WORLD FERTILITY SURVEY

TECHNICAL BULLETINS

OCTOBER 1976

NO. 1/TECH. 441

Some Notes on Statistical Problems Likely to Arise in the Analysis of WFS Surveys

SIR MAURICE KENDALL

INTERNATIONAL STATISTICAL INSTITUTE
Permanent Office · Director: E. Lunenberg
Prinses Beatrixlaan 428
Voorburg, The Hague,
Netherlands

WORLD FERTILITY SURVEY
Project Director:
Sir Maurice Kendall, Sc. D., F.B.A.
35-37 Grosvenor Gardens
London SW1W OBS, U.K.

The World Fertility Survey is an international research programme whose purpose is to assess the current state of human fertility throughout the world. This is being done principally through promoting and supporting nationally representative, internationally comparable, and scientifically designed and conducted sample surveys of fertility behaviour in as many countries as possible.

The WFS is being undertaken, with the collaboration of the United Nations, by the International Statistical Institute in cooperation with the International Union for the Scientific Study of Population. Financial support is provided principally by the United Nations Fund for Population Activities and the United States Agency for International Development.

This paper is one of a series of Technical Bulletins recommended by the WFS Technical Advisory Committee to supplement the document *Strategies for the Analysis of WFS Data* and which deal with specific methodological problems of analysis beyond the Country Report No. 1. Their circulation is restricted to people involved in the analysis of WFS data, to the WFS depositary libraries and to certain other libraries. Further information and a list of these libraries may be obtained by writing to the Information Office, International Statistical Institute, 428 Prinses Beatrixlaan, Voorburg, The Hague, Netherlands.

SOME NOTES ON STATISTICAL PROBLEMS LIKELY TO ARISE IN THE ANALYSIS OF WFS SURVEYS

WFS/TECH.441

Prepared by:

SIR MAURICE KENDALL
WFS Project Director
International Statistical Institute
35-37 Grosvenor Gardens
LONDON SW1W OBS
United Kingdom

CONTENTS

INTRODUCTION

A full analysis of a WFS survey will probably require statistical expertise of quite varied kinds, from simple tabulation and number-crunching to sophisticated mathematical techniques. It is not possible in this memorandum to discuss the full methodology which may be needed - to do so would require at least one text-book. The following discussion is accordingly confined to some of the major problems which are foreseen at this stage, and sounds some warnings about the dangers of an undiscriminating use of certain accepted statistical routines.

Much of the analytical process consists of fitting models to the data, or of seeing whether the data are consonant with hypotheses which the demographer advances for testing. The statistician tends to look for models by examining the data, without in many cases having a prior notion of the causality of the system, although he has to examine the logical consistency of his assumptions; the demographer tends to approach the analysis with a background of possible hypotheses which are suggested from his previous knowledge and experience. But both are really aiming at the same objective: the construction of an explanatory model. And although 'explanation' is a relative term and 'causality' an elusive concept, it seems evident that the most productive outcome of WFS studies will arise from continual dialogue between the statistician and the demographer. This document deals with some of the statistical topics which will need to be discussed between them.

CATEGORIZED VARIABLES IN REGRESSION ANALYSIS

1. In many demographic contexts it is required to regress a variable y (the regressand) on a set of variables $x_1, x_2 \ldots x_p$ (the regressors) where some of the x's are not continuous variables but are categorized. For example, the individuals under study may be dichotomized by sex, classified by an ordered grouping, say by educational standard or by social class, or classified by an unordered grouping, say by religion or ethnicity. It is sometimes proposed that in such cases the discontinuous classes can be represented by a pseudo-variable; for example,male and female by a (1,0) variable; favourable, neutral and unfavourable attitudes by a three-way variable (+1,0,-1); three religious groups A, B, C by three variables, one bearing the value 1 if the subject is an A (and zero otherwise), one bearing the value 1 if the subject is B (and zero otherwise), one bearing the value 1 if the subject is C (and zero otherwise). Other variations are possible.

2. These pseudo-variables, especially the dichotomies, are often referred to as 'dummies'. This is not a very exact usage because, strictly speaking, a dummy should remain constant, but it is fairly deeply embedded in the literature and has the merit of brevity.

