or more variables were under scrutiny, we would first analyze each of the two-variable interactions. Then we would obtain an adjusted estimate of each interaction effect so that these effects could be removed from Table 4 prior to an investigation of three-variable interactions.

Step 6. Variance Ratios. The crux of the analysis at each stage is compilation of indices of mean differences which can be compared with a residual index of high order interaction mean differences or with an error variance obtained independently. In any event, a series of ratios is derived for interpretation as F-ratios are in the analysis of variance. We must recognize, of course, that these ratios are not precisely distributed as F.

Results of Method Applied in a Study of Physician Visits in Taiwan

As indicated earlier, we have used the multi-sort technique to evaluate differences in physician utilization rates among various segments of the Taiwanese population. A survey of 66,000 individuals provided data for the 288 cells which result from four economic, three education, four age, three residence (city, town, village), and two sex categories. (Observations appeared in 285 cells. Responses for three were estimated.)

The indices of mean differences are summarized in Table 5. Interactions involving more than three variables have been judged meaningless. They have all been lumped together to provide a residual index of 8.1 with 132 degrees of freedom. By dividing each of the indices of mean differences by this residual we obtain ratios which can be treated as F-values for testing the significance of the effects. Tables of critical F-ratios clearly verify the influence of only four effects in annual visits per capita (V): age group (A), economic class (W), residence (R), and the economic-age interaction.

At this stage we found the multiple regression technique extremely useful, for the preceding analysis had given a good indication of the nature of the regression model. Specifically, we hypothesize that, predicted

$$V = B_0 + B_1A + B_2A^2 + B_3W + B_4R + B_5AW. \qquad (9)$$

Numerically the coefficients which resulted were, predicted

$$V = 12.47 - 11.84A + 2.68A^2 + 1.87W + 0.72R - 0.42AW \qquad (10)$$

Squared terms involving wealth and residence could have been included but they produced a negligible reduction in the standard error of estimate from 0.164 to 0.162. (Standard errors are based upon the cell means, not the individual observations.)

Discussion: Comparison with Other Methods

Other methods of anaylsis of multiple variables have been used in the health field[11,12]. Results of application of some of these methods to our hypothetical example and Taiwan data are discussed.

Table 5. ANALYSIS OF MEAN DIFFERENCES (TAIWAN DATA)

Block	Index of Total Diff. (1)	Degrees of Freedom (2)	Index of Mean Diff. (3)	$(4) = \frac{(3)}{8.1}$
Economic (W)	433.2	3	144.4	17.8
Education (E)	22.4	2	11.2	1.4
Age (A)	2,081.2	3	693.7	85.6
Residence (R)	156.4	2	78.2	9.7
Sex (S)	5.5	1	5.5	0.7
W-E	44.2	6	7.4	0.9
W-A	272.3	9	30.3	3.7
W-R	26.6	6	4.4	0.5
W-S	11.8	3	3.9	0.5
E-A	85.1	6	14.2	1.8
E-R	41.9	4	10.5	1.3
E-S	6.7	2	3.4	0.4
A-R	29.2	6	4.9	0.6
A-S	5.2	3	1.7	0.2
R-S	35.6	2	17.8	2.2
W-E-A	169.4	18	9.4	1.2
W-E-R	81.4	12	6.8	0.8
W-E-S	28.3	6	4.7	0.6
W-A-R	86.0	18	4.8	0.6
W-A-S	38.7	9	4.3	0.5
W-R-S	28.1	6	4.7	0.6
E-A-R	78.6	12	6.6	0.8
E-A-S	106.3	6	17.7	2.2
E-R-S	32.6	4	8.1	1.0
A-R-S	90.3	6	15.1	1.9
Other	1,069.2	132	8.1	

Analysis of Variance

Although the multi-sort technique has much in common with analysis of variance, there are important differences. Specifically, the multi-sort method offers a streamlined computational procedure, but in doing so, certain features of orthogonality are lost.

