DOCUMENT RESUME

ED 100 958 TM 004 042

AUTHOR	Burstein, Leigh
TITLE	Issues Concerning Inferences from Grouped Observations.
PUB DATE	[Apr 74]
NOTE	73p.; Paper presented at the Annual Meeting of the American Educational Research Association (Chicago, Illinois, April 1974). Not available in hard copy due to marginal legibility of original document
EDRS PRICE	MF-$0.75 HC Not Available from EDRS. PLUS POSTAGE
DESCRIPTORS	College Freshmen; *Data Analysis; *Grouping Procedures; *Inductive Methods; *Prediction; *Statistical Bias
IDENTIFIERS	*Inference (Statistical)

ABSTRACT

The investigation focused on the effects of using grouped data to estimate the relations that exist in data on individuals. Different research contexts were identified in which researchers group observations though interested in relations among measurements on individuals. The consequences of estimating regression coefficients from grouped data were examined from a "structural equations" perspective. A simple linear regression model was hypothesized and then modified by the incorporation of a "grouping variable." A taxonomy was generated from the modified model so that every possible grouping variable fit into one of four categories defined by the relations of the grouping criterion to other variables in the system. Each category was then examined for bias and efficiency of estimation. General principles were determined for choosing a grouping method which minimizes loss of information. The complications that arose in the multiple regression case were also delineated. (Author/SM)

ED 100958

BEST COPY AVAILABLE

ISSUES CONCERNING INFERENCES FROM GROUPED OBSERVATIONS+

Leigh Burstein [1]
Department of Educational Psychology
University of Wisconsin-Milwaukee

1. INTRODUCTION

This presentation focuses on the effects of using data from groups of individuals to estimate relations that exist in data on individuals. Such discussions occur in the research literature under the names "data aggregation", the "grouping of observation", or simply "grouping", which all refer to the replacement of numbers representing observations on individuals with a smaller set of numbers representing obervations aggregated (in the present context, averaged) over groups on individuals. For example an investigator may group observations by classroom and analyze between-class relations.

The study of grouped data introduces no special obstacles when inferences are restricted to the level at which the data are collected and analyzed. Complications can arise, however, when investigators turn to data on groups of individuals to estimate regression and correlation coefficients at the individual level. An attempt to estimate the relation between student achievement and student aspiration level from class means for achievement and aspiration can result in seriously misleading estimates and faulty inference. The types of problems considered in this paper are called "change in the units of analysis" problems

+ Paper presented at the Annual Meeting of the American Educational Research Association, April 19, 1974, Chicago, Illinois.

TM 004 042

BEST COPY AVAILABLE

(Blalock, 1974) where, like the example above, inferences about the relations among individuals are desired but the data are analyzable at the group level only.

The objectives of this discussion are:

(1) to identify research contexts in which investigators estimate regression and correlation parameters from grouped observations though interested in relations among measurements on individuals;

and

(2) to clarify the conditions under which estimates of regression and correlation coefficients obtained from grouped data are consistent with estimates that would be obtained from ungrouped data.

First, the different research contexts in which grouping can arise are discussed and earlier investigations of each context which considered data aggregation methods are cited. A framework is offered to clarify certain similarities among the different research contexts and thus simplify the process of identifying whether a particular grouping strategy is applicable for a given context.

Next, the existing results from three different approaches ("clustering", "optimal grouping", and "structural equations") for examining the effects of grouping observations are summarized. This discussion focuses on the parallels and contradictions among these different lines of inquiry. Of the approaches contrasted, the "structural equations" appears to be most promising and will receive the most attention throughout the paper.

In Section 4 a more general approach which subsumes all others is presented. This approach is an extension of the "structural equations"

BEST COPY AVAILABLE

approach originally articulated by Blalock (1964) and Hannan (1970, 1971, 1972). We concentrate on the simple linear regression model and describe systematic procedures for examining the consequences of different methods of grouping in this two-variable case.

The general strategy described in Section 4 is to modify the structural regression model at the individual level by incorporating the grouping characteristic directly into the model. This modified causal structure leads logically to a taxonomy whereby every possible grouping characteristic fits into one of several mutually exclusive categories defined by the relations of the characteristics to the variables in the original regression model. The different characteristics within a given category then have similar implications for the precision of estimating individual-level relations from grouped observations.

In Section 5 data from a study of incoming freshmen at a large midwestern university are used to illustrate the procedures developed here. The results for both regression and correlations coefficients are found to conform with the predictions from extended "structural equations" approach. Additionally, a compositing procedure is described which generates a more stable estimate of the individual-level regression coefficient from the separate estimates from data grouped by several of the best grouping variables.

In the concluding section, the suggestions for improving inferences from grouped data are summarized. In addition, promising strategies for treating unordered grouping characteristics such as classroom are suggested. Complications that arise in extending the results to the

ERIC

BEST COPY AVAILABLE

multivariate case are also delineated. The prediction of effects under nominal grouping conditions and the further elaboration of the consequences of aggregation within a multivariate causal structure represent the kind of aggregation problems where more research and attention will be needed before investigators can analyze all kinds of change-in-units problems with confidence.

2. Data Aggregation in Different Research Contexts

The analysis of grouped data is becoming increasingly common in educational research as investigators contemplate massive census-like data in addition to school and classroom aggregate measures. Carefully chosen grouping methods can also be applied when the question of confidentiality of data arises, when data is missing from some individuals in a study, and when the variables in the study are fallibly measured.[2]

The degree of investigator control over the aggregation of data is an important consideration for each kind of "change-in-the-units-of-analysis" problem. In certain contexts group membership is determined in some natural way by, e.g., school attended or state of residence, and is thus beyond the investigator's control except for exclusion of certain sampling units and individuals (limited or no investigator control). In other contexts the investigator can manipulate the formation of groups at least in part (complete or partial control). Obviously there are more options in the latter contexts for improving the precision of estimates from grouped observations.

BEST COPY AVAILABLE

Table 1 presents a detailed account of the different research contexts and the investigator's options for controlling the formation of groups. Explanations regarding why data aggregation methods are needed, how such methods are applied, where they are principally applied, and who previously conducted research related to each context are also included.

Table 1

Contexts Allowing Complete Investigator Control Over Grouping Membership. Despite the seemingly extensive lists of references, aggregation procedures have been applied infrequently in Contexts (A) and (B), at least in recent years. Perhaps further simplification and clarification of of the grouping methods may be necessary to extend their use in these contexts. However, a more likely explanation for their limited use is that more powerful statistical methods have been proposed (see Affifi and Elashoff (1966, 1967) on the missing information problem and Madansky (1959), Blalock et al. (1970), Blalock (1971), and Wiley and Wiley (1971) on the measurement error problem.) It is still an open question whether these other methods will warrant more consideration once further refinements in grouping methods have been made.

Econometricians have already developed and demonstrated sound principles for data aggregation where the size and economy of analysis (Context (C)) is the chief concern (Prais and Aitchinson (1954), Theil (1954), Cramer (1964), Green (1964).) The other social and behavioral sciences ble 1. Research contexts for data aggregation procedures as a function of the degree of investigator control of groupir

Research Context	Reasons for Use of Data Aggregation	Description of Application of Data Aggregation	Principal Applications	Important Related Literature-
I.	Complete Investigator Control--Group membership can be defined by any characteristic in the data set which is measured for all individual level units.			
(A) MISSING OBSERVATIONS	Missing observations on primary variables for some individuals inhibit confidence in analytical results.	Each missing observation on a primary variable is replaced by the mean response on that variable for some group to which the individual belongs.	Longitudinal and cross-sectional analysis of survey data.	Affifi and Elashc (1966, 1967) Elashoff and Ela (1970,1971) Kline et al.(197
(B) FALLIBLY MEASURED VARIABLES	Random errors of measurement associated with independent variables attentuate estimates of regression coefficients.	Different approaches have been suggested as part of the general refinement of statistical procedures for handling "errors-in-variables" problems.	Statistical treatment of measurement errors in many studies.	Wald (1940) Bartlett (1942) Tukey (1951) Madansky (1959) Blalock (1964) Chai (1968, 1971) Blalock et al.(19
(C) SIZE AND ECONOMY OF ANALYSIS	Budgetary and equipment constraints make analysis of massive data base at the individual level impractical	Data are collapsed into smaller administrative units by some grouping rule.	Analysis of census data and national, regional, and state school statistics.	Prais and Atichir son (1954) Theil (1954) Cramer (1964) Green (1964)
II.	Partial Investigator Control--Group membership can be defined by any characteristic which has been measured simultaneously with each promary variable.			
(D) ANONYMOUSLY COLLECTED INFORMATION	Data on certain primary variables is collected anonymously making it impossible to match observations on primary variables at the individual level.	Characteristics measured simultaneously with both anonymously collected and identifiable information can be used to aggregate the data	Confidential student records and responses to attitudinal questionnaires.	Feige and Watts (1970, 1972) Boruch (1972)

BEST COPY AVAILABLE

Research contexts for data aggregation procedures as a function of the degree of investigator control (continued)

esearch ontexts	Reasons for Use of Data Aggregation	Description of Application of Data Aggregation	Principal Applications	Important Related Literature
III.	<u>Limited or No Investigator Control</u>--Group membership is determined prior to the collection and analysis of data; group membership is directly pertinent to the study of primary variables.			
E) ECOLOGICAL INFERENCE	The sampling units of the investigation constitute "natural" aggregates of individuals.	Disaggregation efforts are generally a necessary precondition to reasonable inferences at the individual level.	Analysis of school and classroom means where the school and the class are the samp-units; data organized by census tract or demographic region.	Gehkle and Biehl (1934) Thorndike (1939) Yule and Kendall (1950) Robinson (1950) Duncan and Davis (1953) Goodman (1953, 195 Selvin (1958) Blalock (1964) Haitovsky (1966, 1968, 1971) Chai (1968, 1971) Alker (1969) Cartwright (1969) Shively (1969) Iverson (1973)

BEST COPY AVAILABLE

BEST COPY AVAILABLE

are just beginning to tap the tremendous wealth of methodological advances made by the econometricians. The methodology for handling data aggregation problems is no exception to the slow pace at which educational, psychological, and sociological researchers are incorporating the econometricians' "power tools" into their theory-building enterprises.

In the attempt to expand the conceptual theory for handling change-in-units problems, this investigator incorporates the econometric results that simplify the present efforts and builds on their framework where the special problems of dealing with individuals, rather than institutions or commodities, dictate modifications. As will be shown, however, the work of Prais and Aitchinson (1954), and later of Cramer (1964), in Context (A) is an essential part of any adequate conceptualization of the problems of data aggregation discussed in this paper.

Partial Investigator Control over Group Membership--Confidential Data in Social Research. The use of partial aggregation techniques for analyzing confidentially collected information is a relatively new notion. Feige and Watts (1970) apparently were the first to recognize the utility of grouping methods in this context. The procedure itself is quite simple. The investigator collects information on potentially suitable grouping characteristics in addition to those of primary interest. The individual observations can then be collapsed into different groups and the parameters of interest can be estimated from the between-group relations. This procedure is viable as long as the grouping characteristics are measured simultaneously with each primary variable (regard-

ERIC

BEST COPY AVAILABLE

less of whether the information on the primary variable is collected anonymously or with the individual subjects identified), and satisfy certain conditions necessary for precise estimation in change-in-units analyses.

Conducting research on confidential data presents very complicated social and political problems. Boruch (1971(a), 1971(b), 1972(a), 1972 (b)) brings into focus the ethical and legal considerations associated with research under confidentiality constraints besides suggesting alternatives to partial aggregation methods. Although the need for the for the privacy and protection of subjects in social research is recognized, this presentation does not deal directly with such complications. The procedures suggested in this investigation offer individuals assurances of their anonymity while maintaining the possibility of carrying out research on topics that can further understanding of the complex interactions among individuals and institutions within our society. The premise is that a person can maintain his or her individual identity and still cooperate with efforts to clarify the cornerstones of social processes through analysis methods designed to allow examination of relations among human characteristics without directly identifying the participating individuals.

Limited or No Investigator Control -- "Ecological Inference". The topic of ecological inference has received a lot of attention in the sociological literature. There was an extended exchange of ideas on the subject among

BEST COPY AVAILABLE

sociologists and social statisticians (Robinson, 1950; Duncan and Davis, 1953; Goodman, 1953; Goodman, 1959) in the 1950's. Though the debate centered around methods for overcoming the "ecological fallacy", there were many who just wondered what all the fuss was about. After all, the group and inter-group relations occupy a prominent position in sociology, and thus group-level analyses should be acceptable.[3]

The educational and psychological literature hardly reflects an awareness of the "ecological fallacy" of inferring correlations between properties of individuals from the ecological correlations derived from group data.[4] Important reports (Coleman _et al_., 1966) and papers in respected journals of educational and psychological research (Goldberg, 1969; Rock _et al_., 1970; Baird and Feister, 1972) perform between-group analyses without directly considering whether the relations estimated at the group level are applicable at the individual level.

The correct response to the query regarding appropriate level of inference is not obvious. However, the dramatic change cited by Robinson (1950) in size of the correlation between illiteracy and race as a function of the coarseness of the units of analysis (.95 at the regional level, .77 at the state level, .20 at the individual level), should warn researchers not to take the query lightly. The investigator definitely needs to understand the rules governing his grouping process in any empirical study in the social or behavioral sciences (Levy, 1972).

In any case, aggregate sampling units present a particularly complex type of aggregation problem since questions regarding sampling bias

BEST COPY AVAILABLE

arise in addition to concerns about level of inference. One question may be whether the sampled schools are representive of the schools in the universe to which one wants to generalize. The investigator must clearly understand the basis for his inferences to the individual level in order to be at all confident about his estimates. Otherwise, it may be best to make inferences at the group level or to examine the individuals within groups (Wiley, 1970).