3. The question for examination is whether these pseudo-variables can be fed into an ordinary least-squares regression analysis and produce meaningful results. The situation in general is by no means straightforward. Consider first of all the simple case where a regressand variable Y is regressed simply on one continuous regressor X but the individuals are classified by sex. A naive (but commonly encountered) approach would be to analyze the model

$$Y = \beta_0 + \beta_1 X + \beta_2 Z + \varepsilon \tag{1}$$

where Z is the sex-variable, say equal to 1 for men (n_1 in number) and 0 for women (n_2 in number). A straightforward least-squares analysis leads to the estimators

$$b_0 = \bar{y} - b_1\bar{x} - b_2\bar{z} \tag{2}$$

$$b_1 = \frac{n_1 \operatorname{cov}_1 (y,x) + n_2 \operatorname{cov}_2 (y,x)}{n_1 \operatorname{var}_1 x + n_2 \operatorname{var}_2 x} \tag{3}$$

$$b_2 = \bar{y}_2 - b_1\bar{x}_2 - b_0 \tag{4}$$

where the bars denote means of the observations and the subscripts 1 and 2 applied to their x's refer to the male and female categories respectively.

Now had we analyzed the two groups separately by the same procedure we should have obtained

for the male group:

$$b_1 \text{ (male)} = \frac{\operatorname{cov}_1 (y,x)}{\operatorname{var}_1 x} \tag{5}$$

and for the female group:

$$b_1 \text{ (female)} = \frac{\text{cov}_2 (y,x)}{\text{var}_2 x} \tag{6}$$

Comparison with equation (3) then shows that the regression coefficient b_1 for the two groups together is a weighted average of the coefficient obtained by treating the two groups separately.

4. The effect of the dummy variable has therefore been to average two relationships which may be quite different. It would clearly be more informative to keep these relationships distinct,unless it can be shown that they are sufficiently alike to justify amalgamation.

5. A more elaborate model requires the addition to equation (1) of an 'interaction' term XZ, so that the model becomes

$$Y = \beta_0 + \beta_1 X + \beta_2 Z + \beta_3 XZ + \varepsilon \tag{7}$$

A least-squares solution now gives

$$\begin{aligned} Y = \bar{y}_2 &- \frac{\text{cov}_2 (y,x)}{\text{var}_2 x} + \frac{\text{cov}_2 (y,x)}{\text{var}_2 x} X \\ &+ \left\{ \bar{y}_1 - \bar{y}_2 + \frac{\text{cov}_2 (y,x)}{\text{var}_2 x} \bar{x}_2 - \frac{\text{cov}_1 (y,x)}{\text{var}_1 x} \bar{x}_1 \right\} Z \\ &+ XZ \left\{ \frac{\text{cov}_1 (y,x)}{\text{var}_1 x} - \frac{\text{cov}_2 (y,x)}{\text{var}_2 x} \right\} \end{aligned} \tag{8}$$

If $x_2 = 1$ this becomes

$$Y = \bar{y}_1 - \frac{\text{cov}_1 (y,x)}{\text{var}_1 x} \bar{x}_1 + \frac{\text{cov}_1 (y,x)}{\text{var}_1 x} X \tag{9}$$

the ordinary regression of y on x_1 in the male group. And similarly if $x_2 = 0$ we get the ordinary regression in the female group. The pseudo-variable x_2 has, in appearance, amalgamated the two groups in equation

(8) but in fact it only provides a summary expression for the two (possibly different) relationships of which (9) is one.

6. Equation (8) does, however, provide a test whether the two relationships have the same β_1. In fact, the coefficient of the last term in that equation is the difference of the estimated β_1's for the two groups. If it is zero or negligibly small the two regressions have the same slope and can be amalgamated so far as concerns b_1. They may still have different values of the estimated β_0, i.e.,may be represented by parallel lines.

7. Similar effects will appear when a pseudo-variable consists of more than two classes or when there are several pseudo-variables. Two approaches are possible: to keep the regression lines within each category distinct before attempting any kind of amalgamation; or to write down a full interactive model, examine whether any terms can be discarded, and use the result to derive the individual regressions. However,

(i) If there are many categorized variables the frequencies in the sub-categories may become small, so small that the regressions within them are subject to so much sampling variability as to be unreliable. For example, a sample of 5,000, divided by sex, three ethnic groups, four educational categories and five geographical regions, giving a total of 2 x 3 x 4 x 5 = 120 sub-categories, would only have an average sample number within categories of 42, and some would be smaller than that. It may be worthwhile melding some of the categories to achieve greater sampling reliability, as a trade-off against the possibility of merging relationships which are not identical.

