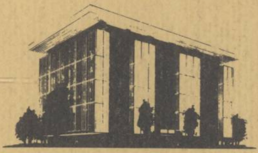
L. R. DUFFUS

Multivariate Model Building

The Validation of a Search Strategy

JOHN A. SONQUIST



SURVEY RESEARCH CENTER

ISR

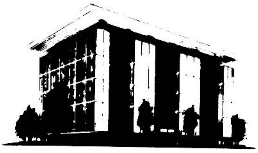
INSTITUTE FOR SOCIAL RESEARCH
THE UNIVERSITY OF MICHIGAN
ANN ARBOR, MICHIGAN

Lee R. Duffus

Multivariate Model Building

The Validation of a Search Strategy

JOHN A. SONQUIST



SURVEY RESEARCH CENTER

ISR

INSTITUTE FOR SOCIAL RESEARCH
THE UNIVERSITY OF MICHIGAN
ANN ARBOR, MICHIGAN

Library of Congress Card No. 76-630039

Printed by Braun & Brumfield
Ann Arbor, Michigan

Manufactured in the United States of America

PREFACE

The first, and primary objective of this monograph has been to provide additional validation of a formal, sequential data-analysis procedure; that is, to verify exactly what it does with certain kinds of variables under certain well–defined conditions. However, in developing, carrying out, and assessing the meanings of the validation experiments reported here, it became even clearer that the way in which one interprets the relationships between measured variables (as revealed by analytic techniques), has to depend upon several factors; among others, the match between these variables and the concepts to which they refer, and the patterns of causality that are implicit in one's theoretical orientation. Moreover, whether one makes assumptions about patterns of causality and measurement or tries to infer what they are depends, at least in part, upon whether one is trying to test theory, or, alternatively to generate or discover it. Thus, a second objective of this monograph took form; the provision of some additional guidelines for the use of such a formal sequential analysis procedure in inductive sociological research.

The assumption behind this second objective is that sociological theory ought to emerge from sociological data, to be based on it. According to this view, then, one legitimate objective of data analysis can be the generation or discovery of hypotheses, of propositions, of new conceptual frameworks. This requires the use of appropriate statistical procedures and corresponding logic and strategy; yet, the best developed parts of the methodological apparatus of sociology appear to be those dealing with hypothesis testing and the logic of verification rather than those concerned with the discovery and formulation of propositions. Consequently, a focus on the methods used to develop models and to arrive at hypotheses, would help achieve a basic objective; that is, an improved ability to put forth sociological theory within a rhetoric of generation and discovery complementing that of testing and verification.

However, a full treatment of problems of induction in sociology is far beyond the scope of this investigation. Hence, the actual tasks have been limited to the above-mentioned validation; and, using the information gained in the validation process, putting forth a series of suggestions for using this specific type of procedure in such a way as to capitalize on its ability to reveal the structure of relationships implicit in a set of data. The information thus obtained provides the starting point for the analyst in his task of model development.

This investigation had its origin in what, in retrospect, was a rather remarkable conversation between Professor James Morgan, the author, and several others, in which the topic was whether a computer ever could replace the research analyst himself, as well as replacing many of his statistical clerks. Discarding as irrelevant whether or not a computer could "think," we explored the question whether or not it might simply be programmed to make some of the decisions ordinarily made by the scientist in the course of handling a typical analysis problem, as well as doing the computations. This required examining decision points, alternative courses of action, and the logic for choosing one rather than the other; then formalizing the decision-making procedure and programming it, but with the capacity to handle many variables instead of only a few.

That this algorithm works (as is demonstrated here), and that it is a relatively simple one, led to a conjecture made elsewhere, but of such significance that it bears repeating; namely, that programming more complex research analysis algorithms of this type is likely to provide a significant increase in the analytic power available to sociologists. It is hoped, therefore, that this experimentation will also serve to suggest even better validation procedures to future students of research methods who may undertake to use simulation techniques.

I am particularly indebted to the members of the University of Chicago Sociology Department who reviewed and commented on this manuscript when it was submitted to the University of Chicago in partial fulfillment of the requirements for a doctorate. They include Professor Nathan Keyfitz, Chairman, and Professors Duncan MacRae and Jack Sawyer. Their suggestions are gratefully acknowledged. My thanks are also due, in particular, to Professor Phillip Hauser, and to Professor Leo Goodman, under whose tutoring I first came to an understanding of the analysis of variance.