Applicability of the Taxonomic Approach to the Different Research Contexts. The strategies developed here are suitable mainly for Contexts (A) through (D). However, they do have implications as to how one can proceed when the grouping characteristic has a nominal scale which most often happens in Context (E). To take full advantage in Context (E) of the procedures described below, the investigator must first determine how to express the relations of the grouping characteristic to primary variables in an ordinal fashion. How this can be done is a topic requiring more research, but once it is done, the strategies are applicable.

BEST COPY AVAILABLE

3. The Basic Model and Accompanying Review of the Literature

A basic summary of the different approaches to change-in-units analysis is provided in Hannan and Burstein (1974). A more complete accounting of the problems and strategies of grouping for individual-level inferences can be found in Burstein (1974). In the present paper, the effects of grouping on the estimation of the relation between two variables within a simple linear regression framework is examined. The reason for the restriction to the two-variable case is that the different approaches achieve their purest and simplest forms when only two variables are considered. Forecasting results for higher-order relations is possible, but the strategies can not be as clearly delineated (See Section 6).

The regression model is examined because it results in formulation suitable for estimating both regression and correlation coefficients. If the preferred "structural equations" approach is followed, the investigator need only conduct his analysis on the individual observations in standardized form if he wishes to estimate correlations. Once the individual observations are standardized, the comparison of the regression coefficients before and after data aggregation becomes essentially a comparison of zero-order correlation coefficients at the individual and group levels. This slight twist of procedure enables the investigator to apply the same basic model to both regression and correlation coefficients.

The analysis that follows might best be viewed in the context of a substantive problem. Assume that an investigator wished to study the

BEST COPY AVAILABLE

relation of achievement (X) to students ratings of their intellectual abilities (Y). To be specific, he wants to estimate the linear regression coefficient β_{YX} from the simple linear model

$$(1) \qquad Y = \alpha + \beta_{YX}X + u \ ,$$

where α is the intercept, β_{YX} is the standardized regression coefficient from the regression of Y on X, and u represents the lack of fit of a linear model and the effects of other variables on Y, independent of X.

To estimate β_{YX} the investigator normally collects paired measurements (X_p, Y_p) from a sample of N students (p=1,...,N) and then uses ordinary least-squares (OLS) procedures to estimate β_{YX} from the equation

$$(2) \qquad Y_p = \alpha + \beta_{YX}X_p + u_p$$

under the assumptions that

1) $E(u_p) = 0$, for all p.

2) $E(u_p u_{p'}) = \begin{cases} \sigma_u^2, & \text{if } p=p', \text{ constant for all p.} \\ 0, & \text{otherwise.} \end{cases}$

3) $E(X_p u_p) = 0$, for all p.

BEST COPY AVAILABLE

The OLS estimator b_{YX} of β_{YX} for this model is known to be

$$(3) \qquad b_{YX} = \frac{C(X_p, Y_p)}{V(X_p)} = \frac{C(X,Y)}{V(X)} \quad ,$$

where C(X,Y) and V(X) are the sample covariance of X and Y and the sample variance of X, respectively. (From here on, subscripts are dropped where the interpretation of the values will remain unambiguous.)

Under the assumptions for (2),

$$E(b_{YX}) = \beta_{YX}$$

with estimated mean squared error (MSE) equal to the variance of b_{YX}; i.e.,

$$(4) \qquad MSE(b_{YX}) = V(b_{YX}) = E(b_{YX} - \beta_{YX})^2 = \sigma_u^2 E(\frac{1}{SS(X)})$$

where the SS(X) is the sum of squares for X.

The next step is to estimate β_{YX} from observations grouped on some characteristic. Earlier researchers have approached the problem of grouped estimation in several ways. Three relatively distinct perspectives are discussed below.

BEST COPY AVAILABLE

The Clustering Perspective. The earliest treatment of aggregation problems in the social sciences developed from concerns over the inflation of correlation coefficients as individual observations are grouped together. This effect was noticed in a wide variety of investigations. Robinson's (1950) data on the ecological correlation between race and illiteracy were cited above. Prior to him, Gehkle and Biehel (1934) showed the inflationary effects of grouping for data on rental values and delinquency rates, Thorndike (1939) cited the same fallacy in the psychological literature over the correlation between family size and delinquency, and Yule and Kendall (1950) examined correlations among consolidated regional crop yields.

Each investigator tried to uncover the mechanism responsible for what he perceived to be the grouping artifact. Most arrived at essentially the same conclusions from their different algebraic formulations. Since Robinson's work on ecological correlation has reveived much attention (See Alker (1969); Hannan (1970,1971)), it will be summarized here, with some modification in notation, as an example of the clustering perspective.

Robinson employs "covariance theorems" to decompose the sum of squares and sums of cross-products into their within-group and between-group (ecological) components. Given a sample of size N comprised of groups equal size n,

$$SS(XY) = WSS(XY) + SS(\bar{X}\bar{Y}),$$

where WSS(XY) and $SS(\bar{X}\bar{Y})$ are the within-groups and between-groups sums of cross-products, respectively. Similarly,

BEST COPY AVAILABLE

$$SS(X) = WSS(X) + SS(\bar{X}) \quad ,$$

where WSS(X) and SS($\bar{X}$) are the within-group and between-groups sums of squares for X.

The above equations are substituted into the formula for the correlation coefficient r_{XY} from ungrouped observations, and the terms are rearranged to yield

$$(5) \qquad r_{XY} = Wr_{XY}\sqrt{1-\eta^2_X}\sqrt{1-\eta^2_Y} + r_{\bar{X}\bar{Y}}\sqrt{\eta^2_X\eta^2_Y}$$

In Equation (5), Wr_{XY} and $r_{\bar{X}\bar{Y}}$ are the within-group and ecological correlations, respectively, and $\eta^2_X \left(= \frac{SS(\bar{X})}{SS(X)}\right)$ and $\eta^2_Y \left(= \frac{SS(\bar{Y})}{SS(Y)}\right)$ are the familiar correlation ratios for X and Y.

The relationship between $r_{\bar{X}\bar{Y}}$ and r_{XY} is complex but describable. In his interpretation of Equation (5), Robinson identified several typical effects of consolidating units:

(i) The ecological correlation decreases as groups become more heterogeneous since the within-group correlation increases directly with increasing coarseness and the between-group proportion of the variation equals 1 - Wr^2_{XY}.

(ii) The correlation ratios η^2_X and η^2_Y decrease as the between-groups variation becomes smaller.

(iii) Of the first two effects, the changes in the correlation ratios are considerably more important than the changes in Wr_{XY} so that the numerical value of the ecological correlation increases with increasing consolidation of units.

Thus, according to the clustering approach, grouping always inflates coefficients. But, as Hannan and Burstein (1974) have pointed out, the clustering approach fails to explicate the nature of the grouping process,

BEST COPY AVAILABLE

and thereby misses certain key distinctions among ways of consolidating units. This is especially unfortunate since the clustering approach is mainly applicable to Context (E) where group membership is "naturally determined", and the discovery of the grouping mechanisms is a complicated but necessary endeavor.

The "Optimal Grouping" Approach. Optimal grouping proponents were motivated by the need to reduce their over-abundant data (Content (A)) by a grouping strategy which optimized the retention of the ungrouped information. Prais and Aitchinson (1954) and Cramer (1964) are largely responsible for the basic work in this area. Cramer's formulation of the single regressor case is summarized below.

Cramer started with Equation (2) above for his model at the individual level with one important change. He relabeled the individual observations by letting X_{ij} (replacing X_p in (2)) represent the achievement score of the jth student in the ith group and Y_{ij} (replacing Y_p represent the student's corresponding academic self-rating.[5] This was done to reflect the underlying, as yet unspecified, group membership. Cramer subsequently arrived at Equations (3) and (4) above for his individual-level b_{YX} and $V(b_{YX})$.

Next, the group means ($\overline{X}_i$, $\overline{Y}_i$.) are found by averaging over X_{ij} and Y_{ij} within groups, and these "grouped observations" become the units of analysis. For the substantive example, this is equivalent to grouping by, say, classroom and using class mean achievement and class mean self-rating in the analysis.

BEST COPY AVAILABLE

The equation at the group level is then

$$(6) \qquad Y_{i.} = \alpha + \beta_{YX}X_{i.} + u_{i.} \ .$$

Under the assumptions for (2), the regression coefficient in the population for (6) has the same value as in (2).

The weighted least-squares estimator $B_{\bar{Y}\bar{X}}$ of β_{YX} from (6) is

$$(7) \qquad B_{\bar{Y}\bar{X}} = \frac{C(\bar{X}_{i.},\bar{Y}_{i.})}{V(\bar{X}_{i.})} = \frac{C(\bar{X},\bar{Y})}{V(\bar{X})} \ .$$

Under the assumptions above,

$$E(B_{\bar{Y}\bar{X}}) = \beta_{YX}$$

and

$$(8) \qquad V(B_{\bar{Y}\bar{X}}) = \sigma_u^2 E\left(\frac{1}{SS(\bar{X})}\right) \ .$$

Thus, according to Cramer, to Prais and Aitchinson, the grouped estimator $B_{\bar{Y}\bar{X}}$ is an unbiased estimate of β_{YX} with relative efficiency

$$(9) \qquad Eff(b/B) = \frac{V(b)}{V(B)} = E\left(\frac{SS(\bar{X})}{SS(X)}\right) = X \ ,$$

the familiar correlation ratio which has a value less that 1.

However, Cramer indicates that the estimate of the correlation coefficient ρ_{XY} between X and Y is inflated if the groups are not formed randomly.

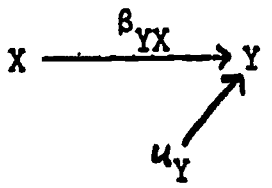
<u>The "Structural Equations" Approach.</u> Blalock (1964) considered the problems of the grouping of observations from a causal perspective. He started with the hypothesis that systematic grouping can lead to differential effects on the regression estimates of causal relations. Blalock argues

BEST COPY AVAILABLE

convincingly (later amplified by Hannan (1972)) that if the investigator, for reasons beyond his control, groups on the dependent variable Y, or by a variable highly related to Y, other than X, $B_{\bar{Y}\bar{X}}$ will be a biased estimate of β_{YX}. He cites the facts that (1) the value $r^2_{XY} = b_{YX}b_{XY}$ is inflated by any systematic grouping procedure (that is, $r^2_{\bar{X}\bar{Y}} > r^2_{XY}$), and (2) grouping on the independent variable X (or by a variable highly related to X) does not produce bias (that is, $B_{\bar{Y}\bar{X}} = b_{YX}$ for grouping on X). Taking (1) and (2) together implies that $B_{\bar{Y}\bar{X}} > b_{XY}$. Thus grouping on X inflates the regression coefficient when X is the dependent variable. Similarly, grouping on Y inflates the estimate of β_{YX}.

Blalock also describes the phenomenon in another way. Grouping on Y results in a proportional reduction in the variation of Y and the covariation of X and Y. But the variation in X exhibits a greater proportional reduction unless X and Y are extremely highly related. Since $V(\bar{X})$, and not $V(\bar{Y})$, is the denominator of the sample estimator $B_{\bar{Y}\bar{X}}$, bias will result from this type of grouping.

Hannan (1972) uses a different argument to arrive at the same conclusion. He starts with the causal model



where $u_{\bar{Y}}$ represents the influence of other causes of Y. He then states that when variation in Y is maximized by ranking observations by their Y values and then grouping "adjacent" observations, observations that have high X and high u_Y values will be placed in the highest Y groups.

BEST COPY AVAILABLE

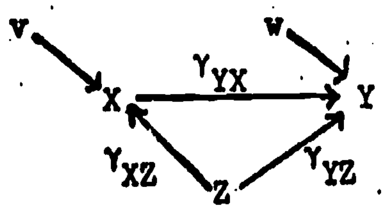
Similarly, observations with both low X and low u_Y values are placed in the groups lowest on Y. (assuming B_{YX} is positive) Thus, the other causes of Y are confounded with X so that σ^2_{Xu} is no longer zero in the probability limit. Hannan calls this a specification error introduced by grouping and calls the bias in the OLS estimates at the group level aggregation bias.

4. Reconsideration of the Basic Model -- Introduction of a "Grouping Variable"

Neither Blalock nor Hannan offer formal mathematical arguments for their findings. However, their causal thinking suggests that 'the role of the grouping rule (see Theil, 1954) might best be explicitly identified in the model even though its presence is strictly dictated by its use for group formation.by introducing a grouping variable into the causal structure.[6] In other words, the criterion by which the individual observations are to be grouped is treated as a random variable which may be related to other variables in the system. Furthermore, if the grouping variable Z is related to another variable, the causal structure specifies that Z is causally prior to that variable. It does not matter that Z may appear to be caused by,say X in the sense that X would be logically or temporally prior to Z if the three-variable model Y=f(X,Z) were under study. We visualize the grouping process as one in which Z can "select" or "force" the observations from the bivariate distribution of X and Y into groups. It is in this sense that Z is always causally prior to X and Y.

BEST COPY AVAILABLE

Structural Equations Incorporating the Grouping Variable. Once this direction of causation has been specified, the structural model for the relations among X, Y, and Z is easily defined. The path diagram for the causal structure when Z is prior to X and Y is



In this diagram, v is the disturbance term representing all causes of X that are not linearly related to Z, and w is the disturbance term representing all causes of Y that are not related to X and Z. γ_{YX}, γ_{XZ}, and γ_{YZ} are the path regression coeffcients.