(ii) Sometimes it may be of no interest to keep categories distinct. If, for example, we have a sample of women and regress their fertility (e.g.,numbers of children ever born) on income in four different geographical areas, and if the sample is representative as to numbers in these regions, and if we are concerned only with the relation between fertility and income <u>for the whole area</u>, then one relation may be sufficient.

(iii) Sometimes ordered categorizations can, to a satisfactory degree of approximation, be represented by rank numbers, which are then treated as ordinary variables. Suppose, for example, we have 1,000 individuals classified by social grade as follows, A being the highest.

A	B	C	D	E
50	150	500	200	100

If this were a ranking of 1,000 people we might regard the first 50 as ranked from 1 to 50 and allocate to each member the average of those ranks $\frac{1}{50}(1 + 2 \ldots + 50) = 25.5$, treating them as tied ranks. The next group would each have a score $\frac{1}{150}(51 + \ldots 200)$ = 125.5 and so on. More sophisticated scores can be assigned in such cases but they depend on some assumption about the distribution which has given rise to the observed categorization.

(iv) It may, nevertheless, be of interest to work with the equation of type (7), where a number of dummies are concerned, and to subject it to one of the regression routines which reject non-contributory variables, in order to arrive at the most parsimonious representation. This procedure has to be followed with caution and continual regard to the realities of the situation.

8. As an example of the 'averaging' effect which can result from a dummy variable, the following data relate to the WFS survey in Fiji. The age at marriage y is regressed on age of wife (x_1), her years of education (x_2), and race (x_3), (Fijian = 0, Indian = 1). The result for the whole group was

$$y = 0.09x_1 + 0.30x_2 - 0.18x_3 \qquad (10)$$

the variables being measured about their mean.

For the Fijian group alone the result was

$$y = 0.20x_1 + 0.16x_2 \qquad (11)$$

and for the Indian group alone

$$y = -0.04x_1 + 0.33x_2 \qquad (12)$$

Clearly the relationships expressed by (11) and (12) are quite different and are obscured if they are run together in equation (10).

This example, though based on real data with nearly 5,000 cases in (10), more than 2,500 in (12) and more than 2,000 in (11), is advanced for illustration purposes only. A study in greater depth would be required to explore the relation of age at marriage with other variables.

RELATIONSHIPS AMONG REGRESSOR VARIABLES

9. In an equation such as

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + \varepsilon \qquad (13)$$

it has been customary to refer to Y as the 'dependent' variable and the X's as 'independent' variables. It is relatively rare for the X's to be independent in the statistical sense and quite often they are highly correlated. This creates special problems in the interpretation of such an equation and in particular the relative contributions to Y of individual X's.

To avoid the tendentious word 'independent' to describe mutually dependent variables, it is also customary to call Y the 'explained' and the X's the 'explanatory' variables. This is an improvement in terminology but is not entirely free from objection because the 'explanatory' variables may not themselves account for the variation of Y but may only be linked to it by some circuitous causal mechanism. An entirely neutral terminology is to call Y the regressand and the X's regressor variables.

10. It is therefore desirable to explore the relations among the X's before entering on a regression analysis. This requires a somewhat sophisticated approach which is difficult to summarize in non-technical terms. An examination of the individual correlations between pairs of

variables is not sufficient. What is required is an analysis of the whole set of correlations or covariances. One of the best methods is to compute the covariance or correlation matrix of the regressors and to determine the constants known as latent roots or eigenvalues. Any zero eigenvalue will imply a linear relation among some of the X's and therefore a redundancy among them. A small eigenvalue indicates near-collinearity among the X's and warns that the coefficients b, the estimators of β, will individually be unreliable.

In fact the estimators b, in matrix terms are,

$$\underset{\sim}{b} = \underset{\sim\sim}{yx}^{1} \; (\underset{\sim\sim}{xx}^{1})^{-1} \tag{14}$$

where x is the $p \times n$ matrix of observations x (p variables, n observations) x^1 is its transpose, y is the $1 \times n$ vector of observations of y and the covariance matrix accordingly (with x's measured about their means) is $(xx)^{-1}$. The fact that this matrix appears as an inverse implies that if it has a small determinant (corresponding to one or more small eigenvalues) the estimators b will be inflated and individually unreliable. Should this situation arise, it is preferable to delete some of the variables.

