Problems in the Analysis of Survey Data, and a Proposal

James N. Morgan, John A. Sonquist

Journal of the American Statistical Association, Volume 58, Issue 302 (Jun., 1963), 415-434.

Stable URL:
http://links.jstor.org/sici?sici=0162-1459%28196306%2958%3A302%3C415%3APITAOS%3E2.0.CO%3B2-6

PROBLEMS IN THE ANALYSIS OF SURVEY DATA, AND A PROPOSAL

JAMES N. MORGAN AND JOHN A. SONQUIST*
University of Michigan

Most of the problems of analyzing survey data have been reasonably well handled, except those revolving around the existence of interaction effects. Indeed, increased efficiency in handling multivariate analyses even with non-numerical variables, has been achieved largely by assuming additivity. An approach to survey data is proposed which imposes no restrictions on interaction effects, focuses on importance in reducing predictive error, operates sequentially, and is independent of the extent of linearity in the classifications or the order in which the explanatory factors are introduced.

A. NATURE OF THE DATA AND THE WORLD FROM WHICH THEY COME

THE increasing availability of rich data from cross section surveys calls for more efficient methods of data scanning and data reduction in the process of analysis. The purpose of this paper is to spell out some of the problems arising from the nature of the data and the nature of the theories which are being tested with the data, to show that present methods of dealing with these problems are often inadequate, and to propose a radical new method for analyzing survey data. There are seven things about the data or about the world from which they come which need to be kept in mind.

First, there is a wide variety of information about each person interviewed in a survey. This is good, because human behavior is motivated by more than one thing. But the very richness of the data creates some problems of how to handle them.

Second, we are dealing not with variables for the most part, but with classifications. These vary all the way from age, which can be thought of as a variable put into classes, to occupation or the answers to attitudinal questions, which may not even have a rank order in any meaningful sense. Even when measures seem to be continuous variables, such as age or income, there is good reason to believe that their effects are not linear. For instance, people earn their highest incomes in the middle age ranges. Expenditures do not change uniformly with changes in income at either extreme of the income scale.

Third, there are errors in all the measures, not just in the dependent variable, and there is little evidence as to the size of these errors, or as to the extent to which they are random.

Fourth, the data come from a sample and generally a complex one at that. Hence, there is sample variability piled on top of measurement error. The fact that almost all survey samples are clustered and stratified leads to problems of the proper application of statistical techniques. Statistical tests usually assume simple random samples rather than probability samples. More ap-

* The authors are indebted to many individuals for advice and improvements. In particular, Professor L. J. Savage noticed that some interactions would remain hidden, and Professor William Ericson proved that locating the best combination of subclasses of a single code was simple enough to incorporate into the program. A Ford Foundation grant to the Department of Economics of the University of Michigan supported the author's work on some substantive problems which led to the present focus on methods. Support from the Rockefeller Foundation is also gratefully acknowledged.

propriate tests have been developed for simple statistics such as proportions, means, and a few others.

Fifth, and extremely important, there are intercorrelations between many of the explanatory factors to be used in the analysis—high income goes along with middle age, with advanced education, with being white, with not being a farmer, and so forth. This makes it difficult to assess the relative importance of different factors, since their intercorrelations get in the way. Since many of them are classifications rather than continuous variables, it is not even easy to measure the extent of the intercorrelation. Measures of association for cross classification raise notoriously difficult problems which have not really been solved in any satisfactory way.[1]

Sixth, there is the problem of interaction effects. Particularly in the social sciences, there are two powerful reasons for believing that it is a mistake to assume that the various influences are additive. In the first place, there are already many instances known of powerful interaction effects—advanced education helps a man more than it does a woman when it comes to making money; and it does a white man more good than a Negro. The effect of a decline in income on spending depends on whether the family has any liquid assets which it can use up. Women have their hospitalizations at different ages than men. Second, the measured classifications are only proxy variables for other things and are frequently proxies for more than one construct. Several of the measured factors may jointly represent a theoretical construct. We may have interaction effects not because the world is full of interactions, but because our variables have to interact to produce the theoretical constructs that really matter. The idea of a family life cycle, unless arbitrarily created out of its components in advance, is a set of interactions between age, marital status, presence, and age of children.[2] It is therefore often misleading to look at the over-all gross effects of age or level of education. Where interaction effects exist, the concept of a main effect is meaningless, and it is our belief that in human behavior there are so many interaction effects that we must change our approach to the problems of analysis.

Another example of interaction effects appeared in the attempt to build equivalent adult scales to represent the differences in living expenses of families of different types. After many years of analysis, one of the most recent studies in this field has concluded "when its size changes, families' tastes appear to change in more complicated ways than visualized by our hypothesis."[3] More

[1] One seemingly appropriate measure for two classifications both being used to predict the same variable is one called lambda suggested by Goodman and Kruskal. With many kinds of survey data this measure, which assumes that an absolute prediction has to be made for each individual, is too insensitive to deal with situations where each class on the predicting characteristic has the same modal class on the other characteristic that is to be predicted. An effective and properly stochastic measure would be derived by assigning a one-zero dummy variable to belonging to each class of each of the two characteristics and then computing the canonical correlation between the two sets of dummy variables.

See Leo A. Goodman and William H. Kruskal, "Measures of association for cross classifications," *Journal of the American Statistical Association*, 49 (December, 1954), 732–64.

[2] John B. Lansing and James N. Morgan, "Consumer finances over the life cycle," in *Consumer Behavior*, Volume II, L. Clark (Editor) (New York: New York University Press, 1955).

See also Leslie Kish and John B. Lansing, "Family life cycle as an independent variable," *American Sociological Review*, XXII (October, 1957), 512–9.

