PERGAMON

Evaluation and Program Planning 22 (1999) 375–384

EVALUATION and PROGRAM PLANNING

www.elsevier.com/locate/evalprogplan

The utility of 'tree-generating' statistics in applied social work research[1]

Kathryn G. Wambach*

School of Social Work, University of Texas, 1925 San Jacinto, Austin, TX 78712, USA

Received 29 October 1996; received in revised form 17 June 1997; accepted 28 July 1998

Abstract

Because of the complex nature of phenomena examined in applied research, there continues to be a need to identify multivariate analytic techniques which are sensitive to interaction effects and which minimize resource and technical demands. One solution to this dilemma involves the use of a family of statistical techniques sometimes referred to as 'tree-generating' statistics. The purpose of this paper is to demonstrate the utility of such statistics for public sector service planning. Specifically, an example using an early version of 'tree-generating' statistics, Automatic Interaction Detection, is presented to provide a model for the data analysis needed to plan substance abuse programs on a local level, targeting poor, culturally diverse, adult women.

Keywords: Multivariate statistics; Program evaluation

1. Introduction

Because of the complex nature of phenomena examined in applied research, there continues to be a need to identify multivariate analytic techniques which are sensitive to interaction effects and which minimize resource and technical demands. Dumas (1989) identified two procedures sensitive to contextual variables and, consequently, ideal for planning efforts: structural equation modeling (e.g., LISREL) and meta-analysis. In applied program planning and evaluation settings, however, few local communities are likely to have numerous, relatively recent, local data sets that are needed for meta-analysis regarding target problems which are known to have substantial regional and temporal variability (e.g., substance abuse), have the resources to gather the large sample size preferred for structural equation modeling, or have the expertise and resources available to utilize either procedure. One solution to this dilemma involves the use of a family of statistical techniques, sometimes referred to as 'tree-generating' statistics, as an alternative approach.

2. Overview of 'tree-generating' statistics

Derived from the work of Morgan and Sonquist (1963), this family of statistical techniques is often characterized as 'tree-based' or 'tree-generating' because the typical display of results resembles a tree. Designed to detect complex statistical relationships, a number of specific programs now exist which can accommodate both categorical and quantitative dependent variables, operate in both mainframe and PC environments, and include options in addition to those addressed in this discussion of an early, mainframe example of this statistical family. This early version

* Tel.: +1-512-471-7287; fax: +1-512-471-9600.

E-mail address: kwambach@mail.utexas.edu (K.G. Wambach).

[1] The data discussed in this paper were drawn from research supported by the US Department of Health and Human Services, Centers for Disease Control (Grant #U62/CCU403278) through subcontract with the Florida Department of Health and Rehabilitative Services by the Florida State University, Center for Human Services Policy and Administration. See Imershein et al. (1989) for details of the original study.