11. An example, also drawn from the Fijian data, will serve to illustrate the point. The regressand variable y is the parity (number of children). The regressor variables are age of mother in years, x_1; years of education of mother, x_2; desired family size, x_3; and marital duration x_4.

The correlation matrix of the regressor variables is as follows

	x_1	x_2	x_3	x_4
1	1.000			
2	.914	1.000		
3	.500	.548	1.000	
4	-.321	-.430	-.280	1.000
Corr. with y	.635	.700	.774	-.342

The regressors are highly correlated and a principal component analysis gives for the eigenvalues

Component	Eigenvalue % of total	Cumulative %
1	2.56 (64%)	64.0
2	0.77 (19%)	83.3
3	0.59 (15%)	98.0
4	0.08 (2%)	100.0

The smallness of the least eigenvalue indicates that the four regressors are nearly collinear and warns us that the coefficients of a regression of y on them are completely unreliable.

In fact the regression of y on all four (not measured about their means) is

$$y = -.006x_1 - .015x_2 + .842x_3 + .130x_4 - 1.123 \qquad (15)$$
$$R^2 = 0.77$$

The regression on x_1 x_2 and x_3 is

$$y = .145x_1 - .077x_2 + .809x_3 - 3.606 \qquad (16)$$
$$R^2 = .74$$

The regression on x_2, x_3, x_4 is

$$y = -.021x_2 + .744x_3 + .159x_4 \qquad (17)$$
$$R^2 = .77$$

So far as concerns the goodness of fit, as measured by the square of the multiple correlation coefficient R^2, (15) (16) and (17) are about as good as one another. Clearly, no precise meaning can be assigned to the individual coefficients.

12. It is important to realize the force of the argument here. The classical inference drawn from equation (13) would be, for example, that if $X_2 \ldots X_p$ stay fixed, a variation of δ in X_1 would entail a variation of $\delta\beta_1$ in Y. This is true, but in many cases is not relevant. Since the X's are inter-correlated a variation in X_1 will, in general, imply variations in other X's so they will not, in fact, remain constant.

13. It follows that unless the regressors are uncorrelated or only very weakly correlated, no particular meaning can be attached to the individual coefficients in a regression equation. It is the equation as a whole which matters, namely, the excellence of fit as judged by the size of the multiple correlation coefficient R^2. It follows further that in general we cannot use those coefficients to measure the relative contribution to Y of the individual regressors. A failure to appreciate this point has impaired a good deal of published regression analysis.

14. It will naturally be asked whether, if there are dependencies among the regressors, it is possible to sort out their relative contribution to the regressand. The answer, in general, is in the negative so far as concerns the regression technique alone. Further progress in the direction of causal explanation requires the setting up of a causal model for analysis, as to which see Technical Bulletin Number 2.*

GROUPING EFFECTS

15. It frequently happens in demographic work that the members under study are grouped into frequency classes. For example, in a set of 5,000 women, it would be standard practice to group them into age categories, say 15-19, 20-24, 25-29 and so on. The question arises whether correlations or regressions, based on such data are seriously different from what they would be if the data were not grouped. The subject has been extensively discussed in a rather sophisticated way by Haitovsky (1973).

16. Consider first the correlation between two variables x_1 and x_2 which are grouped respectively in intervals of h_1 and h_2. Computation of the variance of the grouped data (the value in any group being taken as the point of that group) exaggerates the true (ungrouped) variance by an amount which is, nearly enough, represented by a corrective term

* M.G. Kendall and C.A. O'Muircheartaigh, "Path Analysis and Model Building"

known as Sheppard's.

$$\text{var(ungrouped)} = \text{var(grouped)} - h^2/12 \tag{18}$$

On the other hand the covariance needs no such correction.

Thus the estimated correlation for ungrouped data is

$$\frac{\text{cov } (x_1, x_2)}{\left\{\text{var } x_1 \quad \text{var } x_2\right\}^{\frac{1}{2}}} \tag{19}$$

where var x_1 and var x_2 are calculated from ungrouped material.