I also wish to thank Rensis Likert, Director of the Institute for Social Research and his colleagues for providing a stimulating multidisciplinary environment in which a continued concern for improved methodology has sparked numerous thought-provoking discussions. Computer time and a powerful programming language (MAD) have been provided by The University of Michigan Computing Center, directed by Dr. R. C. F. Bartels.

Of the many whose thoughts and suggestions were of great value, some friends and colleagues deserve special mention. George Katona, John Lansing, Leslie Kish, William Dunkelberg, Wade Clifton, Frank Stafford, and Marvin Snowbarger reacted vigorously to many of these ideas. Donald Pelz, Frank Andrews, and Martin Gold provided additional valuable insights into the problems faced by the analyst wishing to use multivariate techniques.

Mildred Dennis, Linda Davidson, Marlys Schaeffer, Maryon Wells, and Anita Grob provided, at various times, a remarkable ability to combine speed and accuracy in typing of the manuscript. Lee Behnke's help in drawing the figures is gratefully acknowledged, and editorial help from Carolyn Hauser, Tracie Brooks and William V. Haney was invaluable.

Finally, I wish to acknowledge special debts to Professor James Morgan, at whose suggestion I first began to think about the problems of interaction effects in multivariate analysis; to my wife, Hanne, for her care and patience over an extended writing schedule; and to many good friends for their encouragement.

FOREWORD

This document describes the validation of the newly developed automatic interaction detection technique (AID). The validation is done by means of a comparison of its results with those obtained using conventional multiple classification analysis (MCA) methods. The AID analysis algorithm explored here was implemented in a large scale computer program and involves the successive segregation of sample sub-groups through the step-wise application of one-way analysis of variance techniques. A primary objective was to test each algorithm's ability to lead the analyst to a correct assessment of the structure of the predictive model implicit in the data. An additive and two interactive models were used. The data employed were generated so that the actual structure of all of the relationships was completely known beforehand. The behavior of the techniques was studied under several levels of random noise.

It is concluded that MCA cannot produce correct functional representations for interactive models, but can produce a precise, accurate, and stable model when applied to additive data. AID's representation of interactive models is clear and accurate, and it produces information for the analyst on whether to introduce additivity assumptions immediately or to develop and use interaction terms in the equation representing the model. The technique discriminates between additive and interactive models even in the presence of noise in the data. In addition, it provides information suggesting the forms these interaction terms should take. Unlike MCA, it cannot report main effects adjusted for intercorrelations.

A joint strategy is therefore suggested for the use of AID as a scanning device to locate interaction terms for use in a subsequent MCA analysis which actually estimates the model. A typology of multivariate models based on Boolean operators is also outlined, incorporating all the additive and interactive models into a single framework. A discussion of the place of interaction effects in sociological research is included.

TABLE OF CONTENTS

Page

APPENDIXES

LIST OF TABLES

LIST OF TABLES—*Continued*

Table Page

LIST OF ILLUSTRATIONS

Figure | Page

Chapter 1

DATA ANALYSIS PROBLEMS

Introduction

The research described here is focused on a particular data-analysis situation characteristic of much present-day research in sociology, a situation in which the purpose of the analysis is considerably more sophisticated than the mere reporting of descriptive statistics, but may not necessarily involve the testing of specific hypotheses formally deduced from theory. The pattern of most current sociological research is that theoretical orientations provide only guidelines as to which measurable variables and constructs are probably responsible for variations in the phenomenon to be examined. Existing theory often is not stated in such a fashion that specific, precise hypotheses can be deduced and tested statistically against the null hypothesis: at most, alternative hypotheses can be suggested.

Riley (1964) noted that, in actuality, a researcher may not even be proceeding from an interpretive point in his work but from an empirical one; or, he may alternate these phases in a single study, working back and forth between theory and data. Sometimes, she points out, it is useful to work from data to model after some of the findings are in hand. As exploratory research uncovers empirical regularities, one can look for clues to new ideas and explanations that might account for these findings. These can then be used to amplify and specify the conceptual model.