The equations corresponding to the causal structure with Z incorporated can be written

$$(10a) \qquad Y = \alpha + \gamma_{YX}X + \gamma_{YZ}Z + w \quad ,$$

$$(\ b) \qquad X = \lambda + \gamma_{XZ}Z + v \quad .$$

Once again, γ_{YX}, γ_{XZ} and γ_{YZ} are regression parameters, and w and v are disturbance terms. w is assumed to be independent of X, Z and v, and v is also assumed to be independent of Z. Both disturbance terms are homoscedastic and independent. ($\sigma^2_{w_1} = \sigma^2_{w_2} = \sigma^2_{w}$ and $\sigma^2_{v_1} = \sigma^2_{v_2} = \sigma^2_{v}$; $\sigma_{w_1w_2}=0$ and $\sigma_{v_1v_2}=0$ for any two persons.).

The notation α is retained for the intercept term in (10a) though its value may differ from that in earlier equations. The notation λ will represent the intercept term in the second equation of the structural system; its value may also vary according to the specified model.

BEST COPY AVAILABLE

Equation (10b) can be substituted into (10a) to obtain a single equation for the regression of Y on Z and v:

$$(11) \qquad Y = (\alpha + \gamma_{YX}\lambda) + (\gamma_{YX}\gamma_{XZ} + \gamma_{YZ})Z + \gamma_{YX}v + w \ .$$

Equation (11) is actually a reparameterization of (1) where X has been divided into two parts -- the part predictable from the grouping variable Z and a residual part v. Equations like (11) are generally called the "reduced forms" for the causal structures. Equation (11) is in a form that cannot be reduced further by substitution of other equations.

The regression of Y upon X still has the regression coefficient β_{YX}. This coefficient can be expressed in terms of the fixed parameters of the modified causal structure. Besides the intercepts, the fixed parameters are the three regression coefficients (γ_{YX}, γ_{YZ}, γ_{XZ}). and the variances of X, v and w (σ_Z^2, σ_v^2, σ_w^2). Substitution of the reduced form expressions into the formula for β_{YX} yields

$$(12) \qquad \beta_{YX} = \frac{\sigma_{YX}}{\sigma_X^2}$$

$$= \frac{\gamma_{XZ}(\gamma_{YZ} + \gamma_{YX}\gamma_{XZ})\sigma_Z^2 + \gamma_{YX}\sigma_v^2}{\gamma_{XZ}^2\sigma_Z^2 + \sigma_v^2}$$

$$= \gamma_{YX} + \gamma_{YZ}\gamma_{XZ}\left(\frac{\sigma_Z^2}{\sigma_X^2}\right)$$

Estimator of β_{YX} from Individual Data. Under the modified causal structure, a simple random sample of $N(=\sum_{i=1}^{g} n_i)$ observations is drawn from the trivariate distribution $f(X_{ij}, Y_{ij}, Z_{ij})$. The sample regression estimator of β_{YX} is given by

BEST COPY AVAILABLE

$$(13) \qquad b_{YX} = \frac{\sum_{i=1}^{g} \sum_{j=1}^{n_i} (X_{ij} - X_{..})(Y_{ij} - Y_{..})}{\sum_{i=1}^{g} \sum_{j=1}^{n_i} (X_{ij} - X_{..})^2}$$

Under OLS assumptions,

$$(14) \qquad E(b_{YX}) = \beta_{YX}$$

$$= \gamma_{YX} + \gamma_{YZ}\gamma_{XZ}\left(\frac{\sigma_Z^2}{\sigma_X^2}\right) \qquad \text{(from (12))} .$$

Note that when X and Z have unit variance, (14) becomes

$$(15) \qquad E(b_{YX}) = \gamma_{YX} + \gamma_{YZ}\gamma_{XZ} \quad .$$

This equation is simply the estimate of the net effect of X on Y along all paths connecting the two variables. When all variables are standardized b_{YX} from (15) is an unbiased estimate of the standardized regression coefficient.

Tha sample variance of b_{YX} can be derived from a theorem in Hansen-Hurwitz-Madow (1953, Vol. II, P.65):

$$V(b) = E_2V_1\ (b) + V_2(E_1(b)) \quad ,$$

where V_1 (b) is the variance of b conditional on X and Z and V_2 is the variance conditional on Z. The resulting variance formula is

$$(16) \qquad V(b_{YX}) = (\ \sigma_w^2 + \gamma_{YZ}^2\sigma_v^2)E\left(\frac{1}{SS(X)}\right)$$

The expression for the sample variance of b_{YX} for the modified causal structure is more complicated that for the simple model. Equations

BEST COPY AVAILABLE

(4) and (16) are similar in form, and when group does not result in bias, they are equal.

Revised Structure for the Weighted Group Means. The structural equations for the group means based on Z can be written as

$$(17a) \qquad \bar{Y} = \alpha + \gamma_{YX}\bar{X} + \gamma_{YZ}\bar{Z} + \bar{w} \; ,$$

$$(\;\; b) \qquad \bar{X} = \lambda + \gamma_{XZ}\bar{Z} + \bar{v} \; .$$

These equations are the same as (10a) and (10b) except that grouped quantities have been substituted for their ungrouped counterparts. There are still six fixed parameters in addition to the intercepts: γ_{YX}, γ_{YZ}, γ_{XZ}, $\sigma^2_{\bar{Z}}$, $\sigma^2_{\bar{v}}$, and $\sigma^2_{\bar{w}}$.

The regression coefficient for the regression of $\bar{Y}$ upon $\bar{X}$ is

$$(18) \qquad \beta_{\bar{Y}\bar{X}} = \frac{\sigma_{\bar{Y}\bar{X}}}{\sigma^2_{\bar{X}}}$$

$$= \gamma_{YX} + \gamma_{YZ}\gamma_{XZ}\left(\frac{\sigma^2_{\bar{Z}}}{\sigma^2_{\bar{X}}}\right) .$$

$$B_{\bar{Y}\bar{X}} = \frac{\sum_{i=1}^{g} n_i \bar{x}_{i.} \bar{y}_{i.}}{\sum_{i=1}^{g} n_i \bar{x}^2_{i.}} .$$

The grouped regression coefficient $\beta_{\bar{Y}\bar{X}}$ can no longer be tacitly assumed equal to the ungrouped coefficient β_{YX}. $\beta_{\bar{Y}\bar{X}}$ and β_{YX} differ in that between-group variances replace the total variance.

Regression Estimator from Grouped Data. A simple random sample of N observations is drawn from the trivariate distribution $f(X_{ij}, Y_{ij}, Z_{ij})$. The X_{ij} and Y_{ij} are then grouped on the basis of the values of Z_{ij}, and each observation replaced by the group mean corresponding to its Z_{ij} value; that is, $\bar{X}_{i.}$ replaces X_{ij} and $\bar{Y}_{i.}$ replaces Y_{ij}.

BEST COPY AVAILABLE

Let $B_{\bar{Y}\bar{X}}$ denote the estimator for the regression coefficient from grouped data. Then

$$B_{\bar{Y}\bar{X}} = \frac{\sum_i n_i \bar{x}_i \bar{y}_{i.}}{\sum_i n_i \bar{x}^2_{i.}}$$

where the lower case letters denote the deviations of the group means from the grand means of the sample.

Under OLS assumptions,

$$(19) \qquad E(B_{\bar{Y}\bar{X}}) = \beta_{YX}$$

$$= \gamma_{YX} + \gamma_{YZ}\gamma_{XZ}(\frac{\sigma^2_{\bar{Z}}}{\sigma^2_{\bar{X}}}) \qquad \text{(from (18))} .$$

Ohe only differences between equations (12) and (18), and consequently between (14) and (19), are that the variances of the group means of Z and X replace the variances of the ungrouped observations. The unbiased estimators of β_{YX} and $\beta_{\bar{Y}\bar{X}}$ are b_{YX} and $B_{\bar{Y}\bar{X}}$ respectively, but the investigator wants to estimate β_{YX} from $B_{\bar{Y}\bar{X}}$. Under certain conditions $\beta_{\bar{Y}\bar{X}} = \beta_{YX}$ and thus $B_{\bar{Y}\bar{X}}$ is an unbiased estimator of β_{YX}.

Bias and Efficiency Formulas. Equations (14) and (19) express the expected values of the regression slope from ungrouped and grouped data in terms of the parameters of the modified causal structure. The expectation of the difference between $B_{\bar{Y}\bar{X}}$ and b_{YX} is the bias that results from grouping, and it will be denoted by θ.

Then

$$(20)\qquad \theta = E(B_{\bar{Y}\bar{X}} - b_{YX})$$

$$= E(B_{\bar{Y}\bar{X}}) - E(b_{YX})$$

$$= [\gamma_{YX} - \gamma_{XZ}\gamma_{YZ}(\frac{\sigma^2_{\bar{Z}}}{\sigma^2_{\bar{X}}})] - [\gamma_{YX} + \gamma_{XZ}\gamma_{YZ}(\frac{\sigma^2_{\bar{Z}}}{\sigma^2_{X}})]$$

$$= \gamma_{XZ}\gamma_{YZ}(\frac{\sigma^2_{\bar{Z}}}{\sigma^2_{\bar{X}}} - \frac{\sigma^2_{Z}}{\sigma^2_{X}}) \quad .$$

The bias term θ has a straightforward interpretation. It implies that the grouping of observations leads to biased estimation if all three of the following conditions hold:

(a) The grouping variable Z has a direct relation to X ($\gamma_{XZ} \neq 0$).

(b) The grouping variable Z has a direct relation to Y ($\gamma_{YZ} \neq 0$).

(c) The ratio of the between-groups variances of Z and X does not equal the ratio of the total variances of Z and X.

Since Z has been defined in such a way that $Z_{ij} = \bar{Z}_{i\cdot}$, $\sigma^2_{\bar{Z}} = \sigma^2_{Z}$.

Whence,

$$(\frac{\sigma^2_{\bar{Z}}}{\sigma^2_{\bar{X}}} - \frac{\sigma^2_{Z}}{\sigma^2_{X}}) = \sigma^2_{Z} E[1/SS(\bar{X}) - 1/SS(X)].$$

Hence, condition (c) can be restated as

(c') The between-groups sum of squares of X does not equal the total sum of squares of X.

The magnitude of the bias increases directly with the increasing relation of Z to both X and Y•X and with the reduction in the variation of X from grouping. These three conditions are not independent, and some interesting ramifications of their interrelations are explored elsewhere.

A formula for the variance of the grouped estimator must be presented before efficiency can be considered. By reasoning like that used to find $V(b_{YX})$, the sample variance of $B_{\bar{Y}\bar{X}}$ can be shown to be

$$V(B_{\bar{Y}\bar{X}}) = (\sigma_w^2 + \gamma_{YZ}\sigma_v^2)E(\frac{1}{SS(\bar{X})}) \tag{21}$$

The efficiency of the grouped estimator is given by

$$E(B/b) = \frac{MSE(b_{YX})}{MSE(B_{\bar{Y}\bar{X}})} \tag{22}$$

where MSE(T) denotes the mean square error for estimator T. Note that

$$MSE(T) = V(T) + (bias(T))^2.$$

So when both estimators are unbiased, the efficiency of the grouped estimator is the ratio of (16) to (21) which reduces to

$$\frac{E[SS(\bar{X})]}{E[SS(X)]}$$

A Taxonomy for Grouping Variables. Figure 1 presents the path diagrams which can result from setting various combinations of γ_{YZ} and γ_{XZ} in (10a-b) equal to zero. Each diagram represents a given set of constraints on the relations of Z to Y and X.

BEST COPY AVAILABLE

Figure 1

A taxonomy for comparing grouping variables can also be derived which will be parallel to the diagrams in Figure 1. Four categories of grouping variables can be distinguished:

I. Z is directly related to both X and Y. ($\gamma_{XZ} \neq 0$, $\gamma_{YZ} \neq 0$)

II. Z is directly related to Y but not to X. ($\gamma_{XZ} = 0$, $\gamma_{YZ} \neq 0$)

III. Z is directly related to X but not to Y. ($\gamma_{XZ} \neq 0$, $\gamma_{YZ} = 0$)

IV. Z is not related to either X or Y. ($\gamma_{XZ} = 0$, $\gamma_{YZ} = 0$)

These categories include all possible relations linking causally prior grouping variables to the regression of Y on X. Certain of these categories represent broader classes of variables. For instance, any random grouping method will satisfy the conditions for Category IV. Most systematic grouping variables belong to Category I. Any grouping variable can be uniquely categorized if the variances and covariances of X,Y, and Z are known.

Examination of Bias and Efficiency for Each Category. Equations (20) and (22) can now be used to examine each category of grouping variables for bias and efficiency. The taxonomic categories are considered in order.

1. Category I -- $\gamma_{YZ} \neq 0$, $\gamma_{XZ} \neq 0$

Category I includes all grouping variables which have direct relations to both X and Y. An obvious example is that scholastic aptitude (Z) may be directly related to both achievement (Y) and student academic interests (X).