[3] In other words family composition had different effects on different expenditures. F. G. Forsythe, "The relationship between family size and family expenditure," *Journal of the Royal Statistical Society*, Series A, vol. 123 (1961), 367–97, quote from p. 386.

recently in analyzing factors affecting spending unit income, it has become obvious that age and education cannot operate additively with race, retired status, and whether the individual is a farmer. The attached table illustrates this with actual average incomes for a set of nonsymmetrical groups. The twenty-one groups account for two-thirds of the variance of individual spending unit incomes, whereas assuming additivity for race and labor force status even with joint age-education variables produces a regression which with 30 variables accounts for only 36 per cent of the variance. A second column in the

TABLE 1. SPENDING UNIT INCOME AND THE NUMBER IN THE UNIT WITHIN VARIOUS SUBGROUPS

Group	Spending unit average (1958) income	Number in unit	Number of cases
Nonwhite, did not finish high school	$ 2489	3.3	191
Nonwhite, did finish high school	5005	3.4	67
White, retired, did not finish high school	2217	1.7	272
White, retired, did finish high school	4520	1.7	72
White, nonretired farmers, did not finish high school	3950	3.6	87
White nonretired farmers, did finish high school	6750	3.6	24
The Remainder			
0–8 grades of school			
18–34 years old	4150	3.8	72
35–54 years old	4670	3.8	240
55 and older—not retired	4846	2.2	208
9–11 grades of school			
18–34 years old	5032	3.7	112
35–54 years old	6223	3.4	202
55 and older—not retired	4720	2.1	63
12 grades of school			
18–34 years old	5458	3.3	193
35–54 years old	7765	3.8	291
55 and older—not retired	6850	2.0	46
Some college			
18–34 years old	5378	3.0	102
35–54 years old	7930	3.8	112
55 and older—not retired	8530	2.0	36
College graduates			
18–34 years old	7520	3.8	80
35–54 years old	8866	2.9	150
55 and older—not retired	10879	1.8	34

Source: 1959 Survey of Consumer Finances.

table gives the average number of people in the unit, and it can be seen that this particular breakdown is not particularly useful for analyzing the number of people in a unit. On the other hand, if each group were to be used to analyze expenditure behavior, income, and family size are likely to operate jointly rather than additively.

In view of the fact that intercorrelation among the predictors on the one hand and interaction effects on the other are frequently confused, it seems useful to give a pictorial example indicating both the differences between them and the way in which they operate when both are present. Our concern is not with statistical tests to distinguish between them, but with the effects of ignoring their presence.

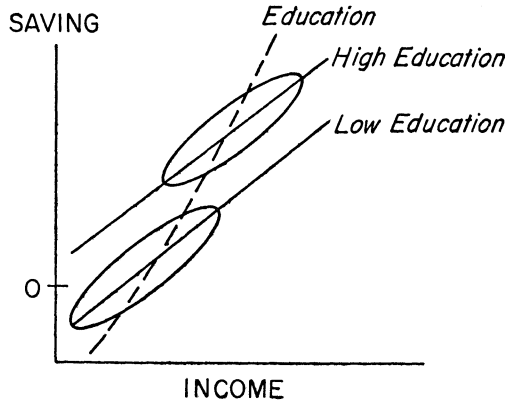
Chart I shows pictorially three cases, real but exaggerated. First, there is a case where the two explanatory factors, income and education, are correlated with one another, but do not interact. Second, a case where income and being self-employed interact with one another but are not correlated, and third, a situation where income and asset holdings are correlated with one another and also interact in their effect on saving. The ellipsoids represent the area where most of the dots on a scatter diagram would appear. In the first case, it is clear that a simple relation between income and saving would exaggerate the effect of income on saving by failing to allow for the fact that high income people have more education, and that highly educational people also save more. An ordinary multiple regression, however, using a dummy variable representing high education would adequately handle this difficulty. In the second case there is no particular correlation, we assume, between income and being self-employed, but the self-employed have a much higher marginal propensity to save than other people. Here, the simple relationship between income and saving becomes a weighted compromise between the two different effects that really exist. A multiple correlation would show no effect of being self-employed and the same compromise effect of income. Only a separate analysis for the self-employed and the others would reveal the real state of the world. In the third case, not only do the high-asset people have a higher marginal propensity to save, but they also tend to have a higher income. Multiple correlation clearly will not take care of this situation in any adequate way. It *will* produce an "income effect" which can be added to an "asset effect" to produce an estimate of saving. Here the income effect is an average of two different income effects. The estimated asset effect is likely to come out closer to zero than if income had been ignored. Of course, where interactions exist, there is little use in attempting to measure separate effects.

Finally, there are logical priorities and chains of causation in the real world. Some of the predicting characteristics are logically prior to others in the sense that they can cause them but cannot be affected by them. For instance, where a man grows up may affect how much education he gets, but his education cannot change where he grew up. We are not discussing here the quite different analysis problem where the purpose is not to explain one dependent variable but to untangle the essential connections in a network of relations.

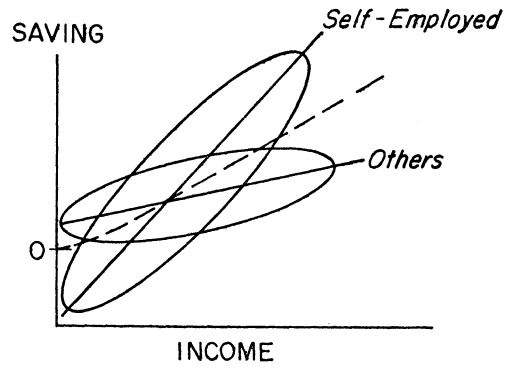
In dealing with a single dependent variable representing some human behavior, we might end up with at least three stages in the causal process—early

childhood and parental factors, actions and events during the lifetime, and current situational and attitudinal variables. If this were the end of the problem we could simply run three separate analyses. The first would analyze the effects of early childhood and parental factors. The second would take the residuals from this analysis and analyze them against events during a man's lifetime up until the present, and the third would take the residuals from the

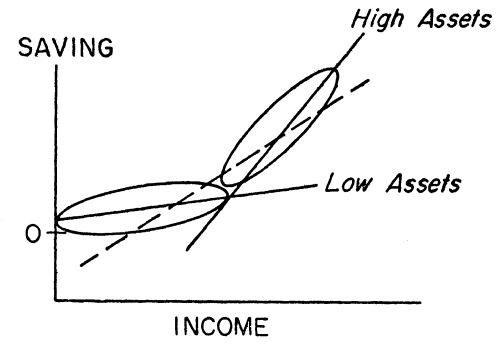
Muticollinearity, i.e., correlation between income and education but no interaction



Interaction, but no multicollinearity (no correlation between income and self-employment)



Both



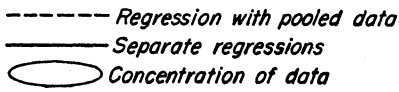


CHART I. Combinations of Multicollinearity and Interaction and Their Effects.

second analysis and analyze them against current situational and attitudinal variables. But the real world is not even that simple, because some of the same variables which are logically prior in their direct effects may also tend to mediate the effect of later variables. For instance, a man's race has a kind of logical priority to it, but at the same time it may affect the way other things such as the level of his education operate to determine his income.