If we worked from grouped material without correction the denominator in (19) would be too large and the correlation should be

$$\frac{\text{cov } (x_1, x_2)}{\left\{(\text{var } x_1 - h_1^2/12)(\text{var } x_2 - h_2^2/12)\right\}^{\frac{1}{2}}} \tag{20}$$

In (20) var x refers to the estimated variance from the grouped material. Without grouping corrections, therefore, the calculated correlation will be too small, to an extent which depends on the coarseness of the grouping grid. The grouping effect is not uniform, but depends on the frequency distribution of the X's. Cases can occur where the attenuation due to grouping is reversed.

17. It would appear better practice, then, to work on ungrouped data where possible. Similar considerations apply to regressions. In general, grouping sacrifices information and may distort relations between variables.

18. Another rather subtle effect appears when grouping is carried out. In the classical model of equation (13) it is assumed that the random residual ε is homoscedastic, that is to say has the same variance regardless of the value of Y. When observations in Y are grouped this property tends to be lost because a set of Y's, say n, in number, clustered together at a single point have a variance dependent on n; so that, if the numbers in class-frequencies differ (as they almost always will) the error term has different variances at different points of the range of Y. This is another

reason for working with ungrouped data.

19. In the monograph under reference Haitovsky examines the possibilities (a) of restoring equal variances to grouped data by a linear transformation and (b) of estimating regression coefficients when only marginal frequencies are available, for example, in the case of two variables, when the full classification is not available but a classification of each by itself is given. There are some serious pitfalls in this part of the subject, although very often, as when one is working from published tables, no better resources are available. In the context of the World Fertility Survey it would seem desirable to work with ungrouped data where means, variances, correlations, regressions,or similar types of statistics are under study.

MISSING VALUES

20. In ordinary tabular work missing values can be handled without imputation by columns headed 'not available' or 'not given' or some similar category. For more sophisticated routines involving multivariate analysis missing values are a nuisance and it is desirable to have a systematic way of dealing with them. This also is a subject with some serious pitfalls.

21. To fix ideas, suppose we are regressing Y on four variables X_1 to X_4 and that some of the Y's and some values of the X's are missing. One simple way, of course, is to ignore all the records which are not complete. But this may involve the sacrifice of a good deal of information. Another way is to hunt through the complete records and find one which matches the incomplete record on the surviving information, and to replace the missing values by those in the matching record: the so-called 'hot-deck' method. A further procedure is to replace the missing values by numbers chosen at random within the permissible range of the variable which is missing: the 'cold-deck' method. Both hot- and cold-deck methods are methods of imputation, to which there may be objection on ethical or political grounds and in any case require a fairly large supply of complete data to provide the necessary matches.

22. More sophisticated methods are available which attempt to use all the existing information, including that in the incomplete records, by estimating the missing values from the surviving complete records. For example, if we have a number of complete records with the values of X_1

to X_4 known, we might regress X_4 on X_1 to X_3 and use it to estimate X_4 in those cases where X_4 is missing but X_1 to X_3 have survived. The subject has been reviewed by Beale and Little (1975) who discuss six different approaches and come to the conclusion, on the basis partly of theory and partly of simulation studies, that the best procedure is one which they describe as modified maximum likelihood. Effectively it is one of iteration to convergence. The complete records are used to estimate the means and covariances of all the variables. This result is used to estimate the missing quantities, which are substituted and the estimation of means and covariances repeated; and so to convergence.

23. One important danger to be avoided is the use in a single analysis of estimates of means and covariances from different sample sizes. For example, if of 1,000 records there are 900 cases where X_1 and X_2 survive, it appears plausible to calculate the means and covariance of X_1, X_2 from those values; and if there are 950 cases where X_1 and X_3 survive to do the same; and so on. The different covariances can then be substituted into covariance matrix and the resulting equations of regression theory solved. This procedure can be disastrous if the missing values are not a random set (as for instance if high incomes tend to be omitted). Haitovsky (1968) constructed 100 observations according to the formula.

$$Y = 150 + 5.0X_1 - 2.0X_2 + 0.3X_3 + 3.0X_4 + \varepsilon \qquad (21)$$

where the X's were correlated normal variables.