BEST COPY AVAILABLE

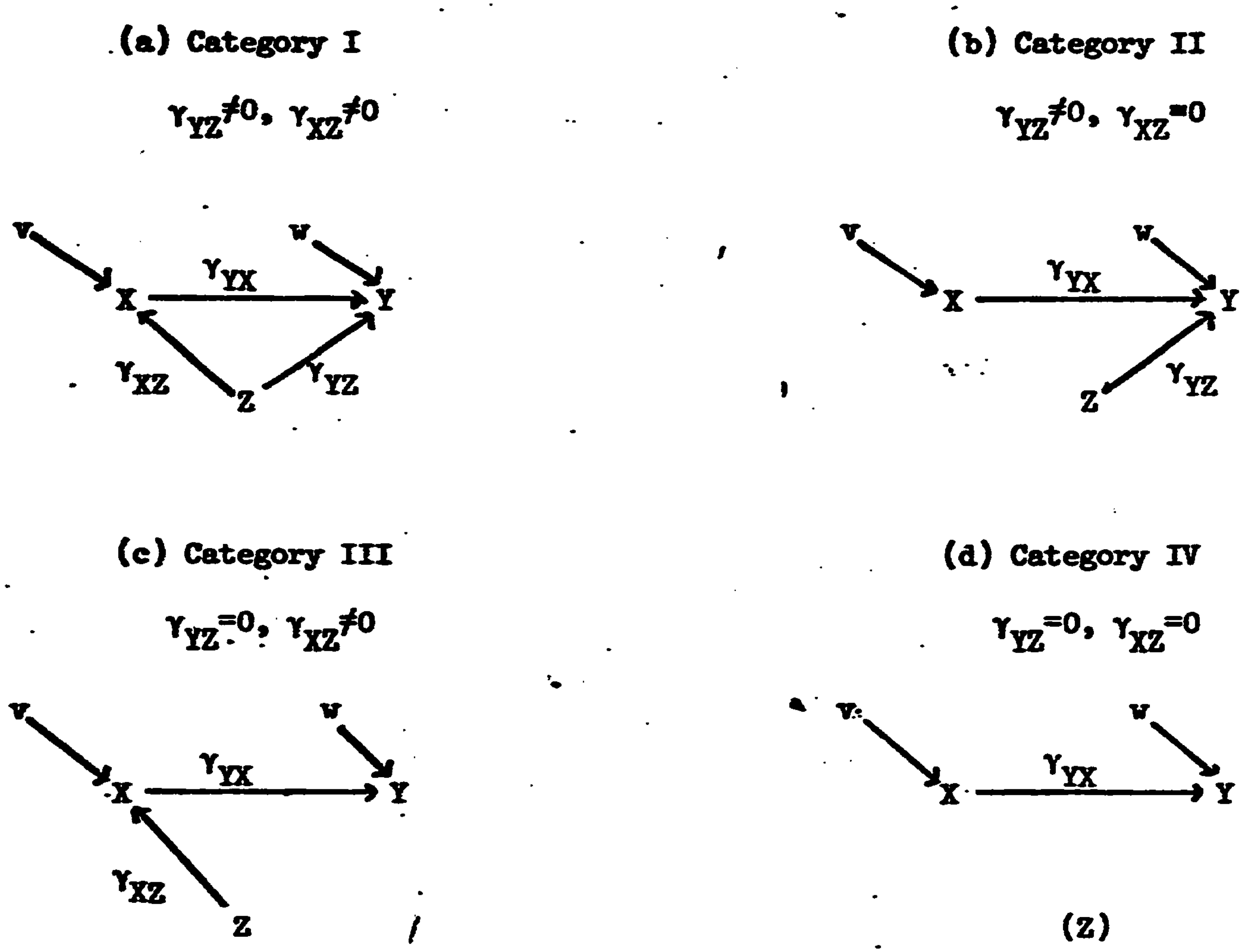


Figure 1. Path Diagrams of the Structural Equations for the Categories of the Taxonomy

BEST COPY AVAILABLE

In general, the slope estimated from data grouped on a Category I variable is a biased estimator of β_{YX}. The magnitude of this bias is given exactly by (20) for known values of γ_{YZ}, γ_{XZ}, σ^2_Z, σ^2_X, and $\sigma^2_{\bar{X}}$. Thus,

$$\theta(Z_I) = \gamma_{YZ}\gamma_{XZ}\sigma^2_Z\left[\frac{\sigma^2_X - \sigma^2_{\bar{X}}}{\sigma^2_{\bar{X}}\sigma^2_X}\right]$$

Some idea about the bias for Category I variables becomes evident from an examination of the bias in estimating standardized regression coefficients. Assume that the original observations are standardized and also that g groups of equal size are formed on discrete values of Z so that $\sigma^2_Z = \sigma^2_{\bar{Z}}$. Under these conditions,

$$(1)\quad \sigma^2_Z = \sigma^2_{\bar{Z}} = 1,$$

$$(2)\quad \sigma^2_v = \sigma^2_X - \gamma^2_{XZ}\sigma^2_Z = 1 - \gamma^2_{XZ}\ ,$$

$$(3)\quad \sigma^2_{\bar{v}} = \sigma^2_v/n\ ,$$

and

$$(4)\quad \sigma^2_{\bar{X}} = \gamma^2_{XZ}\sigma^2_{\bar{Z}} + \sigma^2_{\bar{v}} = \gamma^2_{XZ} + (1 - \gamma^2_{XZ})/n$$

$$= ((n-1)\gamma^2_{XZ} + 1)/n\ .$$

After substitution and simplification, (20) can be written

$$(20')\qquad \theta^*(Z_I) = \gamma_{YZ}\gamma_{XZ}\left[\frac{(n-1)(1 - \gamma^2_{XZ})}{(n-1)\gamma^2_{XZ} + 1}\right]\ ,$$

BEST COPY AVAILABLE

where θ^* denotes the bias of the grouped estimator of the standardized regression coefficient.

The asymmetrical properties of θ^* over the range of γ_{XZ} are illustrated by Figure 2. Predicted bias θ^* is plotted versus γ_{XZ} for fixed γ_{YZ}(=0.1) and selected values of n. A comparable family of curves can be generated for any value of γ_{YZ}. The curves become highly skewed (right) for large n and are roughly symmetrical for small n. This occurs because the groupings become coarser and less representative of the ungrouped observations as n gets larger, no matter what relations exist between Z and X and Y.

Figure 2

Table 3 indicates the bias θ^* for several values of γ_{YZ}, γ_{XZ}, and n. An examination of the tabled values leads to the following conclusions:

1.) Bias increases with n (unless γ_{YZ} is 0 or γ_{XZ} is 0 or 1).

2.) For fixed γ_{XZ} (not 0 or 1) and n, bias increases with γ_{YZ}.

3.) For fixed γ_{YZ} (not 0 or 1) and n, bias first increases and then decreases with γ_{XZ}.

Table 2

Minimizing the direct relation to Y and maximizing the direct relation to X is the safest way to achieve small bias. θ^* approaches its maximum rapidly even for small values of n. Large n is less damaging when γ_{XZ} is large and γ_{YZ} is small, though the necessary value of γ_{XZ} increases rapidly with γ_{YZ}. For n=500 and γ_{YZ}=0.1, γ_{XZ} must be greater

BEST COPY AVAILABLE

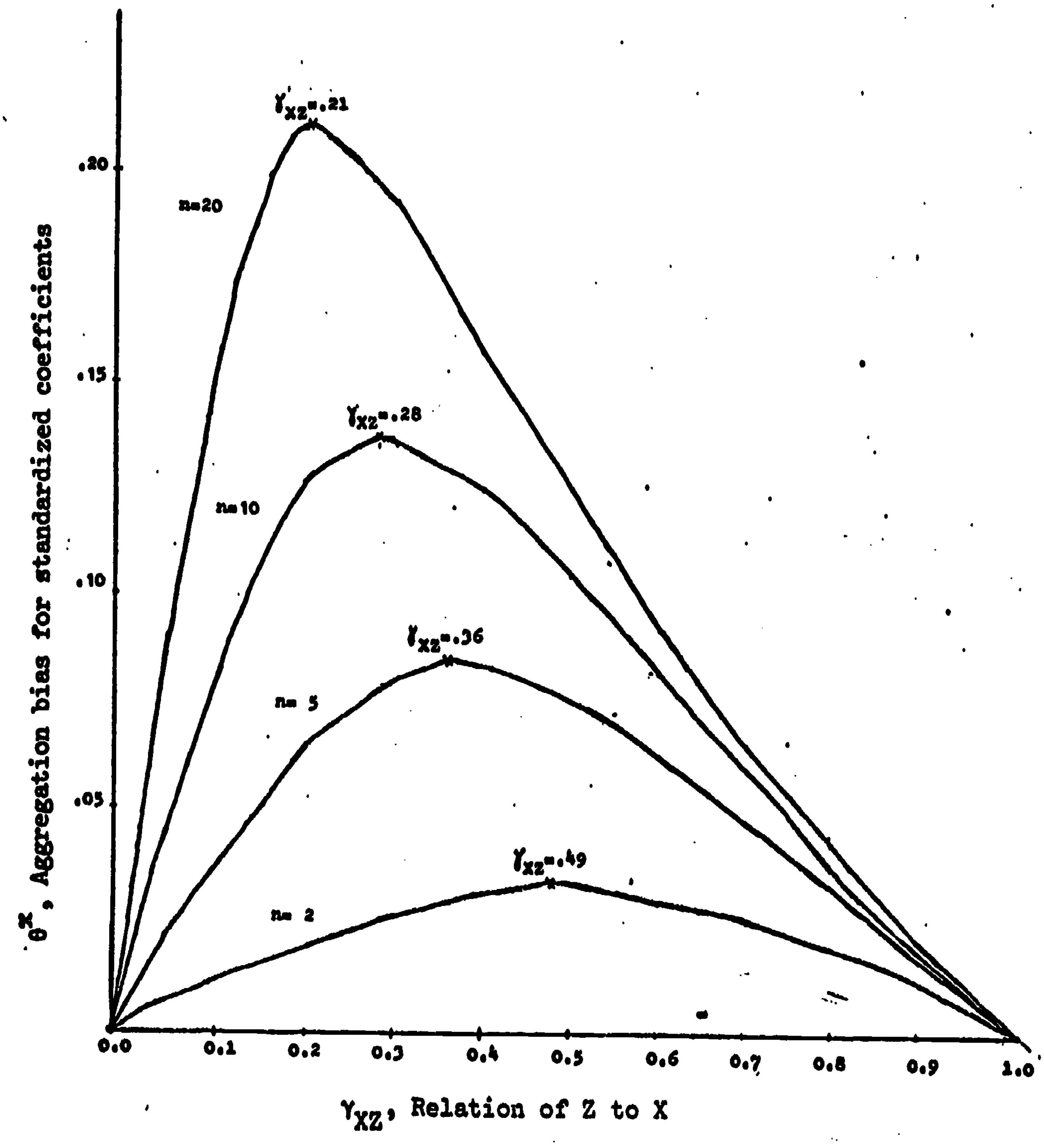


Figure 2. Aggregation bias θ^* as a function of γ_{XZ} and group size n (γ_{YZ}=0.10)

ERIC

BEST COPY AVAILABLE

Table 2. Predicted Bias in Estimating Standardized Regression Coefficient θ^*_{YX} from Grouped Data as a Function of Group Size n, γ^*_{YZ}, and γ^*_{XZ}

Group size n	θ^*- Magnitude of the Bias $[E(B^*_{YX})-\beta^*_{YX}]$									
	γ^*_{YZ}=	0.2			0.5			0.8		
	γ^*_{XZ}=	0.2	0.5	0.8	0.2	0.5	0.8	0.2	0.5	0.8
2		0.037	0.060	0.035	0.092	0.150	0.088	0.148	0.240	0.140
4		0.103	0.129	0.059	0.257	0.321	0.148	0.411	0.615	0.237
5		0.132	0.150	0.065	0.331	0.375	0.162	0.529	0.600	0.259
11		0.274	0.214	0.078	0.686	0.536	0.195	1.110	0.857	0.311
20		0.415	0.248	0.083	1.036	0.620	0.208	1.658	0.991	0.333
50		0.636	0.277	0.087	1.589	0.693	0.218	2.543	1.109	0.349
100		0.766	0.288	0.089	1.916	0.721	0.222	3.066	1.153	0.354
500		0.914	0.298	0.090	2.285	0.735	0.223	3.657	1.190	0.359

BEST COPY AVAILABLE

than 0.60 to have bias less than 0.1. For n= 500 and γ_{YZ}=0.2, γ_{XZ} must be greater than 0.78 to achieve the same results.

The expected bias can exceed 1 with large n and $\gamma_{YZ} > _{XZ}$. This should be a further warning against choosing a grouping variable strongly related to Y·X and against having a large number of observations per group. On the other hand, the relatively small bias associated with small γ_{YZ} offers some hope for reasonable estimates from data grouped by a Category I variable.

For Category I variables grouping bias affects efficiency in addition to the variance of $B_{\bar{Y}\bar{X}}$. The mean squared error for $B_{\bar{Y}\bar{X}}$ for Category I is

$$MSE(B_{\bar{Y}\bar{X}}) = V(B_{\bar{Y}\bar{X}}) + \theta^2(Z_I).$$

So,

$$Eff_I(B/b) = \frac{MSE(b_{YX})}{MSE(B_{\bar{Y}\bar{X}})}$$

$$= \frac{V(b_{YX})}{\theta^2 + V(B_{\bar{Y}\bar{X}})}$$

$$= \frac{E[SS(\bar{X})]}{E[SS(X)]} \quad \frac{1}{1+\theta^2 E[SS(X)]}$$

$$< \frac{E[SS(\bar{X})]}{E[SS(X)]} = \eta^2_X \quad .$$

That is, the correlation ratio is an upper bound for the efficiency of Category I grouping.

Though Category I grouping is generally less efficient that Category III grouping, it can be more efficient that Category II or Category

BEST COPY AVAILABLE

IV grouping. When γ_{XZ} is large and γ_{YZ} small, Category I variables behave like Category III variables. Though the resulting regression estimator from Category I grouping includes a small amount of bias, it will most likely be more stable (smaller mean square error) than an unbiased estimate from either Category II or Category IV grouping under similar conditions (number of groups and distribution of observations among groups).

2. Category II -- $\gamma_{YZ}\neq 0$, $\gamma_{XZ}=0$

Category II contains grouping variables Z_{II} which are related to $Y(\gamma_{YZ}\neq 0)$ and are not related to X $(\gamma_{XZ}=0)$. Since $\gamma_{XZ}=0$,

$$\begin{aligned}\theta(Z_{II}) &= E(B_{\bar{Y}\bar{X}(Z_{II})}) - E(b_{YX}) \\ &= \gamma_{YX} - \gamma_{YX} \\ &= 0\end{aligned}$$

Thus estimates derived from data grouped by a Category II variable are unbiased.