In line, then, with the pattern of most current sociological research, the tasks with which we shall be concerned are primarily associated with the inductive phase of model development rather than with deductive model testing. The problem is one of determining which of the variables for which data have been collected are actually related to the phenomenon in question, and under what conditions and through which intervening processes, with appropriate controls for spuriousness. More specifically, it is the problem of actually obtaining "research findings" with which we shall be concerned.

In the inductive phase, ex post facto explanations of the relationships found within the data form a basis for assembling a set of interrelated propositions, a middle-range theory, which describes the functioning of a specific aspect of a social system. This forms the basis for developing much more specific hypotheses to be tested later, and usually specifies which additional relationships between variables should be found, given certain conditions.

Later, in the model testing phase, verification of the correctness of these predictions provides evidence supporting the interpretations or indicating their inadequacies, and provides the basis for the construction of much more precise, formal models of behavior.

Theory then may be said to consist of those various sets of propositions which describe, at an abstract level, the functioning of a social system. The task of much research is to estimate statistically the relationships within the system.

The first step in the estimation is the determination of which variables are to enter into the various propositions; the second is to specify the form of the relations between the variables. The particular theoretical orientation which suggests the inclusion of particular variables may also suggest the precise functional form to use, or it may merely suggest certain limitations on such things as the intercept, slope, and curvature of the function. It may suggest whether cross-product terms are to be included; or it may suggest very little. In the case where precise functional form is not suggested, it is necessary to look to the statistical analysis of data for help in choosing between alternative forms. However, this is a problem faced by all analysts. We shall deal with a specific variant.

Research Objectives and the Problem of Functional Form

There are many research designs for which functional form is a relevant problem.

The research design with which we are concerned may be termed a "sample survey model," in which values of a set of predictors, X_1, X_2..., X_j, ...X_p and a dependent variable Y, have been obtained over a set of observations, or units of analysis, U_1, U_2, ...U_i, ...U_N. A weight, W_j, may also be established for each U_i if multi-stage sampling methods have been used. In particular, we shall be concerned with the case in which the X_j are nominal or ordinal scales and Y is either dichotomous or is a continuous variable or equal-interval scale.[1] We shall not be concerned with data collected at several points in time.

The X_j variables may consist of a mixture of "independent" variables, those presumed to "cause" high or low values of Y, and also specifiers (condi-

[1] Much of what is to be dealt with here also applies to other "micro" units or semi-aggregated data, such as that pertaining to counties, etc.

tions) and elaborators (intervening variables).[1] The problem is basically the explanatory analysis described by Hyman (1954).

Stated in a more concise notation, following Ezekiel and Fox (1959), we have

$$Y = f\ (X_1,\ X_2,\ ...X_p) \tag{1.1}$$

The objective is to explain the variation of the dependent variable Y by means of a function representing the joint effects of the X_j.

Theory postulates exact functional relationships between variables. But points do not lie exactly on straight lines or on other smooth functions. Thus formula (1.1), and the various forms associated with it are inadequate. The method of dealing with the problem is the introduction of a stochastic (error or disturbance) term

$$Y = f\ (X_1,\ X_2,\ ...,X_p) + e \tag{1.2}$$

The objective is to find a stable function that keeps the stochastic disturbance term at a minimum.[2] It is the methods which are used by the analyst to achieve this objective that constitute the central subject matter of this report.

However, before turning to a discussion of the problems the analyst must face in pursuing his task, it will be useful to examine the idea of a disturbance term in somewhat more detail.

According to Johnston (1960) there are three ways of rationalizing the idea of inserting a stochastic term and then trying to find ways to minimize it. First, it seems reasonable to argue that the value of Y for each and every observation could be predicted with complete accuracy if we knew all the factors responsible for variations in Y and had all the necessary data. However, in explaining human behavior the list of relevant factors may be extended almost ad infinitum. Many of the factors may not even be measurable; but even if they were, it would not be possible in practice to obtain data on them all. And even if one could do that, the number of factors would still almost certainly exceed the number of observations that it would be feasible to examine. Consequently, no statistical means exist for estimating their influence. Moreover, many variables may be presumed to have only slight effects, so that even with substantial quantities of data, the statistical estimation of their influence would be difficult and uncertain. Since many factors may be at work

[1] This report will not attempt to deal with the basic scientific problems of conceptualizing causal links or with latent and manifest functions, but only with the apparent relations between measured constructs and their congruence with an underlying causal structure.