This is an impressive array of problems. Before we turn to a discussion of current attempts to solve these problems and to our own suggestions, it is essential to ask first what kind of theoretical structure is being applied and what the purposes of analysis are.

B. NATURE OF THE THEORY AND PURPOSES OF ANALYSIS

Perhaps the most important thing to keep in mind about survey data in the social sciences is that the theoretical constructs in most theory are not identical with the factors we can measure in the survey. The simple economic idea of ability to pay for any particular commodity is certainly a function not only of income but of family size, other resources, expected future income, economic security, and even extended family obligations. A man's expectations about his own economic future, which we may theorize will affect his current behavior, might be measured by a battery of attitudinal and expectational questions or by looking at his education, occupation, age, and the experience of others in the same occupation and education group who are already older. The fact that the theoretical constructs in which we are interested are not the same as the factors we can measure, nor even simply related to them, should affect our analysis techniques and focus attention on creating or locating important interaction effects to represent these constructs.

Second, there are numerous hypotheses among which a selection is to be made. Even if the researcher preferred to restrict himself to a single hypothesis and test it, the intercorrelations among the various explanatory factors mean that the same result might support any one of several hypotheses.[4] Hence, comparisons of relative importance of predictors, and selecting those which reduce predictive errors most, are required.

When we remember that there are also variable errors of measurement, the problem of selecting between alternative hypotheses becomes doubly difficult, and ultimately requires the use of discretion on the part of the researcher. Better measurement of a factor might increase its revealed importance.

Finally, researchers may have different reasons why they wish to predict individual behavior. Most will want to predict behavior of individuals in the population, not just in the sample, which makes the statistical problem somewhat more complicated. But some may also want to focus on the behavior of some crucial individuals by assigning more weight to the behavior of some rather than others. Others may want to test some explanatory factors, however small their apparent effect, because they are important. They may be important because they are subject to public policy influences or because they

[4] For an excellent statement of the application of this problem to the economists' concern with the permanent income hypothesis versus the relative income hypothesis, see Jean Crockett, "Liquid assets and the theory of consumption" (New York: National Bureau of Economic Research, 1962) (mimeographed).

are likely to change over time, or because they are crucial to some larger theoretical edifice. The nature of these research purposes thus combines with the nature of the data and their characteristics to make up the problem of how to analyze the data.

C. THE STRATEGY CHOICE IN ANALYSIS

One can think of a series of strategies ranging from taking account of only the main effects of each explanatory classification separately or jointly, to trying to take account of all possible combinations of all the classifications at once. Even if there were enough data to allow the last, however, it would not be of much use. The essence of research strategy then consists of putting some restrictions on the process in order to make it manageable. One possibility is to cut the number of explanatory factors utilized, and another is to restrict the freedom with which we allow them to operate.[5] One might assume away most or all interaction effects, for instance, and keep a very large number of explanatory classifications. Still further reduction in the number of variables is possible, if one assumes linearity for measured variables or, what amounts to the same thing builds arbitrary scales, incestuously derived out of the same data in order to convert each classification into a numerical variable. Clearly, the more theoretical or statistical assumptions one is willing to impose on the data, the more he can reduce the complexity of the analysis. A difficulty is that restrictions imposed in advance cannot be tested. There seems some reason to argue that it would be better to use an approach which developed its restrictions as it went along. In any case keeping these problems in mind we turn now to a summary of how analysis problems in using survey data are currently being handled and some of the difficulties that present methods still leave unsolved.

D. HOW PROBLEMS IN ANALYSIS ARE CURRENTLY BEING HANDLED—AN APPRAISAL

We take the seven problems in section A in the same order in which they are presented there plus the major problem in section B, that of theoretical constructs not measured directly by the factors on which we have data. The first problem was the existence of many factors. The simplest procedure has been to look at them one at a time always keeping in mind the extent to which one factors is intercorrelated with others. Another technique, particularly with attitudes, has been to build indexes or combinations of factors either arbitrarily or with the use of some sort of factor analysis technique.[6] The difficulty is that the first of these is quite arbitrary, and the second is arbitrary in a different sense, in that most mechanical methods of combining factors are based on the intercorrelations between the factors themselves and not in the way in which they may affect the dependent variable. It is quite possible for two highly correlated factors to influence the dependent variable in opposite ways. Building a combination of the two only on the basis of their intercorrelation would create a factor which would have no correlation at all with the dependent

[5] For a discussion of alternative strategies made while commenting on a series of papers, see James Morgan, "Comments," in *Consumption and Saving*, Volume I, I. Friend and R. Jones (Editors.) (Philadelphia: University of Pennsylvania Press, 1960), pp. 276–84.

[6] Charles Westoff and others, *Family Planning in Metropolitan America* (Princeton: Princeton University Press, 1961).

variable. With highly correlated attitudes, however, some such reduction to a few factors may be required and meaningful.

With the advent of better computing machinery, the problem of multiple factors has frequently been handled by using multiple correlation techniques. The use of these techniques, of course, required solving the second problem, that arising from the fact that in many cases we have classifications rather than continuous variables. This has been done in two ways, first, by building arbitrary scales. For instance, one could assign the numbers one, two, three, four, five, and six to the six age groups in order. Or if age were being used to predict income, one could assign a set of numbers representing the average income of people in those age groups.[7] But unless machine capacity is extremely limited, a far more flexible method which is coming into favor is to use what have been called dummy variables.[8] The essence of this technique is to assign a dummy variable to each class of a characteristic except one. It is called a dummy variable because it takes the value one if the individual belongs in that subclass or a zero if he does not. If ordinary regression procedures are to be used, of course, dummy variables cannot be assigned to every subclass of any characteristic, since this would overdetermine the system. However, at the Survey Research Center we have developed an iterative program for the IBM 7090, the output of which consists of coefficients for each subclass of each characteristic, the set for each characteristic having a weighted mean of zero. This means that the predicting equation has the over-all mean as its constant term, and an additive adjustment for each characteristic, depending on the subclass into which the individual falls on that characteristic. This is the standard analysis of variance formulation when all interactions are assumed to be zero. Of course, the coefficients of dummy variables using a regular matrix inversion routine can easily be converted into sets of this sort. There remain two difficulties with this technique. One is the problem of interaction effects, which are either assumed away or have to be built in at the beginning in the creation of the classes. A second arises from the nature of the classifications frequently used in survey data. Even though association between, say, occupation and the incidence of unemployment faced by an individual is not terribly high, the occupation code generally includes one or two categories such as the farmers and the retired who, by definition, cannot be unemployed at all. When dummy variables are assigned to these classes, it may easily occur that there is a perfect association between a dummy variable representing one of these peculiar (not applicable) groups in one code and a dummy variable representing something else in another classification (not unemployed). If the researcher omits one of each such pair of dummy variables in a regression routine, he is all right.