The conclusions for Category II grouping are not surprising. When Z is a Category II variable, Equation (10a) is the original model (Equation (1)) where the "other" causes represented by u have been divided into two parts (Z and w), both independent of X. Thus unbiased estimates are expected under the OLS assumptions.

BEST COPY AVAILABLE

Category II variables are hard to find. None of the more the 200 pairs of parameter estimates, γ_{YZ} and γ_{XZ}, from the empirical data described in Section 5 satisfactorily meet the conditions for Category II grouping. No doubt such variables can be constructed by some orthogonalization procedure, but there are other categories of variables which yield unbiased estimators with greater efficiency. Henceforth, Category II will receive little attention.

3. Category III -- $\gamma_{YZ}=0$, $\gamma_{XZ}\neq 0$

Category III consists of all variables Z_{III} which are related to Y only through X. Systematic grouping on the independent variable falls in Category III. Z_{III} may be an explicit ordered function of X such as the decile rank of X. With this kind of grouping, the within-group distributions of X do not overlap. It is also possible that Z involves some random component (v) which allows the within-group distributions of X to overlap. The presence or absence of overlap is irrelevant in the determination of bias but can affect efficiency.

Since $\gamma_{YZ}=0$, equations (10a) and (17a) reduce to

$$Y= \alpha+\gamma_{YX}X + w$$

and

$$\bar{Y} = \alpha+ \gamma_{YX}\bar{X} + \bar{w} \quad .$$

These equations are like (1) though the regression parameters and disturbance term have been relabeled. Thus for Category III grouping, the simple model and the modified model with the grouping variable incorporated are the same and estimate the same β_{YX}.

BEST COPY AVAILABLE

Thus, the least-squares estimators from data grouped on a variable Z which is related to X but not to Y•X are unbiased for any value of β_{YX}, which can be expressed in equation form by saying that

$$\theta\ (Z_{III}) = 0 .$$

The bias and efficiency of Category III grouping has been studied extensively, dating back at least to Prais and Aichinson's work (1954). Most variables systematically related to X do not strictly satisfy the condition $\gamma_{YZ} = 0$ and thus exhibit some minimal bias. If this condition is relaxed so that γ_{YZ} is considered zero if it does not exceed three times its standard error ($\gamma_{YZ}<3SE(\gamma_{YZ})$), there are generally several Category III variables in any large study. This is fortunate since Category III estimates are always unbiased and can be highly efficient (Prais and Aitchinson, 1954). If such variables do exist in a study, the remaining decision should focus on how to best utilize them to optimize the precision of parameter estimation.

4. Category IV -- $\gamma_{YZ}=0$, $\gamma_{XZ}=0$

Category IV contains all variables Z_{IV} which have no relation to either X or Y. Student weight in ten-pound units for the study of the achievement-on-aptitude regression is an example of a Category IV variable. Z_{IV} might also be a variable generated by assigning numbers randomly to individual observations, such as a student ID. Category IV grouping, alternatively called unsystematic or random grouping, results in random groups of (X,Y) observations.

BEST COPY AVAILABLE

When $\gamma_{YZ} = 0$ and $\gamma_{XZ} = 0$, it follows that

$$E(B_{\bar{Y}\bar{X}}) = E(b_{YX}) = \gamma_{YX} = \beta_{YX}.$$

Hence,

$$\theta\ (Z_{IV}) = 0$$

So for any Z_{IV}, $B_{\bar{Y}\bar{X}}$ is an unbiased estimator of β_{YX}.

The interpretation of this result is straightforward. Estimating β_{YX} from the means of g randomly formed groups is statistically equivalent to estimating β_{YX} from a sample of size g drawn randomly from the N observations or from the g strata means where the strata have been randomly formed (Hansen et al., 1953). In either case, the random process does not alter any preexisting relations among the variables. All variances and covariances among the variables decrease in proportion to the number of observations in a group for fixed group size n for Category IV grouping. This proportional reduction in magnitude does not alter the estimate of the regression coefficient.

Category IV variables may not be the best choices for grouping when efficient estimates are desired because of the difficulty of obtaining an adequate number of groups to overcome the marked efficiency reduction (Feige and Watts, 1973). Unfortunately, in many cases Category IV variables may be the only ones for which the investigator has sufficient understanding to form groups.

Using β_{XZ} and β_{YZ} to Predict Bias. One interesting finding is that the investigator need not eactually estimate γ_{YZ} and γ_{XZ} to determine the possible bias from a given grouping method. If β_{XZ} $(=\gamma_{XZ})$ and $\beta_{YZ}(= \frac{\sigma_Y}{\sigma_Z}\rho_{YZ})$

BEST COPY AVAILABLE

are determinable, (this is the case when either X or Y has been collected anonymously so that pairs of observations cannot be matched at the individual level), an upper bound for aggregation bias can be estimated under most prevailing conditions. This is accomplished by substituting β_{YZ} for γ_{YZ} in the bias formula (20) to get

$$\pi = \frac{\beta_{YZ}}{\gamma_{YZ}} (\theta).$$

Thus the investigator need not be hampered by response anonymity in choosing the "best" grouping variable for his study.

Estimating ρ_{XY} From a Systematic Grouping Procedure. The results for the bivariate case also suggest that estimates of ρ_{XY} can be from a systematic grouping procedure. The standardized regression coefficient β^*_{YX}(The "*" designates standardized parameters and estimators.) and the zero-order correlation coefficient ρ_{XY} are equal so that the grouped estimator B_{YX} is an estimate of both B^*_{YX} and r_{XY} (and thereby an estimate of both the regression and correlation coefficients in the standardized case). So "good" methods of estimating b are also good methods for estimating r when the original observations have first been standardized. This result should prove useful for persons interested only in correlation coefficients.

The Taxonomy as a Guide for Investigation. The main implication from the above discussion is that the investigator should consider the relations of the alternative grouping variables to the study variables before collecting his data, using such prior knowledge as is available. This will

BEST COPY AVAILABLE

enable him to collect information only on those grouping variables which yield estimates having the desired properties.

If the investigator demands an unbiased estimate of β_{YX}, then, under the assumptions of the model, variables from Categories II, III, and IV are satisfactory. While Category IV variables can always be found or created, they are relatively inefficient. Category III variables can be highly efficient, yielding large values of SS($\bar{X}$). The efficiency of Category II grouping is no better than for Category IV grouping because observations are assigned to groups essentially randomly with respect to X. Category III variables are clearly the best choices for data aggregation.

Category I variables yield biased estimates though the bias can be small for large γ_{XZ} and small γ_{YZ}. Category I estimators are less efficient than those from Category II or Category IV grouping. If small bias is tolerable and Category III variables are hard to find, Category I grouping may be advisable.

Most of the discussion has assumed that an investigator has the original observations and can choose his own grouping procedure. Data can be available in aggregated form only, however; e.g., when individual data have been aggregated for economy of storage or for confidentiality. The grouping variables that generally appear under these circumstances are geographic variables such as "state" and "census tract", and school system delimiters such as "school", "teacher", and "classroom". These grouping variables are generally related to X and Y•X and are thereby subject to the criticisms of Category I grouping. Regression estimates determined under these conditions should be interpreted cautiously.

BEST COPY AVAILABLE

4. An Empirical Example--The Regression of Academic Self-Rating on Achievement

The literature on aggregation offers very little in the way of concrete demonstrations of the likely magnitude of aggregation bias in realistic cases. This sort of work is quite important in informing the substantive researcher as to the likely consequences of alternative grouping strategies. Information collected on incoming freshmen at a large midwestern university will serve as the data base for an empirical demonstration of the grouping methods described in the taxonomic approach to grouping.

Over 300 measures of the abilities, attitudes, and interests of the students were collected in the original study. Approximately 20 of these measures are used in the present analysis. To avoid confounding missing-data problems and aggregation problems, the analyses are performed only on the 2676 students with complete information on all measures.[8]

Regression Model from Ungrouped Data. The parameter of interest is the regression coefficient from the regression of a composite self-rating academic abilities (SRAA) on achievement test performance (ACH). The main reason for the proposed order is a concern for clarity of illustration as the chosen causal ordering appears to be more informative with regard to the effects of grouping than the reverse order.

The two primary variables were chosen so that the example reflects the kind of study where anonymity of response might be a problem. None of the empirical data was collected anonymously so that the results

BEST COPY AVAILABLE

from treating the data as completely identified can be compared to the results when anonymous collection of some information is assumed.

Using the 2676 observations at the individual level, the equation relating achievement to academic self-rating is

$$SRAA = -29.136 + .344(ACH)$$

so that $b_{YX} = .344$. Also,

$$SE(b_{YX}) = 0.011$$

$$r_{XY} = 0.529$$

and $$R^2_{XY} = 0.281.$$

Identification of Grouping Variables. The grouping variables for the example are described in Table 3. They were selected from available information because (1) previous studies of the relation between achievement and academic self-rating included similar indicators (e.g., parental income (PARINC), father's education (FATHED)) or (2) the frequency distribution of the variable and its correlations with ACH and SRAA (see Table 4) suggested that it might represent a particular taxonomic category.

Table 3

The seventeen grouping variables are mostly student reports of parental characteristics and of their own background and attitudes as measured by single-item scales. These single-item indicators generally have low reliability but are easily manageable for grouping because of their limited number of response choices. Some, however, turn out to be surprisingly good grouping variables.

BEST COPY AVAILABLE

Table 3. Variable i dentifications, descriptions, and the number of groups formed after data aggregation.

Variable Identification	Variable Description	Number of Groups After Aggregation
ID2	Last 2 digits of student indentification	100
ID1	Last digit of student identification	10
HSGPA2	High school's report of student's grade point average on a 4-point scale (highest 2 digits)	23
SAT2	Highest 2 digits of Total Score from the Scholastic Aptitude Test	13
ACH2	Highest 2 digits of Total score from the Achievement Battery	10
PARINC	Student's best estimate of 1970 parental income before taxes	10
REPGPA	Student's report of average grade in secondary school	7
FATHED	Student's report of highest level of formal education obtained by his father	6
SRAA2	Highest digit and sign of composite academic self-opinion	5
ANTHIDEG	Student's anticipated highest academic degree	5
HSMATH	Student's report of number of semesters of high school mathematics	5
HSPHYS	Student's report of number of semesters of high school physical sciences	5
NOBOOK	Student's report of number of books in the home	5
PARASP	"What is the highest level of education that your parents hope you will complete?"	5

BEST COPY AVAILABLE

BEST COPY AVAILABLE

Variable Identification	Variable Description	Number of Groups After Aggregation
CLIMP	"My grades are markedly better in courses that I see I will need later."	4
COLEFF	"I often wonder if four years of college will really be worth the effort."	4
QCJOB	"I often wish that I were offered a good job now so I wouldn't have to spend four years in college."	4

BEST COPY AVAILABLE

Table 4 contains the means, standard deviations, and skewness coefficients for each study variable. Particularly interesting is the behavior of the four- and five-choice scales. Though all seven are similar with respect to the magnitude of their standard deviations, four have highly negatively skewed distributions (HSMATH,QCJOB,PARASP, and NOBOOK). In general the large skewness values result from an uneven distribution of observations among groups and a high degree of consolidation at one end of the scale or the other. This sort of distribution is not conducive to precise estimation. So it might be expected that estimates from data grouped by these variables would be less precise than the estimates from data grouped by variables with a more even spread of observations among groups and a more symmetric distribution.

Table 4

Another factor to consider in examining the grouping variables is the relative coarseness of the grouping as represented by the number of groups formed. In general characteristics resulting in the formation of more groups yield more precise estimates. In fact, the relative efficiency of grouping on two variables with approximately the same relations to SRAA and ACH, is largely determined by the differences in the number of groups formed by each (of course this result is tempered by uneven distribution of observations among the groups).

Categorizing the Grouping Characteristics. The process whereby grouping characteristics are classified into taxonomic categories requires more information about the grouping variables than is provided in Table

BEST COPY AVAILABLE

Table 4. Means, standard deviations, and skewness coefficients of study variables.

Variable Name	Mean	Standard Deviation	Skewness
SRAA	0.008	10.057	0.223
ACH	84.766	15,463	-0.364
ID2	49.561	29,126	0.003
ID1	4.453	2.865	0.011
HSGPA2	3.157	0.469	-0.067
SAT2	10.235	1.798	0.064
ACH2	8.024	1.572	-0.333
PARINC	6.308	2.289	-0.234
REPGPA	3,203	1.284	0.232
FATHED	3.987	1.418	-0.321
SRAA2	0.005	0.689	0.399
ANTHIDEG	3.867	0.959	0.687
HSMATH	4.332	0.879	-1.260
HSPHYS	2.623	0.977	0.319
NOBOOK	4.104	0.978	-0.769
PARASP	4.458	0.626	-1.523
CLIMP	2.201	0,821	0.304
COLEFF	2.695	0.951	-0.209
QCJOB	3.330	.821	-1.151

BEST COPY AVAILABLE

4. Table 5 contains estimates of the unstandardized regression coefficients, γ_{YZ} and γ_{XZ}, their standardized counterparts, γ^*_{YZ} and γ^*_{XZ}, the zero-order correlation of Y and Z, ρ_{YZ}, and the between-groups standard deviation of the independent variable ACH, $\sigma_{\bar{X}}$, for each of the grouping characteristics.

