DOCUMENT RESUME

ED 126 136 TM 005 392

AUTHOR Poynor, Hugh
TITLE Spurious Aggregation and the Units of Analysis.
PUB DATE [Apr 76]
NOTE 12p.; Paper presented at the Annual Meeting of the American Educational Research Association (60th, San Francisco, California, April 19-23, 1976)

EDRS PRICE MF-$0.83 HC-$1.67 Plus Postage.
DESCRIPTORS Analysis of Variance; Correlation; *Data Analysis; Educational Research; *Evaluation Methods; Hypothesis Testing; Multiple Regression Analysis; *Research Design; *Research Problems; Suppressor Variables
IDENTIFIERS *Grouping (Statistics)

ABSTRACT

The degree to which the chosen units of analysis are likely to produce spurious findings in staged combinations of multiple linear regression procedures are examined. The effects of grouping variables (e.g., classroom, school, and school district) on procedures such as Coleman's semipartial regression and Mayeske's commonalities, in light of hypothetical sample homogeneity and heterogeneity are also illustrated. The utility of including grouping variables in the regression models and alternatives such as collecting within-groups regression results in lieu of between-group methods are also considered. (Author/DEP)

ED126136

TM005 392

Spurious Aggregation and the Units of Analysis*

Hugh Poynor

Kirschner Associates, Inc.

Washington, D.C.

In educational research studies the principal investigator takes steps to insure the validity of reported findings by designing the study properly and analyzing the data thoroughly. It is only when failure of design and analysis efforts occurs that spuriour findings are reported. Beyond the reporting stage may lie the scrutiny of collegues, an audience of policy makers with their own urgent perspectives, a need for replication, and a hope of further understanding of reported effects.

Our symposium today is concerned with the research implications of data aggregation and unit of analysis issues. These topics many times are obscured by overly detailed statistical and mathematical methodologies. Yet it is meaningful to characterize their intention with the familiar story of the man seeking to locate his house keys two blocks from where they were lost because of better light where he was looking. In this same vein let me say that those with information needs concerning state-level policy will not find appropriate answers using the pupil as the analysis unit even though that is many times the most conventional place to look. Similarly, one cannot expect to shed light on the effects

* Paper presented at annual AERA convention held in San Francisco, April, 1976. Author's current address is: Suite 1137, 733 15th Street, N.W., Washington, D.C. 20005

15.13

of a particular instructional technique on the pupil by analyzing district-level aggregates. The choice of a unit of analysis in all cases must be directed by the research goals, or by the audience for the findings and their special needs, and by methodological considerations such as independence of the units.

It will now be noted how aggregation and unit of analysis phenomena have potential for creating spurious analysis outcomes. By way of proceeding to these problems, it is necessary to consider several areas that are often thought to be unrelated, and then to re-introduce analysis techniques from a new point of view.

Subsample Boundary (the G variable)

Drawing subsamples of the units of analysis in a data base may be done in several ways and for several reasons. Subsamples are often separated according to identifying features such as school grade level, instructional treatment and so on. Once such samples are identified it is a simple task to devise a mapping variable called G that will serve to define the boundaries of the subsample. This G variable is nothing more than a rule for one of two actions: (1) forming aggregates (or averages, as is usually the case), and (2) identifying interactions among elements of X and Y. Definition of interactions through the use of G is accomplished by simply using the rules to expand the analysis model, or, in general linear model terminology, by expanding the number of predictor variables. The present paper shows several effects of sample boundary variables in routine analysis efforts in light of

hypothetical parent-sample homogeneity or heterogenity.

Traditional Suppressor Phenomena

The action of independent variables adding to prediction of a dependent variable through apparent relationships with other independent variables but not through direct relationships with the dependent variable has been termed the suppressor effect. Two conditions must be met before the independent variable is thought to be operating as a classical or traditional (Cohen & Cohen, 1975; Conger, 1974) suppressor in multiple regression settings: (1) the X variable must be only slightly or not at all related to the dependent variable (Y) and (2) this same X variable must be strongly related to at least one other X variable. Under these conditions multiple prediction will be increased by a factor that usually exceeds the bivariate relation between Y and the X (suppressor) variable. For this reason, suppressors are welcome additions to multiple regression models. Beyond this highly specific pattern of bivariate correlations, a traditional suppressor is identified by its negative regression weight. Thus, the regression weight which a suppressor acquires is positive when it is negatively correlated with the dependent variable. In other language, the raw weight and the standardized (or beta) weight have opposite signs.