A third problem, that of errors in the data, is generally handled by not re-

---

[7] For an example see Jerry Miner, "Consumer Personal Debt—An Intertemporal Analysis," in *Consumption and Saving*, Volume II, I. Friend and R. Jones (Editors) (Philadelphia: University of Pennsylvania Press, 1960), 400–61.

[8] Daniel Suits, "The Use of Dummy Variables in Regression Equations," *Journal of the American Statistical Association*, 52 (December, 1957), 548–51.

T. P. Hill, "An Analysis of the Distribution of Wages and Salaries in Great Britain," *Econometrica*, 27 (July, 1959), 355–81.

jecting hypotheses too easily and by attempting to use some judgment in the assessment of relative importance of different factors or different hypotheses keeping in mind the accuracy with which the variables have probably been measured.

The fact that the data come from a sample has frequently been ignored. As the analysis techniques become more complicated, it becomes almost impossible to keep the structure of the sample in mind too. However, there is some reason to believe that the clustering and stratification of the sample become less and less important the more complex and more multivariate the analysis being undertaken.[9]

What about intercorrelations among the predictors? The main advantage of multivariate techniques like multiple regression is that they take care of these intercorrelations among the predictors, at least in a crude sense. Indeed, if one compares an ordinary subclass mean with the multivariate coefficient of the dummy variable associated with belonging to that subclass, the difference between the two is the result of adjustments for intercorrelations. Where these differences seem likely to be the result of a few major interrelations, some statement as to the factors correlated with the one in question (and responsible for the attenuation of its effect on the multivariate analysis) are often given to the reader. It is, of course, true that where intercorrelations between two predictors are too high, no analysis can handle this problem, and it becomes necessary to remove one of them from the analysis.

Perhaps the most neglected of the problems of analysis has been the problem of interaction effects. The reason is very simple. The assumption that no interactions exist generally leads to an extremely efficient analysis procedure and a great reduction in the complexity of the computing problem. Those of us who have looked closely at the nature of survey data, however, have become increasingly impressed with the importance of interaction effects and the useful way in which allowing for interactions between measured factors gets us closer to the effects of more basic theoretical constructs. Where interaction effects have not been ignored entirely, they have been handled in a number of ways. They can be handled by building combination predictors in the first place, such as combinations of age and education or the combination of age, marital status, and children known as the family life cycle.[10] Sometimes where almost all the interactions involve the same dichotomy, two separate analyses are called for.[11] Interactions are also handled by rerunning the analysis for

---

[9] Actually there are no formulas available for sampling errors of many of the statistics from complex probability samples. Properly selected part-samples can be used to estimate them by a kind of hammer-and-tongs procedure, but this is expensive. See Leslie Kish, "Confidence intervals for clustered samples," *American Sociological Review*, 22 (April, 1957), 154–65. So long as the samples are representative of a whole population the basic statistical model is presumably the "fixed" one, see M. B. Wilk and O. Kempthorne, "Fixed, mixed, and random models," *Journal of the American Statistical Association*, 50 (December, 1955), 1144–67.

See also L. Klein and J. Morgan, "Results of alternative statistical treatments of sample survey data," *Journal of the American Statistical Association*, 46 (December, 1951), 442–60.

[10] Guy Orcutt and others, *Microanalysis of Socioeconomic Systems* (New York: Harper and Brothers, 1961).

[11] For instance, hospital utilization was studied separately for men and women in Grover Wirick, Robin Barlow, and James Morgan, "Population survey: Health care and its financing," *Hospital and Medical Economics*, Volume I, Walter McNerney (Editor) (Chicago: American Hospital Association, 1962).

Participation in recreation was studied separately for those with and without paid vacations; see Eva Mueller and Gerald Gurin, *Participation in Outdoor Recreation: Factors Affecting Demand Among American Adults* (U.S. (U.S.G.P.O., ORRRC Study Report 20, 1962.)

some subgroup of the population. In a recent study of factors affecting hourly earnings, for instance, the analysis was rerun for the white, nonfarmer males only, to test the hypothesis that some of the effects like that of education were different for the non-whites, women, and farmers.[12] A difficulty with this technique, of course, is that if one merely wants to see whether the interaction biases the estimates for the whole population seriously, one reruns the analysis with the group that makes up the largest part of the sample. But if one wants to know whether there are different patterns of effects for some small subgroup, the analysis must be run for that small subgroup.

Another method of dealing with interaction effects is to look at two- and three-way tables of residuals from an additive multivariate analysis. This requires the process, often rather complicated and expensive, of creating the residuals from the multivariate analysis and then analyzing them separately.[13] Where some particular interaction is under investigation, an effective alternative is to isolate some subgroup on a combination of characteristics such as the young, white, college graduates. It is then possible to derive an estimate of the expected average of that subgroup on the dependent variable by summing the multivariate coefficients multiplied by the subgroup distributions over each of the predictors. Comparing this expected value with the actual average for that subgroup indicates whether there is something more than additive effect. It is only feasible to do this with a few interactions, just as it is possible to put in cross product terms in multiple regressions in only a few of the total possible cases. Consequently, most of these methods of dealing with interaction effects are either limited, or expensive and time-consuming.

Still another technique for finding interactions is to restrict the total number of predictors, use cell means as basic data, and use a variance analysis looking directly for interaction effects.[14] Aside from the various statistical assumptions that have to be made, this turns out to be a relatively cumbersome method of dealing with the data. It requires a good deal of judgment in the selecting of the classes to avoid getting empty cells or cells with very small numbers of cases,

[12] James Morgan, Martin David, Wilbur Cohen, and Harvey Brazer, *Income and Welfare in the United States* (New York: McGraw-Hill, 1962).

Malcolm R. Fisher, "Exploration in savings behavior," *Bulletin of the Oxford University Institute of Statistics*, 18 (August, 1956), 201–77.