He then rejected 6 y's, 25 x_1's, 15 x_2's, zero x_3's and 10 x_4's. The procedure was repeated seven times, with the same number of rejections but differently rejected members. The x_1's were rejected partly systematically, ten from the 25 highest values, the other 15 at random. The average results, based on estimating covariances from different sample numbers, were as follows:

	Constant	x_1	x_2	x_3	x_4
True values	150.0	5.0	-2.0	0.3	3.0
Ordinary least squares on full 100 values	150.732	4.968	-1.922	0.514	2.922
Estimated as above	414.443	4.116	-0.660	-6.582	2.699

MULTIVARIATE CONTINGENCY TABLES

24. In the past, material presented to exhibit relationships has usually been in the form of two-way tables, especially for data which are classified into categories. Sometimes three-way, or even four-way tables have been given, especially where the classification is simple (e.g., dichotomy by sex). But difficulties of tabulation, printing and especially interpretation have prevented, or at least restricted, tabulations by more than two variables at a time.

25. During the past two decades a good deal has been learnt about these multiple-way tables and machine methods of analyzing them. Several programs are now available for the purpose. In particular, programs by Goodman (Chicago), Nelder (London) and Brown (Los Angeles) have been specifically designed to this end. Further information about them can be supplied by the WFS on request.

26. These programs cannot, however, be applied blindly and some acquaintance with the underlying theory is desirable if the best use is to be made of them. There exists an extensive literature on the subject. Reference may be made to Plackett's monograph (1974) on Multivariate Contingency and to one of the chapters in Kendall's book (1975) on Multivariate Analysis for a convenient summary; a more comprehensive treatment is given in Bishop *et al* (1975). These books are the most recent but the subject is still developing fairly fast. The remaining part of this section describes very briefly the kind of problems which arise.

27. Contingency tables may consist of ordered categorization, unordered categorization or a mixture of both (e.g.,classification by social class, which is ordered, education which is ordered, ethnic group which is unordered, geographical area which is unordered). The procedure described below applies equally to both types.

28. A further distinction between types of categorization is analogous to that encountered in the theory of continuous variables. On the one hand, some variables may be selected for study as dependent on others (analysis of dependence); on the other hand interest may be in the relation of a group of variables among themselves (analysis of interdependence). Examples of the former are regression and analysis of variance; examples of the latter are component analysis, factor analysis and cluster analysis. Fortunately for the reduction of the number of hypotheses to be considered, the former case is encountered more frequently than the latter.

29. There are two fundamental problems involved in the analysis of multiple contingency. One is to set up a measure of relationship between two or more variables. This is usually done by the use of the X^2 statistic or some function of it. The other is to find ones way through a maze of possible hypotheses in a systematic manner.

For example, in a two-way table it is customary to compare the frequency observed in each cell (say F) with the frequency which would have been observed if the variables were independent (say T) - the latter being calculated by regarding the one-way marginal totals as fixed. There are then two measures (which are asymptotically equivalent) in current use to test the hypothesis of independence:

$$X^2 \text{ (Pearson)} = \Sigma \frac{(F - T)^2}{T} \tag{22}$$

$$X^2 \text{ (Likelihood ratio)} = 2\Sigma F \log \frac{F}{T} \tag{23}$$

where summation takes place over the cells of the table.
Before the days of the pocket computer the former was easier to compute, but the latter is preferable.

30. When we come to a three-way table, (say of variables A, B, C) there is no longer any one single hypothesis to test but 17, of which a few are trivial. They may be exhibited as follows:

A	A,B	A,B,C	AB	AB,C	AB,AC	ABC
B	A,C		BC	BC,A	BC,BA	
C	B,C		CA	CA,B	CB,CA	

Here, for example, A,B refers to a hypothesis based on 'fixing' the univariate margins of A and B. AB represents the hypothesis that the entire three-way table is determined by the joint distribution of A and B. AB,AC is a test 'fixing' the two-way margins AB and AC. In point of fact, the test that a single variable A "explains" the entire table is trivial: it simply tests whether the frequencies in the same category of A are all equal within sampling limits. Similarly for B and for C. Likewise model ABC (which is added for completeness) requires no test because it fixes all the cells in the table; it is referred to as the "saturated" model. The other 13, however, may be of interest. A test based on AB,C, for example, is akin to the test of a partial correlation - are A and B dependent when the effect of C is abstracted?