Suppressor phenomena occur only when an inconsistency exists between the X variable set and the Y variable. Inconsistency arises from the manner in which subsamples of Y are related differentially to X. The variance in Y may be visualized as being composed of a subset of Y scores bearing a

positive correlation with selected X_1 values and another subset of Y bearing no correlation with the remainder of X_1 values in the equation $Y = X_1 + X_2$. This latter subset of Y is highly predictible from the corresponding X_2 values even though on the whole Y and X_2 are not related. In this example, X_2 is the suppressor variable. When all Y values are used to compute r (Y, X_2), the correlation is near zero, but it is possible to find a smaller set of Y which is highly correlated with a corresponding X_2 set. In order for the suppressor effect to operate, then, X_1 and X_2 <u>must share these subsample boundaries.</u>

Guilford (1964) and others do not appeal to this type of explanation of suppressor effects, although a discussion of variance borders on the subsample boundary explanation proposed in the present paper. It is explained that X_1, in spite of a correlation with Y, has some variance that correlates near zero with Y. It is because of this X_1 variance that the correlation between Y and X_1 is prevented from being even larger. Now, variable X_1 correlates highly with X_2 (the suppressor) because they have in common that variance not shared by Y. Thus, including X_2 in the multiple regression model permits this portion of the Y variance to find a capable predictor. A boundary explanation is more generally useful than this traditional variance explanation of suppressors since it leads to the understanding of X-Y relationships in a broader number of analysis situations. Usefulness of this explanation will next be demonstrated with commonality analysis interpretations.

Negative Commonalities

More recently than Guilford, Veldman (1974) and Kerlinger & Pedhazur (1973) have called attention to "negative contributions" in commonality analysis results. The commonality procedure examines each X variable as if it is the last variable in the dependent variable set to be added to the regression equation. The net value of X in the equation is the percent increase in explained Y variance as computed by subtracting the R^2 figures of the before-and-after X models. The commonality name for the procedure arises because pairs of variables, triplets, and so on are similarly treated as the last additions to the model so that their common net value may be determined. Since the output of commonality analysis consists of percents of Y variance explained by both single and joint contributions X's make toward explaing Y variance, then the negative contribution outcome is a signal that less than nothing in Y has been explained. Negative contributions are possible only for commonality values, that is for pairs of X variables while they are impossible for unique (single X) contributions.

Besides being a sad state of affairs after research funds have been expended for the study, explanation of less than nothing can also be a clear signal that suppressors are operating in the data set. Extending the present interpretation of suppressors to negative contributions, then, suggests that they are products of the action of a boundary variable. Just as there are inconsistencies in X-Y relationships which will produce suppressor phenomena, commonality analysis will also register the inconsistencies. By

inconsistency is meant the irregular quality of X-Y correlations when sub-samples of X and Y are considered. In other words the sample is not homogeneous with regard to the manner in which X and Y relate statistically. For example one portion of the sample may be high and positively related on X and Y, while another may be negatively related.

Analysis of Variance

The problem of sample heterogenity can also be considered in the context of analysis of variance situations, where effects are sought as indicators of experimental treatment outcomes. Traditional use of the term "relationships" is made for observational-correlational research studies while use of the term "effects" is usually reserved for experimental studies. In either case the G variable is a useful heuristic for the topic of spurious aggregation and the units of analysis.

Attention to parent sample heterogenity in experimental school effect studies comes from the popular textbook by Glass & Stanley (1970). Their verbal analysis of independence and the sampling units of analysis led them to conclude that degrees of freedom must suffer if the research setting is the intact classroom. A setting such as this is quite frequent in our research, and unfortunately it is many times mistreated in the statistical analysis work. Glass and Stanley (1970) recommend that one form classroom means and adjust degrees of freedom accordingly. That is, the researcher should aggregate the pupil scores, form classroom means, and adjust df to reflect the number of classrooms rather than the number of pupils.

In terms of the G variable approach to ~~this analysis~~ of varience situation, it may be clear already that the G variable is the classroom identification of each pupil. If Glass & Stanley are empirically correct, we must expect there to be considerable subsample heterogenity with respect to G, for if the sample if found to be homogeneous with respect to G then aggregation and loss of df is a questionable procedure.

Because aggregation may be questionable and because it is often done--sometimes automatically--when analyzing both effects and relationships the recommendation is best subjected routinely to empirical test. Poynor (1974) has provided evidence that the untested use of aggregate units of analysis of classroom (and sometimes, individual units of analysis) can lead to grievous Type I and II errors. Either unit of analysis can lead to errors, so exclusive use of either unit accomplishes nothing. A pre-analysis step is required to identify the proper unit of analysis. In an investigation of the independence of analysis units, Glendening (1976) discusses the use of this pre-analysis test, but finds it to be too conservative or too liberal a test in selected situations of independent and dependent units. These two studies show the empirical test of a unit of analysis with respect to sample homogeneity, and the importance of management control in effect studies for protection of the selected units. Because of the great loss of information that occurs with aggregation (Poynor, 1975) it should not be done until after it is proven necessary.