[13] James Morgan, "An analysis of residuals from 'normal' regressions," in *Contributions of Survey Methods to Economics*, L. Klein (Editor) (New York: Columbia University Press, 1954).

[14] F. Gerald Adams, *Some Aspects of the Income Size Distribution* (unpublished Ph.D. dissertation, The University of Michigan, 1956); and a summary, "The size of individual incomes: Socio-economic variables and chance variation," *Review of Economics and Statistics*, XL (November, 1958), 394–8.

James Morgan, "Factors related to consumer savings" in *Contributions of Survey Methods to Economics*, L. Klein (Editor) (New York: Columbia University Press, 1954).

Mordechai Kreinin, "Factors associated with stock ownership," *Review of Economics and Statistics*, XLI (February, 1959), 12–23; "Analysis of liquid asset ownership," *Review of Economics and Statistics*, XLIII (February, 1961), 76–80.

M. Kreinin, J. Lansing, J. Morgan, "Analysis of life insurance premiums," *Review of Economics and Statistics*, XXXIX (February, 1957), 46–54.

Robert Ferber has pointed out that using the highest order interaction as "error" may hide significant main effects or lower-order interaction effects, and that the heteroscedasticity of means based on subcells of different sizes may make the tests nonconservative. He has made use of the more complex method of fitting constants which provides an exact test for interactions but assumes that the individual observations are all independent. Since this assumption is not correct for most multistage samples the results of this method are also nonconservative. See Robert Ferber, "Service expenditures at mid-century," in *Consumption and Saving*, Volume I, I. Friend and R. Jones (Editors) (Philadelphia: University of Pennsylvania Press, 1960), pp. 436–60.

and the unequal cell frequencies lead to heterogeneity of variances which makes the *F*-test nonconservative. Sometimes interaction effects are considered important only when they involve one extremely important variable. In the case of much economic behavior, current income appears to be such a variable. In this case one can rely on covariance techniques, but these techniques tend to become far too complex when a large number of other factors are involved. Also, as more and more questions arise about the meaning of current income as a measure of ability to pay, the separation of current income for special treatment becomes more doubtful.

Finally, it is also true that if we restrict the number of variables, multiple regression techniques, particularly using dummy variables, can build in almost all feasible interaction effects. One way to restrict the number of variables is to make an analysis with an initial set and run the residuals against a second set of variables. However, unless there is some logical reason why one set takes precedence over another, this is treacherous since the explanatory classifications used in the second set will have a downward bias in their coefficients if they are at all associated with the explanatory classifications used in the first set.[15]

All these methods for dealing with interaction effects require building them in somehow without knowing how many cases there are for which each interaction effect could be relevant. The more complex the interaction, the more difficult it is to tell, of course.

The problem of logical priorities in the data and chains of causation can be handled either by restricting the analysis to one level or by conducting the analysis sequentially, always keeping in mind that the logically prior variables may have to be reintroduced in later analyses on the chance that they may mediate the effects of other variables. In practice, very little analysis of survey data has paid much attention to this problem. Perhaps the reason is that only recently has anyone been able to handle the other problems so that a truly multivariate analysis was possible. And it is only when many variables begin to be used simultaneously that the problem of their position in a causal structure becomes crucial.

Finally, there is the problem remaining from section B that the constructs of theories do not have any one-to-one correspondence with the measures from the survey. Sometimes this problem is handled by building complex variables that hopefully represent the theoretical construct. The life cycle concept, for instance, has been used this way. In a recent study, a series of questions that seemed to be asking evaluations of occupations were translated into a measure which was (hopefully) an index measure of achievement motivation.[16] More commonly, the analyst has been constrained to interpret each of the measured characteristics in terms of some theoretical meaning which it hopefully has. This is often not very satisfactory. In the case of liquid assets, the amount of

---

[15] James Morgan, "Consumer investment expenditures," *American Economic Review*, XLVIII (December, 1958), 874–902, Appendix, 898–901.

Arthur S. Goldberger and D. B. Jochems, "A note on stepwise least squares," *Journal of the American Statistical Association*, 56 (March, 1961), 105–11.

[16] Morgan, David, Cohen, and Brazer, *Income and Welfare in the United States*. (New York: McGraw-Hill Book Company, Inc., 1962).

these assets a man has represents both his past propensity to save and his present ability to dissave, two effects which could be expected to operate in opposite directions. In general, the analysis of survey data has been much better than this summary of problems would indicate. Varied approaches have been ingeniously used, and cautiously interpreted.

E. PROPOSAL FOR A PROCESS FOR ANALYZING DATA

One way to focus on the problems of analyzing data is to propose a better procedure. The proposal made here is essentially a formalization of what a good researcher does slowly and ineffectively, but insightfully on an IBM sorter. With large masses of data, weighted samples, and a desire for estimates of the reduction in error, however, we need to be able to simulate this process on large scale computing equipment. The basic idea is the sequential identification and segregation of subgroups one at a time, nonsymmetrically, so as to select the set of subgroups which will reduce the error in predicting the dependent variable as much as possible relative to the number of groups. A subgroup may be defined as membership in one or more subclasses of one or more characteristics. If more than one characteristic is used, the membership is joint, not alternative.

It is assumed that where the problem of chains of causation and logical priority of one variable over another exists, that this problem will be handled by dividing the explanatory variables or predictors into sets. One then takes the pooled residuals from an analysis using the first set of predictors and analyses these residuals against the second set of predictors. The residuals from the analysis using this second set could then be run against a third set. In practice, we might easily end up with three states—early childhood or parental factors, actions and events during the lifetime, and current situational and attitudinal variables.

The possibilities of interactions between variables in different stages can be handled by reintroducing in the second or third analyses, factors whose simple effects have already been removed, but which may also mediate the effects of factors at one of the later stages, that is, nonwhites may have their income affected by education differently from whites.

Temporarily setting aside these complications, we turn now to a description of the process of analysis using the variables from any *one* stage of the causal process. Since even the best measured variable may actually have nonlinear effects on the dependent variable, we treat each of the explanatory factors as a set of classifications. As we said, our purpose is to identify and segregate a set of subgroups which are the best we can find for maximizing our ability to predict the dependent variable. We mean maximum relative to the number of groups used, since an indefinitely large number of subgroups would "explain" everything in the sample. To be more sophisticated, if we use a model based on the assumption that we want to predict back to the population, there is an optimal number of subgroups. However, as an approximation we propose that with samples of two to three thousand we arbitrarily segregate only those groups, the separation of which will reduce the total error sum of squares by at

least one per cent and do not even attempt further subdivision unless the group to be divided has a residual error (within group sum of squares) of at least two per cent of the total sum of squares. This restricts us to a *maximum* of fifty-one groups. It is just as arbitrary as the use of the 5 per cent level in significance tests and perhaps should be subject to later revision on the basis of experience.