PII: S0149-7189(99)00021-X

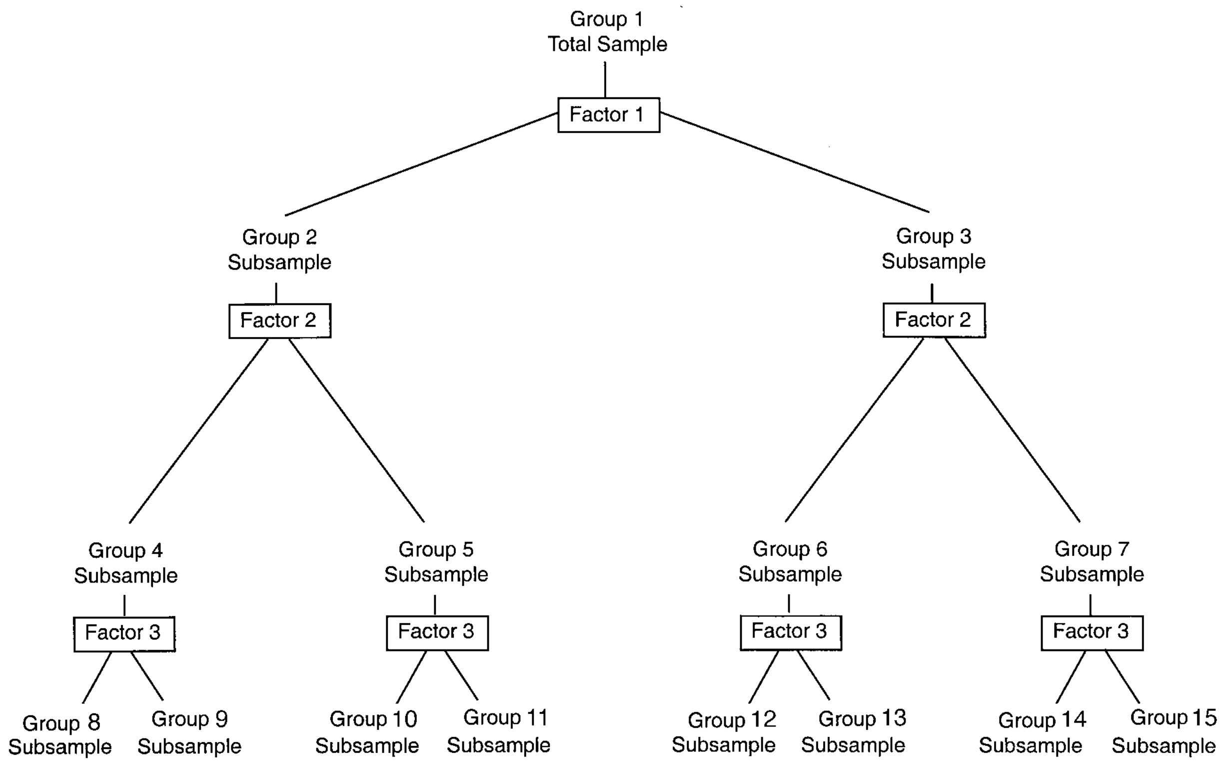

Fig. 1. Sample AID results for additive, linear data.

substantially represents the mathematical logic for later versions and is used in this demonstration of utility to avoid endorsement of any particular current product.[2]

The Automatic Interaction Detector (AID), Version 2 (Sonquist & Morgan, 1964) procedure was designed for "...determining which of the variables, for which data have been collected, are related to the phenomenon in question, under what conditions, and through what intervening processes, with appropriate controls for spuriousness" (p. 2). Utilizing the basic principle of least squares, the AID program performs a sequential one-way analysis of variance process focused on power in reducing error in the dependent variable. The sample is divided into a mutually exclusive series of subgroups through a series of binary splits (see Fig. 1 for an example of an AID tree in which the data are additive and linear so that results from a multiple regression analysis would be identical to an AID analysis). After the initial step when the total sample is the only group, the subgroup with the largest total sum of squares is selected and split into two subgroups on the independent variable which provides the largest reduction in the unexplained sum of squares. The process continues until no subgroup contains a specified fraction of the original total sum of squares, no split can reduce the unexplained variance by a specified amount, and/or the subgroups no longer include specified number of respondents.

The AID procedure consequently avoids the restrictive assumptions of linearity and additivity; further, collinearity does not threaten the integrity of the analysis as in the multiple regression technique. Although R^2 (the proportion of explained variance) is interpreted the same for both methods, the Beta-squared coefficient in AID indicates the proportion of the total variance explained by that variable across the entire procedure. In contrast, the squared multiple partial correlation coefficients in regression express the unique variance accounted for by that variable controlling (i.e., removing the additive and linear effects) for the other independent variables. In the sequential process of the AID analysis, no variables are controlled; groups are simply split into subgroups on the basis of reducing error sum of squares. In summary, the AID procedure was designed to be sensitive to

[2] Current versions of the AID-family tree-generating statistical packages include CHAID (available from SPSS), FIRM (available from the University of Minnesota, School of Statistics), and TREEDISC (available from SAS). Another version, SEARCH, was part of the OSIRIS software (available from the University of Michigan, Institute for Social Research) and is no longer available; SAS, however, is developing a version of SEARCH for marketing. Related families of tree-generating statistics (e.g., CART available from Salford Systems and SYSTAT) differ more substantially from the original AID family. Please note that this listing is based solely on the author's awareness and may not be complete.