For a more concrete comparison of the two methods we return to the hypothetical two-variable illustration presented earlier. With respect to main effects, the indices of mean differences are identical to the analysis of variance mean squares. This is no accident, for the computational procedures of the two methods are alike at this point.

The methods differ somewhat in their assessment of interaction effects. It is of interest, therefore, to compare the two formulations. Although the multi-sort approach has already been described in what seems to be the most straightforward manner, Table 6 presents an algebraically equivalent set of computations which more nearly correspond in form to the analysis of vari-

Table 6. COMPARISON OF INTERACTION EVALUATION IN MULTI-SORT AND ANALYSIS OF VARIANCE

	Multi-Sort Index of Mean Differences	Numerical Example	Analysis of Variance Mean Square	Numerical Example
Total Sum of Squares	$\sum_{i,j} n_{ij}X_{ij}^2$	198.00	$\sum_{i,j} n_{ij}X_{ij}^2$	198.00
Reduction for Other Effects				
Overall	$N\overline{X}^2$	− 1.80	$\left[\sum_i n_{i1}X_{i1} - \sum_i \left(\frac{n_{i1}\sum_j n_{ij}X_{ij}}{N_i}\right)\right]^2$	− 17.62
Column	$\left[\frac{(\sum_j N_jC_j')^2}{N} - \sum_j N_jC_j'^{\,2}\right] + 2\left[\sum_j C_j' \sum_i n_{ij}X_{ij} - \overline{X}\sum_j N_jC_j'\right]$	− 19.15	$\sum_i n_{i1} - \sum_i \frac{(n_{i1})^2}{N_i}$	
Row	$\left[\frac{(\sum_i N_iR_i')^2}{N} - \sum_i N_iR_i'^2\right] + 2\left[\sum_i R_i' \sum_j n_{ij}X_{ij} - \overline{X}\sum_i N_iR_i'\right]$	− 47.23	$\sum_i \left[\sum_j \frac{(n_{ij}X_{ij})^2}{N_i}\right]$	− 50.07
Row and Column	$2\left[\frac{\sum_i N_iR_i' \sum_j N_jC_j'}{N} - \sum_i R_i' \sum_j n_{ij}C_j'\right]$	+ 2.74		
Result		132.56		130.31

ance. The latter formulation is deceptively simple here because we are dealing with only two variables, one of which has only two levels. This special case offers a convenient basis for comparison.

The multi-sort approach does not completely isolate the two main effects, but we are able to confirm the consistent comparability of the two sets of numerical results, even though the algebraic representations are somewhat different.

Multiple Regression

In analyzing the Taiwan data we used the multi-sort technique to sort out the important variables. The relationships were then made more explicit by means of a multiple regression equation. How does this compare with what might have been obtained had the multiple regression approach been employed from the outset?

We shall consider it impractical to deal with a regression equation of more than twenty terms. Within this limitation we can handle the constant term, the first degree terms associated with each of the five demographic variables, four squared terms (excluding sex), and the ten interactions: AW, AR, WR, . . . Considering the possibility that some terms can be safely omitted, we have more than half a million combinations from which to choose. Even if we refuse to admit more than five variables plus the constant, we face 16,663 possibilities.

$$C_1^{19} + C_2^{19} + \ldots + C_{19}^{19} = 524{,}287.$$

This is a common dilemma in regression analyses and forces one to adopt a selection algorithm. A popular one is the stepwise approach whereby variables are admitted only as long as each newcomer succeeds in substantially increasing the multiple correlation coefficient.

Results obtained from the stepwise approach in the present case can be seen in Table 7. Although age is unquestionably the most important determinant of physician utilization, it was not the first to enter the equation. It did enter early, but so did other terms that are conspicuously absent from the multi-sort model.

The importance of the other terms, moreover, was not confirmed by the complete 19-variable regression model. From the complete model we obtained 19 t-values as ratios of the partial regression coefficients to their standard errors. Table 7 records the rank-order of the t-values for comparison with their step-wise order of entry.