We now describe the process of analysis in the form of a series of decision rules and instructions. We think of the sample in the beginning as a single group. The first decision is what single division of the parent group into two will do the most good. A second decision has then to be made: Which of the two groups we now have has the largest remaining error sum of squares, and hence should be investigated next for possible further subdivision? Whenever a further subdivision of a group will not reduce the unexplained sum of squares by at least one per cent of the total original sum of squares, we pay no further attention to that subgroup. Whenever there is no subgroup accounting for at least two per cent of the original sum of squares, we have finished our job. We turn now to a more orderly description of this process.

1) Considering all feasible divisions of the group of observations on the basis of each explanatory factor to be included (but not combinations of factors) find the division of the classes of any characteristic such that the partitioning of this group into two subgroups on this basis provides the largest reduction in the unexplained sum of squares.

Starting with any given group, and considering the various possible ways of splitting it into two groups, it turns out that a quick examination of any possible subgroup provides a rapid estimate of how much the error variance would be reduced by segregating it:

The reduction in error sum of squares is the same size (opposite sign) as the increase in the explained sum of squares.

For the group as a whole, the sum of squares explained by the mean is

$$N\overline{X}^2 = \frac{(\sum X)^2}{N} \tag{1}$$

and the total sum of squares (unexplained by the mean) is

$$\sum (X - \overline{X})^2 = \sum X^2 - \frac{(\sum X)^2}{N} . \tag{2}$$

If we now divide the group into two groups of size N_1 and N_2 and means $\overline{X}_1$ and $\overline{X}_2$, what happens to the explained sum of squares?

$$\text{Explained sum of squares} = N_1\overline{X}_1^2 + N_2\overline{X}_2^2. \tag{3}$$

The division which increases this expression most over $N\overline{X}^2$ clearly does us the most good in improving our ability to predict individuals in the sample.

Fortunately we do not even need to calculate anything more than a term involving the subgroup under inspection, since N and $\sum X$ remain known and constant throughout this search process.

$$N_2 = N - N_1 \tag{4}$$

$$\sum X_2 = \sum X - \sum X_1 \tag{5}$$

$$\therefore \text{explained sum of squares} = N_1\left(\frac{\sum X_1}{N_1}\right)^2 + (N - N_1)\left(\frac{\sum X_2}{N - N_1}\right)^2$$

$$= \frac{(\sum X_1)^2}{N_1} + \frac{(\sum X - \sum X_1)^2}{N - N_1}.$$

The number of cases (or proportion of sample) and the sum of the dependent variable for any subgroup are enough to estimate how much reduction in error sum of squares would result from separating it from the parent group.

If it seems desirable, a variance components model which takes account of the fact that we really want optimal prediction of members of the population not merely of the sample, can be used. Indeed, the expression for the estimate of the explained, or "between" component of variance in the population turns out to be

$$\hat{\delta}_B^2 = \frac{\dfrac{N-1}{N-2}\left[\dfrac{(\sum X_1)^2}{N_1} + \dfrac{(\sum X - \sum X_1)^2}{N - N_1}\right] - \dfrac{\sum X^2}{N-2}}{N - \dfrac{N_1^2 + N_2^2}{N}} \tag{7}$$

which, though it looks formidable, contains only one new element and that is a term from the total sum of squares of the original group which is constant and can be ignored in selecting the best split. The expression in the brackets is the explained sum of squares already derived. N, $\sum X$, and $\sum X^2$ are known and constant. The denominator is an adjustment developed by Ganguli for a bias arising from unequal N's. Where N_1 equals N_2, the denominator becomes equal to N_1. The more unequal the N's, the smaller the denominator, relative to an arithmetic mean of the N's. The ratio of the explained component of variance to the total is *rho*, the intraclass correlation coefficient. Hence, in using a population model, we are searching for the particular division of a group into two that will provide the largest *rho*.[17] Computing formulas for weighted data or a dummy (one or zero) dependent variable can be derived easily.

(2) Make sure that the actual reduction in error sum of squares is larger than one per cent of the total sum of squares for the whole sample, i.e., $>.01$ ($\sum X^2$, $-N\overline{X}^2$) (If not select the next most promising group for search for possible subdivision, etc.)

(3) Among the groups so segregated, including the parent, or bereft ones, we now select a group for a further search for another subgroup to be split off. The selection of the group to try is on the basis of the size of the unexplained

[17] R. L. Anderson and T. A. Bancroft, *Statistical Theory in Research* (New York: McGraw-Hill Book Company, 1952).

M. Ganguli, "A note on nested sampling," *Sankhya* 5 (1941), 449–52.

For an example of the use of rho in analysis see Leslie Kish and John Lansing, "The family life cycle as an independent variable," *American Sociological Review*, XXII (October, 1957), 512–4.

sum of squares within the group, or the heterogeneity of the group times its size, which comes to the same thing. It may well *not* be the group with the most deviant mean.

In other words, among the groups, select the one where

$$\sum X_{ij}^2 - N_i \overline{X}_i^2 \text{ is largest.}$$

If it is less than two per cent of the total sum of squares for the whole sample, stop, because no further subdivision could reduce the error sum of squares by more than two per cent. If it is more than two per cent, repeat Step 1.

Note that the process stops when no group accounts for more than two per cent of the error sum of squares. If a group being searched allows no further segregation that will account for one per cent, the next most promising group is searched, because it may still be possible that another group with a smaller sum of squares within it can be profitably subdivided.

Since only a single group is split off at a time, the order of scanning to select that one should not affect the results. Since an independent scanning is done each time, the order in which groups are selected for further investigation should not matter either, hence our criterion is a pure efficiency one.

Chart II shows how the process suggested might arrive at a set of groups approaching those given earlier in Table 1. The numbers are rough estimates from Table 1.