Correlation and Regression

Thorndike (1939) called attention to the operation of G in correlational-relationship studies by illustrating the spurious development of a relationship

across higher order aggregations. His early example calls attention to real-world manifestations of G. Twelve school districts were each asked to provide two data points on each pupil in the district: the pupil's IQ score (X) and the number of rooms available in the school building where each pupil was taught (Y). Each individual district was quite heterogeneous with respect to these variables. That is, within each district there was a broad range of IQs and rooms yet there was no correlation between these X and Y variables within each district. The problem in the example comes from heterogenity on these X-Y variables across districts. When the 12 districts were combined and a correlation coefficient was computed, it was .45, not .00 as it has been in each district individually. No aggregation had yet taken place. Once school district aggregates (means) were used, the resulitng correlation was .90.

Extensive research work with simulations and literature reviews by the team of Hannan and Burstein have been done in the correlation and regression analysis areas (Hannan, 1971; Burstein, 1975). These studies provided the impetus for the present synthesis of the effects and relationships areas using the G variable concept. This author has sought to apply their detailed statistical research findings to popular analysis models, using nontechnical language. Where the above authors refer to bias and inefficiency (or inaccuracy and inconsistency) of regression weights, these terms have been collected here under the label "spurious." Using their findings to predict spurious correlations, it is necessary to employ the G variable again. Briefly, they conculde that spurious relationships will be products

of situations where strong relationships exist among G-X or G-Y. Recall that both these situations were present in the Thorndike paper which is somewhat classic as an example of confounded or proxy relationships.

Measurement of G

Before ending this paper, let me say that it is very little comfort to know that suppressors, negative commonalities, Type I and Type II errors and inflated correlation coefficients can be explained in terms of a G variable used to establish the unit of analysis. Research which is carefully planned and conducted will rarely be affected by this as a nuisance at the time of data analysis. Still, our understnading of effects and relationships among measures of interest is many times insufficient to control all the potential contaminants of our findings.

An ultimate solution is believed to be the measurement of G itself, for if G is viewed as an abstract rule for selecting or forcing observations into groups prior to analysis, then G is truly a potent treatment variable. This solution does not refer to the practice of simply including dummy variables in a regression equation as indicators of school district, and school building location of the observation.

The importance of fully specified models, or true starting models as they are sometimes called, is well known to data analysts. While the practice of using such dummy variables often increases the percent of explained criterion variance, it does nothing more than acknowledge G as potent.

It is the measurement of the underlying classroom, school building or district differences that will promote our understanding of G and the effects and relationships associated with our criterion variables. Instead of making sterile statements such as "Twenty percent of the criterion variance was explained by school building differences," the researcher may someday be able to offer richer, more meaningful statements relating criterion variance to specific features of conditions, such as teacher warmth, type of disciplinary policy, student body cohesiveness, presence of open classrooms or other substantive learning variables.

Burstein, Leigh. Data Aggregation in Educational Reserarch. Paper presented at the Annual meeting of AERA, Washington, D.C., 1975.

Cohen, Jacob and Patricia Cohen. Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences. John Wiley, New York, 1975.

Conger, A.J. A revised definiton for suppressor variables: a guide to their identification and interpretation. Educ. Psychol. Meas., 1974, 34, 35-46.

Glass, G.R. and Stanley, J.C. Statistics in Education and Psychology. Prentice-Hall, Englewood Cliffs, N.J. 1970.

Glendening, Linda. The Effects of Correlated Units of Analysis: Choosing the Appropriate Unit. Paper presented at the Annual Meeting of AERA, San Francisco, 1976.

Guilford, J.P. Fundamental Statistics in Psychology and Education, McGraw-Hill, New York, 1964.

Hannan, Michael T.and Burstein, Leigh, Estimation from grouped observations. Am. Soc. Rev., 1974, 39, 374-392.

Hannan, Michael T. Problems of Aggregation. In Blalock, H.M. (ed.), Causal Models in the Social Science. Aldine Press, Chicago, 1971.

Hannan, Michael T., Young, A.A. and Finielsen. Specification Bias Analysis of the Effects of Grouping of Observations in Multiple Regression Models. Paper presented at the Annual Meeting of AERA, Washington, D.C., 1975.

Kerlinger, F.N. and Pedhazur, E.J. Multiple Regression in Behavioral Research, Holt, Rinehart and Winston, New York, 1973.

Poynor, Hugh. To Aggregate is to Aggrevate. Paper presented at the Annual Meeting of AERA, Washington, D.C., 1975.

Poynor, Hugh. Selecting units of analysis. In Gary Borich (ed.) Evaluating Educational Programs and Products, Englewood Cliffs, N.J., Educational Technology Press, 1974.

Thurndike, E.L. On the fallacy of imputing the correlations found for groups to the individuals or smaller groups composing them. Am. J. Psychol., 52, 1939, 122-124.

Veldman, D.J. Negative Contributions in Commonality Analysis. Project PRIME working paper, The University of Texas at Austin, 1974.