Table 5

Applying the criterion that a parameter must exceed three times their standard errors to be considered nonzero leads to the following category assignments for the grouping characteristics:

	$\hat{\gamma}_{YZ} \geq 3SE(\hat{\gamma}_{YZ})$	$\hat{\gamma}_{YZ} < 3SE(\hat{\gamma}_{YZ})$
$\hat{\gamma}_{XZ} \geq 3SE(\hat{\gamma}_{XZ})$	Category I HSGPA2 SAT2 ANTDEG REPGPA FATHED SRAA2 HSMATH NOBOOK PARASP COLEFF QCJOB	Category III ACH2 PARINC HSPHYS CLIMP
$\hat{\gamma}_{YZ} \leq 3SE(\hat{\gamma}_{YZ})$	Category II (NONE)	Category IV ID2 ID1

As mentioned previously, no characteristics belong to Category II, and the number falling in Category I is large. SRAA2 and ACH2 are special cases within Categories I and III, respectively. They are the best approximations to what Blalock (1964) and Hannan (1971) have called "grouping on the dependent variable" and "grouping on the independent variable."

BEST COPY AVAILABLE

Table 5. Estimates of the unstandardized regression coefficients (γ_{YZ} and γ_{XZ}), the standardized regression coefficients (γ^*_{YZ} and γ^*_{XZ}), the zero-order correlation of Y and Z (ρ_{YZ}), and the between-groups standard deviation of ACH ($\sigma_{\bar{X}}$) for each grouping variable from the regression of SRAA an ACH.

Variable Name	Parameter Estimates					
	$\hat{\gamma}_{YZ}$	$\hat{\gamma}_{XZ}$	$\hat{\gamma}^*_{YZ}$	$\hat{\gamma}^*_{XZ}$	$\hat{\rho}_{YZ}$	$\sigma_{\bar{X}}$
ID2	.003	.011	.008	.020	.019	2.918
ID1	-.037	-.225	-.011	-.042	-.033	1.208
HSGPA2	2.640	17.636	.123	.535	.370	8.525
SAT2	2.272	7.114	.406	.827	.566	12.833
ACH2	.447	9.670	.070	.983	.522	15.203
PARINC	.121	.470	.028	.070	.064	1.891
REPGPA	-2.025	-5.900	-.258	-.490	-.455	7.874
FATHED	.516	1.512	.073	.139	.145	2.321
SRAA2	11.956	10.690	.819	.476	.885	7.427
ANTHIDEG	1.956	2.512	.186	.156	.264	2.455
HSMATH	-.757	8.429	-.066	.479	.202	7.556
HSPHYS	.469	5.033	.046	.318	.209	5.635
NOBOOK	1.252	2.312	.122	.146	.196	2.281
PARASP	2.212	1.628	.138	.066	.186	1.192
CLIMP	-.043	2.767	.003	.147	.074	2.508
COLEFF	1.277	2.173	.121	.134	.189	2.223
QCJOB	1.775	1.986	.145	.105	.199	1.770

BEST COPY AVAILABLE

Estimates of Regression Coefficients from Different Grouping Methods. Table 6 contains estimates of the unstandardized regression coefficients, their standard errors, the observed and predicted grouping bias (with Θ and π), and estimates of the mean squared error for each grouping method. The grouping variables have been organized by category (in the order IV, III, and I) and by the size of the observed bias within category. ACH2 and SRAA2 have been assigned to special subcategories III' and I' in recognition of their unique relation to the main variables.

Table 6

In general the estimates conform to expectations though the bias and mean squared error (MSE) are enormous for some Category I groupings. Category IV grouping yielded estimates with relatively small bias. In fact, the precision (small bias and mean squared error) of the estimate from ID2 is exceeded only by grouping on the independent variable (by ACH2). But it took ten times as many groups to achieve this level of accuracy.

The magnitude of the bias from grouping by ID1 (10 groups) and its MSE represent, respectively, a ten- and seven-fold increase over the corresponding values from ID2 grouping (100 groups). ID1 performs less well than certain variables from other categories, especially for those variables forming as many groups. The Category III variables that form 10 groups (ACH2 and PARINC) yield smaller bias and smaller MSE than ID1. The two Category I variables that form more than 10 groups (SAT2 and HSGPA2) result in larger bias but smaller MSE. Finally, HSPHYS

BEST COPY AVAILABLE

Table 6. Grouping bias in unstandardized regression coefficients from the regression of SRAA on ACH.
Ungrouped Estimates: $b_{YX} = .344$, $SE(b_{YX}) = .108$, $MSE(b_{YX}) = .0001$

Grouping Variables	$B_{\bar{Y}\bar{X}}$	$SE(B_{\bar{Y}\bar{X}})$	Observed Bias[a]	Predicted Bias θ[b]	π[c]	$MSE(B_{\bar{Y}\bar{X}})$[d]
Category IV						
ID2	.339	.0496	-.005	.003	.007	.0025
ID1	.286	.1184	-.058	.047	.168	.0175
Category III'						
ACH2	.344	.0398	0	.002	.007	.0016
Category III						
PARINC	.362	.0850	.018	.082	.192	.0075
HSPHYS	.370	.0592	.026	.062	.282	.0041
CLIMP	.465	.2568	.121	-.012	.263	.0840
Category I'						
SRAA2	1.200	.0406	.856	.846	.912	.7310
Category I						
HSMATH	.268	.0592	-.076	-.066	.203	.0062
SAT2	.435	.0434	.091	.099	.137	.0100
HSGPA2	.454	.0186	.110	.098	.294	.0122
FATHED	.589	.1057	.245	.285	.567	.0707
REPGAP	.593	.0401	.249	.235	.413	.0631
NOBOOK	.858	.0765	.514	.520	.838	.2690
COLEFF	.944	.0958	.600	.497	.778	.3680
ANTHIDEG	1.058	.1741	.714	.730	1.088	.5401
QCJOB	1.142	.2695	.798	.748	1.025	.7078
PARASP	1.260	.4758	.926	.987	1.240	1.0637

[a] Observed Bias = $B_{\bar{Y}\bar{X}} - b_{YX}$

[b] $\theta = \hat{\gamma}_{YZ}\hat{\gamma}_{XZ}\left(\frac{\hat{\sigma}^2_{\bar{Z}}}{\hat{\sigma}^2_{\bar{X}}} - \frac{\hat{\sigma}^2_Z}{\hat{\sigma}^2_X} \right)$

[c] $\pi = \frac{\hat{\beta}_{YZ}}{\hat{\gamma}_{YZ}} (\theta)$

[d] $MSE(B_{\bar{Y}\bar{X}}) = V(B_{\bar{Y}\bar{X}}) + (B_{\bar{Y}\bar{X}} - b_{YX})^2$

BEST COPY AVAILABLE

yields smaller bias than ID1, and both HSPHYS and HSMATH yield smaller M S E though they form only half as many groups. Clearly, random grouping should be avoided unless many groups can be formed and there are no other variables with systematic relations to the main variables that also yield a larger number of groups.

Three of the four Category III variables yield highly satisfactory estimates of β_{YX} with small M.S E 's. The exception is CLIMP, whose estimate has only the ninth smallest bias and the eleventh smallest M S E. The fact that all of CLIMP's relation to SRAA operates through ACH plus the small number of groups formed (4) might account for the ambiguous results with this grouping variable.

Three Category I variables, HSMATH, SAT2, and HSGPA2, yield relatively precise (within .11) estimates of β_{YX}. All have substantial zero-order correlations with ACH, zero-order correlations with SRAA that are clearly smaller and result in large between-groups standard deviations of ACH. Their M S E are also respectably small. An investigator who uses a grouping variable of their calibre will not reach conclusions that differ in any drastic manner from the investigator who works with the individual-level observations.

The remaining Category I variables, including SRAA, yield particularly poor estimates of the relation of ACH to SRAA. B_{YX} ranges from .59 (FATHED) to an astonishingly large 1.26 (PARASP) for these variables, almost four times the ungrouped effect (5.9)! Their estimates are also unstable with mean squared errors ranging from .06 to 1.06. The M S E for PARASP is 10,000 times larger than the M S E for the estimate from ungrouped data.

BEST COPY AVAILABLE

The example again demonstrates that grouping on the dependent variable is disastrous in terms of bias. The unmeasured factors represented by the disturbance term in the initial linear model (Eq. 1) are confounded with the effects of the primary regressor to such a degree that the relation of ACH to SRAA is unrecognizable.

Overall,there are clear distinctions between the performance of Category I variables and the other grouping variables. In every case, the standard error of the regression estimate from a Category I variable is larger than the observed bias; all regression estimates from Category III and IV grouping fall within one standard error of b_{YX}. So one gains some knowledge of the accuracy of an estimate by simply categorizing grouping characteristics and then examining standard errors.

There appear to be other warning signals for poor estimation from Category I variables, even when confidentiality considerations prevent direct estimation of σ_{YX}. Seven of the eight Category I variables that yielded large bias also had zero-order correlations with SRAA that exceeded their zero-order correlations with ACH (i.e., $\rho_{YZ} > \gamma^*_{XZ} (= \rho_{XZ})$; for REPGA, ρ_{YZ} and γ^*_{XZ} have approximately the same magnitude (-.455 and -.490, respectively).). For SRAA2, ANTHIDEG, QCJOB, and PARASP, γ^*_{YZ} is larger than γ^*_{XZ}, though this is a comparison of a partial correlation with a zero-order coefficient.

There is additional observable warning for the single-item scales in Category I. Grouping by six of the eight Category I variables of this type (excepting HSMATH and REPGPA) results in small between-groups standard deviations in ACH. ANTHIDEG yields the largest $\sigma_{\bar{X}}$ (2.46) and PARASP yields the smallest (1.19).

BEST COPY AVAILABLE

PARASP is an extremely poor grouping characteristic. Of the five choices for the PARASP question, 2612 (97%) responded that their parents hoped that they would complete college (response 4) or obtain a graduate or professional degree (response 5). Thus, PARASP really distinguishes between only two parental aspiration choices and operates as if there were only two groups. It is not surprising, then, that PARASP yields such a poor estimate of β_{YX}. Grouping by PARASP is the negation of the principles for precise estimation that were discussed in earlier sections. The grouping would be coarse even if the observations were evenly distributed among groups; its relations to ACH and SRAA are the reverse of good practice; it barely maintains between-groups variation in ACH, much less maximizing it; and the distribution of observations among the groups is so uneven that two groups rather than five would have been sufficient.

There are other Category I variables that are little better. Essentially the same statements can be made about grouping by QCJOB as were made for grouping by PARASP. Again, there are few initial groups (4), $\gamma^*_{YZ} > \gamma^*_{XZ}$, $\sigma_{\bar{X}}$ is small (1.77), and the observations are unevenly distributed (86% (2272 out of 2637) in the two highest categories.). ANTHIDEG suffers from similar shortcomings with less than 100 observations for its two lowest groups.

a. Predicted Bias vs. Observed Bias.

Despite the high likelihood of specification and measurement errors in the simple model examined, bias predictions stand up well in most cases. For every grouping where the observed bias is greater than

BEST COPY AVAILABLE

.2, the predicted θ is also greater than .2. For every grouping variable yielding observed bias less than .1, the predicted bias is also less than .1.

The predicted value of θ can be considered misleading for only two variables, ID1, and CLIMP. In the case of ID1, it is the matter of sign reversal that troubles us and not the difference in magnitude between predicted and observed bias. The predicted θ for CLIMP would lead one to expect a more precise estimate than actually occurred. The observed bias is not too distressing however.

For the empirical data, the π values are larger than the observed bias in every case. If the grouping variables are ordered from smallest to largest π values, the variables with lowest values (less than .3) are the Category III and IV variables plus the 3 Category I variables with the smallest observed bias (HSMATH, SAT2, HSGPA2).

b. Composite Estimates From Multiple Grouping Variables.

The above findings suggest that an investigator can separate those grouping characteristics which lead to reasonably accurate estimates from those providing extremely misleading ones in empirical studies similar to ours. Once this separation has been accomplished, the investigator can choose the characteristic with the smallest predicted bias. Better yet, he can use the available information about each characteristic and its expected bias to form a weighted composite estimate. The latter can be accomplished by weighting grouped estimates in an inverse proportion to their predicted bias. One can also take the standard errors of the grouped estimates into account by giving additional weight to the more stable estimates.

BEST COPY AVAILABLE

Two examples of the suggested compositing were carried out. In Example (A) knowledge of σ_{YX} was treated as unknown and thus the π values are used for weighting. The five grouping variables, excluding ACH2 and ID2, with the smallest predicted bias were used in each example. Weights were determined by (i) predicted bias only and (ii) by the product of the predicted bias and the standard error of the grouped estimate.

The results of the compositing process are very satisfactory. When σ_{YX} is known, the composite grouped estimates of β_{YX} are (i) .355 and (ii) .345. When σ_{YX} is unknown, the composite grouped estimates of β_{YX} are (i) .362 and (ii) .356. Composite estimate A(ii) performs about as well as grouping on the independent variable ACH2. Composite estimate A(i) is closer to the actual β_{YX} than the estimates from any of the grouping variables except ACH2 and ID2. Composite estimates B(ii) and B(i) do nearly as well, equaled or exceeded only by the estimates from ACH2, ID2, and PARINC. Clearly, judicious use of weighted compositing of grouped estimates, in conjunction with the bias prediction, can lead the conscientious investigator to precise estimates from grouped data.

Estimating the Correlation between ACH and SRAA From Grouped Data. The zero-order correlation between SRAA and ACH, from the grouped observations, ρ_{YX}, can be estimated by employing the procedures prescribed for estimating β_{YX} from grouped data. The only modification is that the investigator standardizes his observations before aggregating them. Once this is done, the coefficient from the regression of ZSRAA on ZACH

BEST COPY AVAILABLE

at the individual level is an unbiased estimate of the correlation coefficient between SRAA and ACH; that is, $E(b^{*}_{YX}) = \beta^{*}_{YX} = \rho_{YX}$. Thus, the ZSRAA-on-ZACH regression using the grouped observations yields estimates of ρ_{YX} from grouped data. Under these circumstances, comparisons of $B^{*}_{\overline{YX}}$ with b^{*}_{YX} are checks for grouping bias in estimating the individual-level correlation coefficient.