Note on Amount of Detail in the Codes

The search for the best single subgroup which can be split off involves a complete scanning at each stage of each of the explanatory classifications, and within each classification of all the feasible splits. This is not so difficult as it seems, for within any classification not all possible combinations of codes are feasible. If one orders the subclasses in ascending sequence according to their means (on the dependent variable), then it can be shown that the best single division—the one which maximizes the explained sum of squares—will never combine noncontiguous groups.

Hence, starting at either end of the ordered subgroups, the computer will sequentially add one subgroup after another to that side and subtract it from the other side, always recomputing the explained sum of squares. By "explained" we mean that the means of the two halves are used for predicting rather than the over-all mean. Whenever the new division has a higher explained sum of squares, it is retained, otherwise the previous division is remembered. But in any case, the process is continued until there is only one subgroup left on the other side, to allow for the possibility of "local maxima."

The machine then remembers the best split, and the explained sum of squares associated with it, and proceeds to the next explanatory characteristic. If upon repeating this procedure with the subclasses of that characteristic, a still larger explained sum of squares is discovered, the new split on the new characteristic is retained and the less adequate one dropped.

The final result will thus be the best single split, allowing any reasonable

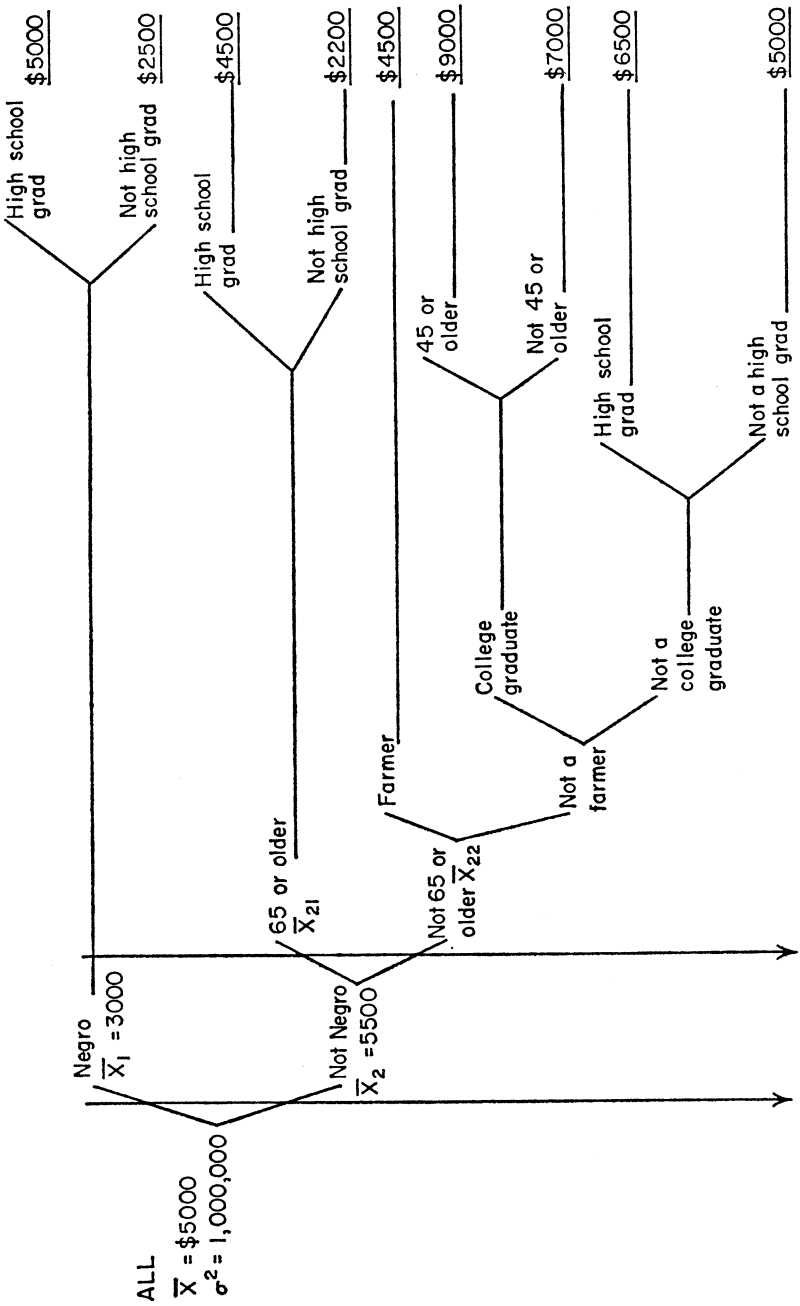


CHART II. Annual Earnings.

combination of subclasses of a single category, to maximize the explained sum of squares. It is easy to see that this choice will not depend on the order in which factors are entered, but may depend on the amount of detail with which they are coded. The number of subclasses probably should not vary too much from one factor to the next.

The authors are planning to try out such a program under a grant from the National Science Foundation. Data which have already been analyzed using dummy variable multiple regressions will be re-analyzed to see whether the new program provides new insights.

DISCUSSION

What is the theoretical model behind this process? Instead of simplifying the analysis by arbitrary or theoretical assumptions that restrict the number of variables or the way in which they operate, this process essentially restricts the complexity of the analysis by insisting that there be a large enough sample of any particular subgroup so that we can be sure it matters, and by handling problems one at a time. This is essentially what a researcher does when first investigating a sample using a sorter and his own judgment. It is assumed that the sample being used in a situation like this is a representative probability sample of a large important population. It is possible that there may be subgroups of the population whose behavior is of more importance than that of other subgroups, in which case it would be easily possible to weight the data to take account of this fact. It may be that there are certain crucial characteristics, the importance of which must be investigated. In this case, either lower admission criteria could be used or an initial arbitrary division of the sample according to this characteristic could be made before starting.

Why not take all possible subsets, in other words, all possible combinations of characteristics, and then start combining subcells where the means are close to one another? The simple reason is that there are far too many possible subsets, and since this is a sample, the means of these subsets are unstable and unreliable estimates. It is true, however, that this is the only way one would avoid all possibility of failing to discover interaction effects. Let us take a simple example of a stituation where the method we propose would fail to discover interaction effects. Suppose we have males and females, old and young, in the following proportions who go to the hospital each year, young females eight per cent, young males two per cent, old females two per cent, old males eight per cent. Assuming half the population is male and half the population is old, the old-young split would give means of five and five per cent, and the male-female split would give means of five and five per cent. Thus we would never discover that it is young females and old males who go to the hospital. One way out of this difficulty which would also vastly increase the efficiency of the machine processes would be to set up a relatively arbitrary division of the sample into perhaps ten groups to start with, groups which are known to be important and suspected to be different in their behavior. The only problem with this is that the remaining procedures will not be invariant with respect to which initial groups were selected.