Table 7. STEPWISE REGRESSION ANALYSIS (TAIWAN DATA)

Variable	Order of Entry	Final Rank-Order of t-Value
Economic—Residence (WR)	1	14
Education—Age (EA)	2	9
Age Squared (A^2)	3	1
Age (A)	4	2

Dummy-variable Approach

We alluded earlier to dummy-variable modifications of multiple regression analysis. One of these, the Lippitt method, was chosen to provide an additional basis for comparison with the multi-sort technique[13]. As the Lippitt technique uses the dummy-variable approach, it also handles blocks of variables. It is not especially geared to the handling of interactions, however.

The Lippitt results were in agreement with the multi-sort conclusions with three exceptions. First, the Lippitt technique did not detect the economic-age interaction because it did not seek it out. Secondly, the iteration procedure employed by the Lippitt method encountered difficulty in achieving convergence with respect to the sex variable, although no importance could be discerned for this factor.

Finally, the Lippitt results attached some importance to education, suggesting that more highly educated individuals tend to make fewer doctor visits. To further test this possibility, we formed a multiple regression equation which added the education term to equation (9). This term reduced the standard error of estimate from 0.164 to 0.163, an even smaller reduction than the slight one obtained from the introduction of non-linear wealth and residence terms.

Summary

An overview of an entire population can be uninformative, or even misleading. Recognizing the need to view components separately, manifests the analytical problem of identifying the subdivisions that are useful in a given investigation.

The multi-sort technique is introduced as a possible approach in cases in which some information is available on nearly all of the potentially important cells of a population. Employing computational procedures that are simple enough to be practicable in a wide variety of applications, the technique is based upon the principle of block-sorting of variables. After the key variables have been identified, multiple regression equations will often be useful in making the important relationships explicit.

References

1. Larson, H. J. and T. A. Bancroft. Sequential model building for prediction in regression analysis I. *Ann. Mathematical Statistics 34*:462, 1963.
2. Larson, H. J. and T. A. Bancroft. Biases in prediction by regression for certain incompletely specified models. *Biometriks 50*:391, 1963.
3. Mallows, C. L. *Choosing variables in a linear regression: a graphical aid.* Paper presented at the Central Regional Meeting of TIMS, Manhattan, Kansas. May, 1964.
4. Schultz, E. F., Jr. and J. F. Goggans. A systematic procedure for determining potent independent variables in multiple regression and discriminant analysis. *Agr. Exp. Sta. Bull. 336.* Auburn University, 1961.
5. Selvin, H. C. and A. Stuart. Data-dredging procedures in surgery analysis. *Amer. Statistician 20* (3):20, 1966.
6. Summerfield, A. and A. Lubin. A square root method of selecting a minimum set of variables in multiple regression: I, the method. *Psychometrika 16* (3):271, 1951.
7. Gorman, J. W. and R. J. Toman. Selection of variables for fitting equations to data. *Technometrics 8* (1):27, 1966.
8. Suits, D. Use of dummy variables in regression equations. *Amer. Statistical Assoc. 52*: 548, 1957.
9. Sonquist, J. A. and J. N. Morgan. *The detection of interaction effects.* Survey Research Center, University of Michigan, Ann Arbor, 1964.
10. Scheffe, H. *The analysis of variance.* New York: John Wiley and Sons, Inc. 1959.
11. Feldstein, P. J. and W. J. Carr. *The effect of income on medical care spending.* Paper presented at annual meeting of The American Statistical Association, December, 1964.
12. Rosenthal, G. D. *Factors affecting the utilization of short-term general hospitals.* Paper presented at annual meeting of The American Public Health Association, New York, October, 1964.
13. Lippitt, V. *Determinants of consumer demand for house furnishings and equipment.* Cambridge: Harvard University Press. 1959.