31. The number of possibilities to be examined increases alarmingly with the number of dimensions. For four-way tables there are 167 and for five-way tables there are thousands. Within the space of this note it is impossible to discuss a systematic approach in detail. Sometimes prior specification of the hypotheses under study will cut down the number of possibilities to be examined. Where this is not so, it seems better to proceed from simpler to more complex models, stopping the analysis when a parsimonious model has been reached (that is to say, one with the fewest parameters).

It is hoped to prepare a Technical Bulletin dealing with this subject in greater detail.

VARIATE TRANSFORMATIONS

32. Since, even in the computer era, linear mathematics are relatively easy mathematics, there has been a tendency on the part of statisticians

to devote most of their attention to models expressed in linear form. (The regression equation (13) is a case in point). Such models impose a severe limitation on the data and it is very desirable to consider at the outset whether linearity is realistic and if not what can be done to improve the model.

33. There are two traditional procedures

(i) If the model is thought to be multiplicative rather than additive (as in the Cobb-Douglas type of demand function in economics) linearity can be achieved by working with the logarithms of the data instead of the original data.

(ii) Even if relationships are not linear the range of interest may be narrow enough for a curvilinear relation to be adequately approximated by a straight line.

34. Apart from this there are many circumstances where it is desirable to make transformations of the variables before submitting them to mathematical analysis. For example, in the context of fertility, consider

(i) *THE MARGINAL IMPACT OF BACKGROUND VARIABLES*

In a number of contexts, such as the relation between stimulus and response in psychophysics (the Weber-Fechner Law) or the relation between income and its utility in economics, the effect of a change in the first variable upon the second, is found to depend on the level as well as the amount of change in the first variable. There may be reasons to expect similar relations among WFS variables. For example, an additional year of education or unit of income may well have a reduced impact upon fertility or fertility intentions if the level of education or of income is already high.

(ii) *THE MARGINAL IMPACT OF INTERMEDIATE VARIABLES*

In studying the effect of some variables on others, it is difficult to specify the mathematical relationships among intervening variables (e.g., the practice of prolonged breast-feeding and the length of post-partum amenorrhea) and their relation to fecundity or fertility. Nevertheless, the usual linear model is clearly

inappropriate for most relationships, and its unexamined adoption will yield, at best, a first approximation. It is seemingly plausible, for example, that the probability a woman will conceive in a month is a linear function of the frequency of sexual intercourse, because the probability is bounded by zero and unity.

In some cases the impact of a change in one variable upon the conditional expectation of another variable can be postulated through earlier, related research. In some cases a relationship has been built in to the definition of the variables. For example, a general fertility rate is the product of a marital fertility rate and the proportion married, and a similar relation holds among Coale's indices of marital fertility. Sometimes one employs a systematic stepwise procedure, already alluded to, including progressively higher order interactions and polynomials until the best 'fit' has been obtained. In other cases one may have the data points plotted by computer and, by inspection, identify a pattern in the conditional expectations.

35. There is another type of transformations whose function is to achieve homoscedasticity of the 'error' term. These are distinct from transformations which produce linearity of the conditional expectations. The two types may be used in conjunction.

If the variables can be transformed to achieve linearity and homoscedasticity, then the usual least-squares estimation procedures may be used. However, there is no longer a limitation to this kind of final form. Nelder and Wedderburn have described a method for estimating parameters when the observations are distributed according to an exponential family. There are also powerful iterative computer programs which permit estimation of the parameters in virtually any form of equation.

REFERENCES

E.M.L. Beale and R.J.A. Little, *Missing Values in Multivariate Analysis.* (J. Roy. Statist. Soc. B, 37, 129, 1975)

Y.M.M. Bishop, S.E. Fienberg and P.W. Holland, *Discrete Multivariate Analysis.* (M.I.T. Press, 1975)

Y. Haitovsky, *Missing Data in Regression Analysis.* (J. Roy. Statist. Soc. B, 30, 67, 1968)

Haitovsky, Y, *Regression Estimation from Grouped Observations.* (Charles Griffin & Co.,1973)

M.G. Kendall, *Multivariate Analysis.* (London: Charles Griffin & Co., 1975)

R.L. Plackett, *The Analysis of Categorical Data.* (London: Charles Griffin & Co., 1974)

J.A. Nelder and H.W.M. Wedderburn, *Generalised Linear Models.* (J. Roy. Statist. Soc. A, 135, 370, 1971)