Table 7 illustrates the results of estimating the correlation between SRAA and ACH from grouped observations. The standardization process resulted in fewer groups for ZID2 (35) than for ID2 (100) and for ZACH2 (6) than for ACH2 (10). The increased coarseness of these two grouping variables may account for their poorer estimation in the standardized case relative to their accuracy in the unstandardized case. Also, HSGPA2 and REPGPA were not used in this phase of the investigation.

Table 7

Most of the statements made about the precision of grouped estimates in the unstandardized case hold for the standardized case. The grouping variables tend to maintain the rank (in terms of observed bias and MSE) that they received in the unstandardized case. The grouping variables yielding the smallest bias and the smallest MSE in the standardized case are the standardized counterparts of the best variables in the unstandardized case. Again, every Category III and Category IV estimate falls within 1 SE($B^{*}_{\overline{YX}}$) of ρ_{YX} while every Category I estimate deviates by more than 1 SE($B^{*}_{\overline{YX}}$) from ρ_{YX}.

BEST COPY AVAILABLE

Table 7. Grouping bias in estimating the correlation coefficient between SRAA and ACH from the standardized regression coefficient estimates from grouped observations.

Ungrouped Estimates: $r_{YX} = b^*_{YX} = .529$, $SE(b^*_{YX}) = .0032$

Grouping Variables	$B^*_{\bar{Y}\bar{X}}$	$SE(B^*_{YX})$	Observed Bias[a]	Predicted Bias θ^*[b]	π^*[c]	$MSE(B^*_{\bar{Y}\bar{X}})$[d]
Category IV						
ZID2	.500	.1020	-.029	.017	.040	.0112
ZID1	.442	.1942	-.087	.075	.225	.0452
Category III'						
ZACH2	.542	.1003	.013	.019	.142	.0101
Category III						
ZPARINC	.558	.1390	.029	.129	.295	.0202
ZHSPHYS	.571	.1057	.042	.095	.433	.0130
ZCLIMP	.717	.4863	.188	-.016	.401	.2718
Category I'						
ZSRAA2	1.832	.0615	1.303	1.395	1.507	1.6940
Category I						
ZHSMATH	.414	.0287	-.115	-.100	.307	.0140
ZSAT2	.671	.0700	.142	.150	.210	.0251
ZFATHED	.911	.1818	.382	.440	.874	.1790
ZNOBOOK	1.335	.1308	.806	.800	1.285	.6668
ZCOLEFF	1.461	.1640	.932	.765	1.194	.8917
ZANTHIDEG	1.631	.3095	1.102	1.117	1.586	1.3102
ZQCJOB	1.853	.4327	1.224	1.188	1.630	1.6854
ZPARASP	1.946	.8473	1.417	1.518	2.048	2.7330

[a] Observed Bias = $B^*_{YX} - b^*_{YX}$

[b] $\theta^* = Y^*_{YZ} Y^*_{XZ} \left(\frac{1 - \sigma^2_X}{\sigma^2_X}\right)$ (remembering that $\sigma^2_X = \sigma^2_Z = \sigma^2_Z = 1.$)

[c] $\pi^* = \frac{\rho_{YZ}}{Y^*_{YZ}} (\theta^*)$

[d] $MSE(B^*_{\bar{Y}\bar{X}}) = V(B^*_{\bar{Y}\bar{X}}) + (B^*_{\bar{Y}\bar{X}} - b^*_{YX})^2$

ERIC Full Text Provided by ERIC

BEST COPY AVAILABLE

a. Predicted Bias vs. Observed Bias.

The predicted bias (θ^* when σ_{YX} is known and π^* when σ_{YX} is unknown) provides as good a guide for selecting grouping variables as it did in the unstandardized regressions. θ^* is more than .05 smaller than the observed bias for only ZCLIMP and ZCOLEFF. π^* is always larger than the observed bias. This underestimation can cause problems for the investigator if he chooses to group by ZCLIMP, but the predicted bias for ZCOLEFF is large enough to eliminate it from further consideration.

b. Composite Estimates of ρ_{YX} from Grouped Observations.

Composite estimates of ρ_{YX} were determined from the weighted average of the estimates from the grouping variables ZID1, ZPARINC, ZHSPHYS, ZCLIMP, and ZHSMATH. They were:

A(i) $\hat{\rho}_{YX} = .548$ B(i) $\hat{\rho}_{YX} = .558$
(ii) $\hat{\rho}_{YX} = .531$ (ii) $\hat{\rho}_{YX} = .544$

The accuracy of the composite estimates based on the products of the expected bias and standard errors is exceeded only by grouping on ZACH2. Moreover, as in the unstandardized case, only ZACH2, ZID2, and ZPARINC provide estimates that are as accurate or more accurate than any of the composite estimates. Compositing of the best (excluding the independent variable) estimates from grouped observations appears to be a robust procedure that will afford greater confidence than the estimate from observations grouped on any single characteristic beside the independent variable.

6. Concluding Remarks

In section 5 the utility of the grouping concepts and methods developed in section 4 were demonstrated under realistic empirical conditions. The empirical evidence regarding the estimation of both β_{YX} and ρ_{YX} conformed to the predictions from the principle of incorporating the grouping characteristics as variables in the structural model, which, in turn, lead to the taxonomic categorization of grouping variables. The latter classification resulted in clusters of readily identifiable "good" and "bad" grouping variables under most aggregation conditions. Furthermore it was shown that if the investigator formed a weighted composite of estimates from several of his best grouping variables, his resulting estimate is invariably highly accurate.

Effective strategies for estimating simple linear regression coefficients and zero-order correlation coefficients have been demonstrated for the case when data aggregation is under the investigator's control and the grouping characteristics under consideration have at least an ordinal scale. To a certain degree, the results are generalizable to naturally aggregated data where some degree of disaggregation is feasible.

The next step in the investigation of change-in-units problems is to come to grips with the complications caused by nominal grouping characteristics and by multiple regression. This paper concludes with comments relevant to these two problems.

Nominal Grouping Characteristics. The development of sound procedures

BEST COPY AVAILABLE

for determining and predicting grouping effects when the grouping characteristic is nominal remains the most pressing aggregation problem in educational research. Cross-level inferences from aggregate sampling units such as schools are too frequent to overlook and condone without careful examination of the consequences. Unfortunately the sociological methods developed to date appear to be either too complicated or not sufficiently applicable to the analyses beyond the level of contingency tables (Goodman (1959), Iversen(1973)).

The approach favored by this investigator is to try to fit structural equation methods to this important case by in some way incorporating the nominal grouping characteristics into the model as was done previously with ordered characteristics. This approach allows the investigator to capitalize on the apparent strength of taxonomic reasoning in the analysis of grouping effects.

To achieve the end of incorporating the nominal grouping characteristics into the model, two approaches seem promising. Wiley (personal communication) points out that nominal characteristics (School) are in reality manifest representations of some latent characteristic or characteristics (community commitment to education, community resources). Latent structural analysis (or multiple discriminant analysis) can be employed to estimate the ordinal true values corresponding each group index and the relations between the primary variables and the latent grouping variable can then be estimated.

ERIC Full Text Provided by ERIC

BEST COPY AVAILABLE

A slightly less complex procedure that may also prove fruitful is to simply represent any nominal grouping characteristic forming g groups by g-1 dichotomous variables in the basic structural model. The model with incorporated grouping characteristic is, then

$$Y = \alpha + \gamma_{YX}X + \gamma_{YZ_{(1)}}Z_{(1)} + \ldots + \gamma_{YZ_{(g-1)}}Z_{(g-1)} + v$$

$$X = \lambda + \gamma_{YZ_{(1)}}Z_{(1)} + \ldots + \gamma_{YZ_{(g-1)}}Z_{(g-1)} + W \;.$$

If $R^2{}_{YX}$ is the coefficient of determination from the regression of Y on X, the the direct strength relation of Y to Z can be estimated from the square root of variation increase accounted for

$$\sqrt{R^2{}_{X.YZ_{(1)}} \cdots Z_{(g)} - R^2{}_{Y.X}}$$

due to the incorporation of the dichotomous regressors based on Z. The relations of Z to X can be estimated from

$$\sqrt{R^2{}_{X.Z_{(1)}} \cdots Z_{(g)}}$$

Neither approach yields perfect indices of the relations of a nominal Z to X and Y but both are worth further consideration as alternatives to those previously proposed. At least they present future investigators a starting point for refining the "structural equations" approach in the nominal case.

Grouping in the Multivariate Case. The examination of the effects of grouping in the multivariate case is a relatively new and still developing line of investigation. For much of the 1950's and 1960's, Prais and Aitchinson's results (no bias, efficency optimized by maximizing

BEST COPY AVAILABLE

the between-groups variation in the regressors) defined the state of knowledge on the topic. Investigators following up on their line of inquiry (Cramer, 196) focussed on strategies for optimal grouping without considering the possibility of bias in estimation.

Haitovsky's (1966) work provided the first major break from the optimal grouping perspective in the multiple regressor case. He studied alternative procedures for estimating multiple regression coefficients when the data are in the form of one-way.classification tables so that frequencies for crossclassification are lacking. Several of Haitovsky's findings are interesting but his most important contribution to the study of grouping effects is his empirical evidence that groupings on one independent variable can lead to biased estimates when the hypothesized model contains two or more independent variables. His data suggests that in the two regressor case, grouping on one regressor yields good estimates of the regressor's corresponding regression coefficient but a distorted estimate of the coefficient from the other regressor.

Recent work by Feige and Watts (1972) is even more informative and definitive in the multivariate case. They develop criteria for evaluating the analytical consequences of what they call "partial aggregation". The context for their results is the problems of performing micro-level analyses while preserving the condifentiality of data. Feige and Watts' (1972) investigation focusses on the differences between grouped and ungrouped estimators of the regression parameters rather than considering

BEST COPY AVAILABLE

bias directly in their equation work. They attribute any differences to one of three sources: (i) specification bias, (ii) bias introduced by a grouping transformation that is not independent of the disturbances or (iii) sampling error induced by the use of less information in the grouped regression.

The empirical example presented by Feige and Watts is illustrative of the variety of grouping possibilities for census or survey data. They established seven grouping rules based on demographic and financial asset indices and then examined three levels of consolidation for each grouping rule.

The one shortcoming of Feige and Watts from the perspective advanced by this present investigation is their failure to explicitly state the mechanism for selecting the "best" grouping rule when several options are available. Thus their methods require knowledge of the initial micro-model relations and thus facilitate description rather than prediction of bias. Investigations employing taxonomic classification with the "Structural equations" approach indicate that grouping bias can be predicted in the multivariate case. The process is complicated but not impossible. For example, there are eight categories of grouping variables included in the taxonomy for the two-regressor case. These categories are generated by the direct relations of Z to the dependent variable (γ_{YZ}) and to each of the independent variables (say, γ_{XZ} and γ_{WZ}). Any association between the regressors (nonzero γ_{XW} or γ_{WX}, depending on which is prior) can also affect bias under certain conditions. Figure 3

BEST COPY AVAILABLE

presents the path diagrams for each of the 16 subcategories from the taxonomy.

Figure 3

The results will not be discussed in detail. Table 8 gives some indications of the expected results with regard to bias. Several conclusions can be drawn from the table:

1. As long as Z has no direct relation to Y ($\gamma_{YZ} = 0$), no grouping bias can result.
2. When Z is directly related to the causally prior regressor (W in this example) and to estimates of its regression coefficient β_{YW} are always biased but unbiased estimates of β_{YX} are possible as long as $\gamma_{XZ}=0$.
3. When Z is directly related to the causally posterior regressor (X) and to Y, estimates of β_{YX} are always biased; in this case estimates of β_{YW} are biased whenever either γ_{WZ} or γ_{XW} is nonzero.

Table 8

The taxonomic strategies presented can easily become cumberso e as more regressors are included. More research is necessary to determine the efficiency of this approach especially compared to the procedures described by Feige and Watts.

BEST COPY AVAILABLE

Table 8. Estimated bias from grouped observations as a function of taxonomic category -- two regressor case.

Category	Specified Values of Parameters				Estimated Coefficients*	
	γ_{YZ}	γ_{XZ}	γ_{WZ}	γ_{XW}	β_{YX}	β_{YW}
1	≠0	≠0	≠0	0	B	B
1	≠0	≠0	≠0	≠0	B	B
2	≠0	≠0	0	0	B	U
2	≠0	≠0	0	≠0	B	B
3	≠0	0	≠0	0	U	B
3	≠0	0	≠0	≠0	U	B
4	≠0	0	0	0	U	U
4	≠0	0	0	≠0	U	U
5	0	≠0	≠0	0	U	U
5	0	≠0	≠0	≠0	U	U
6	0	≠0	0	0	U	U
6	0	≠0	0	≠0	U	U
7	0	0	≠0	0	U	U
7	0	0	≠0	≠0	U	U
8	0	0	0	0	U	U
8	0	0	0	≠0	U	U

* B= Estimator of regression coefficient from grouped data is biased

U= Estimator of regression coefficient from grouped data is unbiased

BEST COPY AVAILABLE

FOOTNOTES

[1] The Schools of Education at Stanford University and the University of Wisconsin--Milwaukee and the International Association for the Evaluation of Educational Achievement partially supported this research through release time and computer funds for data analysis. Kathe Magayne-Roshak and Donald Haumant deserve special thanks for typing the manuscript under horrendous conditions and time constraints.

Suzanne P. Wiviott has offered many helpful comments regarding the paper. Harry Lütjohann, Lars R. Bergman, and Ingram Olkin have also made substantial contributions to certain ideas expressed here. Two persons influenced this work to an extent far beyond which a simple mention can convey: Michael T. Hannan, for his continuing interchange of ideas regarding the problems of data aggregation and for his willingness to collaborate with the author thereby providing a broader forum for new developments, and Lee J. Cronbach, who has spent an enormous amount of time getting the author to develop his ideas more fully and carefully and to communicate his thoughts more clearly. The errors and misrepresentations that remain are solely the responsibility of the author.