Previous Work of a Similar Nature

One can never be sure that there does not exist previous work relevant to any "new" idea. William Belson has suggested a sequential, nonsymmetrical division of the sample which he calls "biological classification," for a different purpose, that of matching two groups on other characteristics used as controls so that they can be compared.[18] His procedure is restricted to the case where the criterion can be converted to a one-zero division, and the criterion for subdivision is the best improvement in discrimination. The method takes account of the number of cases, i.e., focuses on improvement in prediction, not on levels of significance. We have proposed this same focus. No rules are provided as to when to stop, or in what order to keep searching, though an intelligent researcher would intuitively follow the rules suggested here.

Another approach to the problem as been suggested and tried by André Danière and Elizabeth Gilboy. Their approach attempts to keep numerical variables whenever there appears to be linearity, at least within ranges, and to repool groups whenever there does not appear any substantial nonlinearity or interaction effect. The method is feasible only where the number of factors is limited. The pooling both of groups and of ranges of "variables" makes it complicated.[19] In practice, they found it useful to restrict the number of allowable interaction effects.

There are also studies going on in the selection of test items to get the best prediction with a limited set of predictors. But the prediction equation in these analyses always seems to be multiple regression without any interaction effects.[20] Group-screening methods have been suggested whereby a set of factors is lumped and tested and the individual components checked only if the group seems to have an effect. These procedures, however, require knowledge of the direction of each effect and again assume no interaction effects.[21] These group-screening methods are largely used in experimental designs and quality control procedures. It is interesting, however, that they usually end up with two-level designs, and our suggested procedure of isolating one subgroup at a time has some similarity to this search for simplicity.

The approach suggested here bears a striking resemblance to Sewall Wright's path coefficients, and to procedures informally called "pattern analysis." The justification for it, however, comes not fron any complicated statistical theory, nor from some enticing title, but from a calculated belief that for a large range of problems, the real world is such that the proposed procedure will facilitate understanding it, and foster the development of better connections between theoretical constructs and the things we can measure.

One possible outcome, for those who want precise measurement and testing,

---

[18] William A. Belson, "Matching and prediction on the principle of biological classification," *Applied Statistics*, VIII (1959), 65–75.

[19] André Danière and Elizabeth Gilboy, "The specification of empirical consumption structures, in *Consumption and Saving*, Volume I, I. Friend and R. Jones (Editors) (Philadelphia: University of Pennsylvania Press, 1960), pp. 93–136.

[20] Paul Horst and Charlotte MacEwan, "Optimal test-length for multiple prediction, the general case," *Psychometrika*, 22 (December, 1957), 311–24 and references cited therein.

[21] G. S. Watson, "A Study of the group-screening method," *Technometrics*, 3 (August, 1961), 371–88.

G. E. P. Box, "Integration of techniques for process control," *Transactions of the Eleventh Annual Convention of the American Society for Quality Control, 1958.*

is the development of new constructs, as combinations of the measured "variables," which are then created immediately in new studies and used in the analysis. The family life cycle was partly theoretical, partly empirical in its development. Other such constructs may appear from our analysis, and then acquire theoretical interpretation.

F. WHAT NEEDS TO BE DONE?

It may seem that the procedure proposed here is actually relatively simple. Each stage involves a simple search of groups defined as a subclass of any one classification and a selection of one with a maximum of a certain expression which is easily computed. It turns out, however, that the computer implications of this approach are dramatic. The approach, if it is to use the computer efficiently requires a large amount of immediate access storage which does not exist on many present-day computers. Our traditional procedures for multivariate analysis involve storing information in the computer in the form of a series of two-way tables, or cross-product moments. This throws away most of the interesting and potentially fruitful interconnectedness of survey data, and we only recapture part of it by multivariate processes which assume additivity. The implications of the proposed procedure are that we need to be able to keep track of all the relevant information about each individual in the computer as we proceed with the analysis.

Only an examination of the pedigree of the groups selected by the machine will tell whether they reveal things about the real world, or lead to intuitively meaningful theoretical constructs, which had not already come out of earlier "multivariate" analyses of the same data.

It may prove necessary to add constraints to induce more symmetry, such as giving priority to seriatim splits on the same characteristic, since this might make the interpretation easier. Or we may want to introduce an arbitrary first split, say on sex, to see whether offsetting interactions previously hidden could be uncovered in this way.

Most statistical estimates carry with them procedures for estimating their sampling variability. Sampling stability with the proposed program would mean that using a different sample, one would end up with the same complex groups segregated. No simple quantitative measure of similarity seems possible, nor any way of deriving its sampling properties. The only practical solution would seem to be to try the program out on some properly designed half-samples, taking account of the original sample stratification and controls, and to describe the extent of similarity of the pedigrees of the groups so isolated. Since the program "tries" an almost unlimited number of things, no significance tests are appropriate, and in any case the concern is with discovering a limited number of "indexes" or complex constructs which will explain more than other possible sets.

It seems clear that the procedure takes care of most of the problems discussed earlier in this paper. It takes care of any number of explanatory factors, giving them all an equal chance to come in. It uses classifications, and indeed only those sets of subclasses which it actually proves important to distinguish. The results still depend on the detail with which the original data were coded.

Differential quality of the measures used remains a problem. Sample complexities are relatively unimportant since measures of importance in reducing predictive error are involved rather than tests of significance, and one can restrict the objective to predicting the sample rather than the population. Intercorrelations among the predictors are adequately handled, and logical priorities in causation can be.

Most important, however, the interaction effects which would otherwise be ignored, or specified in advance arbitrarily from among a large possible set, are allowed to appear if they are important.

There is theory built into this apparently empiristic process, partly in the selection of the explanatory characteristics introduced, but more so in the rules of the procedures. Where there is one factor of supreme theoretical interest, it can be held back and used to explain the differences remaining within the homogeneous groups developed by the program. This is a severe test both for the effect of this factor and for possible first-order interaction effects between it and any of the other factors used in defining the groups.

Finally, where it is desired to create an index of several related measures, such as attitudinal questions in the same general area, the program can be restricted to these factors and to five or ten groups, and will create a complex index with maximal predictive power.