[2] We shall be concerned with the most commonly used term, the sum of squares associated with the deviations of predicted values from actual values.

(and in a given situation may be pulling in opposite directions), we should expect small values of the stochastic term, e, to occur more frequently than large values. We are thus led to think of e as a random variable with a probability distribution centered at zero and having a finite variance of σ_e; and it is for this reason that e is referred to as a stochastic disturbance or error term.

One way out of this difficulty is to represent Y as an explicit function of a small number of what seem to be the most important X's and let the net effect of all of the excluded variables be represented by the stochastic term.

A second justification for inserting a stochastic disturbance term is the assumption that over and above the total effect of all the relevant factors, there exists a basic and unpredictable element of randomness in human responses. The latter can best be characterized by the inclusion of a random variable. However, for purposes of practical data analysis the distinctions between these two rationalizations does not matter since, for reasons of both theory and data, we hardly ever claim to have included all distinguishable and relevant factors in any functional formulation. Consequently, it can be seen that the insertion of a stochastic term is essential simply because one cannot include all of the relevant factors in any given model. Genuinely random components of behavior merely add to the variance of the term.

A third reason for the inclusion of a disturbance term lies in the errors arising from the measurement process itself. These errors, however, may be thought of as superimposed on the other two sources of disturbance and need no further explanation.[1]

In general, the simplest possible assumptions about the disturbance term appear to be: an assumption that the mean is zero; the variance is constant and independent of the X_j; and the various values of e are to be drawn independently of one another. Therefore, if the observed disturbances are large or are not random, this may be an indication that at least one important explanatory variable has been omitted, that it is correlated with some of the variables in the analysis, and that its inclusion in the disturbance term is preventing that term from displaying random behavior.

The problem faced by the analyst when he omits an important explanatory variable are primarily theoretical and are largely outside the scope of this discussion, although it can be stated that large, non-random disturbance terms arising from this omission can only be remedied by isolating and studying the other important variables. Nor shall we be concerned with measurement problems and random behavior and their effects on the size of the disturbance term.

[1] We shall not be concerned with the effects of measurement errors occurring in the data used as predictors. The effects of error here are considerably more complex and are beyond the scope of the present considerations.

However, large disturbance terms and/or non-randomness in their behavior may also reflect an incorrectly specified form of the model. Which terms should be included? The large error terms reported in much research may be due, in part, to the way in which the joint effects of the predictors are combined in the model. It is this problem of specification that is the starting point of this investigation.

Functional Form as a Data Analysis Problem

The problem of functional form may usefully be considered in the context of the other problems in which it arises. Where the number of predictors is small, the problems of isolating the relationships between the X_j and Y are manageable, even using hand computational techniques. However, when the number of predictors is large, which is typical of sociological data gathered by modern survey methods, then an analysis of the joint effects of the X_j on Y presents serious problems. These include: (1) the fact that many of the X_j are classifications, which makes them more difficult to handle than the normally distributed variables elegantly described in many statistics texts; (2) the existence of various types and amounts of measurement error not only in the dependent variable but in the X_j; (3) the inapplicability of the usual significance tests to multivariate analyses based on stratified, clustered samples; (4) non-linearities in relationships; (5) intercorrelations between the predictors; (6) interaction effects; and (7) logical priorities among variables and chains of causation.

Many of these problems have been discussed extensively in the literature. One review is presented in Morgan and Sonquist (1963). In this paper it was contended that reasonably adequate techniques had been developed for handling most of the problems of analyzing survey data. However, problems revolving around the existence of interaction effects were seen to have less adequate solutions, mainly because of the introduction of simplifying additivity assumptions in multivariate analysis. Morgan and Sonquist suggested that these assumptions were all too often unwarranted and misleading. The argument is summarized in some detail below.

The need for multivariate techniques in sociology results from the reasonable assumption that human behavior is influenced by many factors operating simultaneously and from the availability of a rich body of intercorrelated information generated by modern survey methods.

The problems of analyzing these complex data are compounded because, for the most part, the analyst must deal, not with continuous variables but with classifications; and correlations between classifications, in particular, are notoriously difficult to deal with adequately. These classifications vary all the