[2] The contexts discussed in this section are in no way meant to be exhaustive in the area of data aggregation. Temporal aggregation, aggregation over commodities, aggregation of different responses within the individual have all received consideration in the literature of economics. Econometricians have also treated data aggregation models where the regression parameters are not constant but instead, vary from unit to unit at the micro-level but are constant at the macro-lével (Theil, 1964). In this investigation there is only a single regression parameter when there is one regressor.

[3] Iversen (1973) recently reviewed the methods for estimating ecological regressions and correlations from contingency tables, but the complexity of the approaches he suggests work against their utilization. Also, see Hauser (1969) for a discussion of the use of contextual variables in sociological research on individuals.

[4] Oddly enough, one of the first references to the inflationary effects of estimating correlation coefficients from grouped data was by the eminent psychologist E. L. Thorndike (1934). There appear to be no further comments on the topic from educational and psychological researchers except the papers questioning the appropriateness of estimating individual learning curves from group learning curves.

[5] It is again assumed for simplicity that groups of n observations each are formed such that $gn = N$. Otherwise, the group means $(\overline{X}_{i.}, \overline{Y}_{i.})$ need to be weighted by the group size (n_i) in the least-squares estimation of the parameters.

BEST COPY AVAILABLE

Footnotes--continued

[6] But the parameter of interest remains β_{YX}, the simple linear regression coefficient.

[7] This interpretation for Z is in no sense arbitrary. The process of grouping systematically has much in common with the notion of selection. In fact, Lutjohann (personal communication) has suggested that the grouping bias discussed here is essentially selection bias, the result of a manipulated sampling of the observations of X and Y because of their association with Z. Recent work by Goldberger (1972a, 1972b), on selection bias in evaluating treatment effects with non-random sampling, also hints at this connection.

[8] After the bulk of the analyses was completed, it was discovered that there were missing observations on the grouping characteristics CLIMP, COLEFF, and QCJOB. In addition certain modifications were made in the response categories of ANTHIDEG. In its original form, ANTHIDEG formed nine groups. In the results reported here, however, students responding "Other (9)" were dropped, and students anticipating any professional degree beyond the masters level (responses 5, 6, 7, and 8) were collapsed into a single group numbered "5." The sizes of the subsamples defined by the acceptable responses to CLIMP, COLEFF, QCJOB, and the modified ANTHIDEG were 2632, 2669, 2637 and 2646, respectively. An examination of the means, standard deviations, and intercorrelations of SRAA, ACH, and SAT for these subsamples did not indicate any consistent and important deviations from the estimates based on the entire 2676 observations.

BEST COPY AVAILABLE

REFERENCES

1. Afifi, A. A. and Elashoff, R. M. "Missing observations in multivariate statistics I: review of the literature." Journal of the American Statistical Association, 1966, 61, 595-604.

2. Afifi, A. A. and Elashoff, R. M. "Missing observations in multivariate statistics II: point estimation in simple linear regression." Journal of the American Statistical Association, 1967, 62, 10-29.

3. Alker, Hayward R., Jr. "A typology of ecological fallacies." Pp. 69-86 in M. Dogan and S. Rokkan (eds.), Quantitative Ecological Analysis in the Social Sciences. 1969, Cambridge, Mass.: MIT Press.

4. Baird, L. and Feister, W. J. "Grading Standards: The relation of changes in average student ability to the average grades awarded." American Educational Research Journal, 1972, 9 (3), 431-441.

5. Bartlett, M. S. "Fitting a straight line when both variables are subject to error." Biometrics, 1949, 5, 207-212.

6. Blalock, H. M. Causal Inferences in Nonexperimental Research. Chapel Hill, N. C.: University of North Carolina Press, 1964.

7. Blalock, H. M. "Aggregation and measurement error." Social Forces, 1972.

8. Blalock, H. M.; Wells, Caryll S.; and Carter, L. F. "Statistical estimation with random measurement error." Pp. 75-103 in E. Borgotta and G. Bohrenstedt (eds.), Sociological Methodology, 1970. San Francisco : Jossey-Bass.

9. Boruch, R. F. "Assuring confidentiality of responses in social research." American Sociologist, 1971, 6 (4), 308-311(a).

10. Boruch, R. F. "Educational research and the confidentiality of data: A Case Study." Sociology of Education, 1971, 44, pp. 59-85(b).

11. Boruch, R. F. "Maintaining confidentiality of data in educational research; A systematic analysis." American Psychologist ,1971, 26, pp. 413-430(c).

12. Boruch, R. F. "Strategies for eliciting and merging confidential social research data.'' Policy Sciences, 1972 (in press).

13. Boruch, R. F. "Social research and the confidentiality issue: methodological perspectives." Paper presented at the meetings of the American Educational Research Association, Chicago ,1972.

BEST COPY AVAILABLE

References--continued

14. Burstein, Leigh. The Use of Data from Groups for Inferences About Individuals in Educational Research. Unpublished Ph.D. dissertation, Stanford University, 1974.

15. Cartwright, D. S. "Ecological variables." In E. F. Borgatta (ed) Sociological Methodology 1969, 1969, San Francisco: Jossey-Bass, pp. 155-218.

16. Chai, J. J. A Study of Effects of Correlated Measurement Errors and Grouping on the Ordinary Least-Squares Estimator for the Regression Coefficient and Ordinary Estimator of the Product-Moment Correlation Coefficient of a Finite Bivariate Population. Unpublished Ph.D. Dissertation, University of Minnesota, 1968.

17. Chai,J.J. "Errors of measurement and 'ecological' correlation." Proceedings of the Social Statistics Section of the American Statistical Association, 1971, 272-278.

18. Cramer, J. S. "Efficient grouping: regression and correlation in Engel curve analysis." Journal of the American Statistical Association, 1964, 59(March), 233-250.

19. Dogan, M. and Rokkan, S. (eds.), Quantitative Ecological Analysis in the Social Sciences , 1969, Cambridge, Mass.: MIT Press.

20. Duncan, O. D. and Davis, B. "An alternative to ecological correlation." American Sociological Review, 1953, 18 (December), 665-666.

21. Duncan, O. D.; Cuzzort, R. P. and Duncan, B. D. Statistical Geography, 1961, Glencoe, Illinois: Free Press.

22. Durbin, J. "Errors in variables." Revue de l'Institut International de Statistique, 1954, 22, 23-32.

23. Elashoff, J. S. and Elashoff, R. M. "Missing data problems for two samples on a dichotomous variable." Stanford Center for Research and Development Memorandum No. 73, 1971.

24. Elashoff, R. M. and Elashoff, J. S. "Regression analysis with missing data." In R. L. Bisco (ed). Data Bases, Computers and the Social Sciences, 1970, New York: John Wiley & Sons, 198-207.

25. Feige, Edgar L. and Watts, Harold W. "Protection of privacy through microaggregation." In R. L. Bisco (ed.), Data Bases, Computers and the Social Sciences. 1970, New York: Wiley & Sons, 261-272.

BEST COPY AVAILABLE

References--continued

26. Feige, Edgar L. and Watts, Harold W. "An investigation of the consequences of partial aggregation of micro-economic data." Econometrica, 1972, 40 (March): 343-360.

27. Gehkle, C. and Biehel, R. "Certain effects of grouping upon the size of the co rrelation coefficient in census tract material." Journal of the American Statistical Association Supplement, 1934, 29: 169-170.

28. Goldberg, L. R. "Man versus mean: the exploitation of group profiles for the construction of diagnostic classification systems." Journal of Abnormal Psychology, 1972, 79 (2), 121-131.

29. Goldberger, A. S. "Selection bias in evaluating treatment effects: some formal illustrations." Discussion paper, Institute for Research on Poverty, University of Wisconsin--Madison, 1972 (a),#123-72.

30. Goldberger, A. S. "Selection bias in evaluating treatment effects: the case of interaction." Discussion paper, Institute for Research on Poverty, University of Wisconsin--Madison, 1972 (b), #129-72.

31. Goodman, Leo. "Ecological regression and the behavior of individuals." American Journal of Sociology, 1953, 64 (May), 610-625.

32. Goodman, Leo. "Some alternatives to ecological correlation." American Journal of Sociology 64 (May): 610-625, 1959.

33. Green, H. A. John. Aggregation in Economic Analysis. Princeton: Princeton University Press, 1964.

34. Griliches, Zvi. "Notes on the role of education in production functions and growth accounting." In W. L. Hansen (ed.) Education, Income and Human Capital, Studies in Income and Wealth No. 35, National Bureau of Economic Research, 1971, New York: Columbia University Press, 83-115.

35. Grunfield, Yehuda and Griliches, Zvi. "Is aggregation necessarily bad?" Review of Economics and Statistics 42 (February): 1-13, 1960.

36. Haitovsky, Yoel. "Unbiased multiple regression coefficients estimated from one-way classification tables when the cross-classifications are unknown." Journal of the American Statistical Association 61 (September): 720-728, 1966.

37. Haitovsky, Yoel. "Regression analysis of grouped observations when the cross-classifications are unknown." The Review of Economic Studies, 1968, (1) No. 101, 77-89.

BEST COPY AVAILABLE

References--continued

38. Hannan, Michael T. *Problems of Aggregation and Disaggregation in in Sociological Research.* Methodology Working Paper #4. Institute for Research in the Social Sciences, University of North Carolina, Chapel Hill, N. C., 1970.

39. Hannan, Michael T. *Aggregation and Disaggregation in Sociology.* Lexington, Mass., and Washington, D. C.: Heath, 1971.

40. Hannan, Michael T. "Approaches to the Aggregation Problem." Technical Report #46. Laboratory for Social Research, Stanford University, Stanford, Calif., 1972.

41. Hannan, M. T. and Burstein, L. "Estimation from grouped observations." *American Sociological Review*, 1974 (in press).

42. Hansen, M. H.; Hurwitz, W. N. and Madow, W. G. *Sample Survey Methods and Theory*, Volume II, 1953, New York: John Wiley & Sons, 65-66.

43. Hauser, R. M. "Schools and the stratification process." *American Journal of Sociology*, 1969, 74 (May), 587-611.

44. Ijiri, Y. "Fundamental queries in aggregation theory." *Journal of the American Statistical Association* 66 (December): 766-782, 1971.

45. Iversen, G. R. "Recovering individual data in the presence of group and individual effects." *American Journal of Sociology*, 1973, 79 (2), 420-434.

46. Johnston, J. *Econometric Methods.* 2nd Ed. New York: McGraw-Hill, 1972.

47. Kline, Gerald F.; Kent, K., and Davis, D. "Problems in the causal analysis of aggregate data with applications to political instability." In J. Gillespie and B. Nesvold (eds.), *Macro-Quantitative Analysis.* Beverly Hills: Sage Publications, 1971.

48. Lewy, Arieh. "Correlation Coefficients of three data types: zero order measures, within group deviation scores and group means." Unpublished manuscript, Working Paper No. 16, 1972.

49. Lütjohann, Harry. *Linear Aggregation in Regression and Econometrics.* Research Report, Institute of Statistics, University of Stockholm, 1971.

50. Madansky, Albert. "The fitting of straight lines when both variables are subject to error." *American Statistical Association Journal*, 1959, 54 (March), 173-205.

References--continued

BEST COPY AVAILABLE

51. Nataf, Andre. "Aggregation,"in S. D. Sills (ed.) International Encyclopaedia of the Social Sciences, 1968, 162-168.

52. Prais, S. J. and Aitchison, J. "The Grouping of observations in regression analysis." Review of the International Statistical Institute, 1954, 22: 1-22.

53. Robinson, A. H. "The necessity of weighting values in correlation analysis of areal data." Annals of the Association of American Geographers, 1956, 46 (March): 233-236.

54. Robinson, William S. "Ecological correlations and the behavior of individuals." American Sociological Review, 1950, 15 (June): 351-357.

55. Rock, D. A.; Centra, J. A., and Linn, R. L. "Relationships between college characteristics and student achievement," American Educational Research Journal, 1970, 7 (1), 109-121.

56. Selvin, H. E. "Durkheim's 'Suicide' and problems of empirical research." American Journal of Sociology, 1958, 63 (May), 607-619.

57. Shively, W. P. " 'Ecological' inference: the use of aggregate data to study individuals." American Political Science Review, 1969, 63 (December), 1183-1196.

58. Slatin, G. T. "Ecological analysis of delinquency." American Sociological Review, 1969, 34 (December), 894-906.

59. Theil, Henri. Linear Aggregation in Economic Relations. 1954, Amsterdam: Holland Publishing Company.

60. Theil, Henri. Principles of Econometrics. 1971, New York: John Wiley

61. Thorndike, E. L. "On the fallacy of imputing the correlations found for groups to the individuals or smaller groups composing them." American Journal of Psychology, 1939, 52, 122-124.

62. Tukey, J. W. "Components in regression." Biometrics, 1951, 7, 33-70.

63. Wald, A., "Fitting of straight lines if both variables are subject to error." Annals of Mathematical Statistics, 1940, 11, 284-300.

64. Wiley, D. W. "Design and analysis of evaluation studies." In M. C. Wittrock and D. E. Wiley (ed.) The Evaluation of Instruction: Issues and Problems, 1970, New York: Holt, Rhinehart and Winston, 259-268.

BEST COPY AVAILABLE

References--continued

65. Wiley, D. E. and Wiley, J. A. "The estimation of measurement error in panel data." American Sociological Review, 1970, 35, 112-117.

66. Yule, Udny G. and Kendall, Maurice G. An Introduction to the Theory of Statistics. 1950, London: Charles Griffin.

ERIC Full Text Provided by ERIC