DOCUMENT RESUME

ED 189 155 TM 800 326

AUTHOR Wolfle, Lee M.
TITLE Path Analysis with Categorical Data: Applications to Education.
INSTITUTION Virginia Polytechnic Inst. and State Univ., Blacksburg.
SPONS AGENCY National Center for Education Statistics (DHEW), Washington, D.C.
PUB DATE Apr 80
CONTRACT 300-78-0516
NOTE 12p.: Paper presented at the Annual Meeting of the American Educational Research Association (64th, Boston, MA, April 7-11, 1980).

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Academic Ability: Data: Data Analysis: *Dropouts: *Grade Point Average: Higher Education: Hypothesis Testing: Models: *Path Analysis: *Racial Factors
IDENTIFIERS *Categorical Data: *Causal Analysis: Chi Square: National Longitudinal Study High School Class 1972

ABSTRACT

An extension of the methods of path analysis to include studies of categorical data was described and exemplified in a causal study of college dropouts. The usual models and methods of causal (path) analysis were designed for the study of quantitative variables and are not appropriate when the variables under investigation are categorical. Extensions of the loglinear analysis of contingency tables to include cases with a specified order of priority for variables in a causal model is used to expand the range of theoretical questions that educational researchers can address. The example presented in this document concerns the problem of withdrawal from institutions of higher education using data from the National Longitudinal Study. The model related a respondent's race and ability, two exogenous variables, to postsecondary grades and both grades and ability to the dropout variable. The results support the proposed causal model and indicate that the effects of ability and grades are independent of race. (Author/CTM)

PATH ANALYSIS WITH CATEGORICAL DATA: APPLICATIONS TO EDUCATION

Lee M. Wolfle

Virginia Polytechnic Institute and State University

This paper was prepared for presentation at the annual meetings of the American Educational Research Association, Boston, April 7-11, 1980. Work on this project was in part supported by the National Center for Education Statistics, U.S. Department of Education (No. 300-78-0516). Dianne Robertshaw assisted in some of the computations reported in this paper; her assistance is gratefully acknowledged.

PATH ANALYSIS WITH CATEGORICAL DATA: APPLICATIONS TO EDUCATION

Lee M. Wolfle

The usual models and methods of causal analysis (see, e.g., Wright, 1921; Duncan, 1966, 1975) were designed for the study of quantitative (continuous) variables. However, the usual methods of path analysis are not appropriate when the variables under investigation are categorical.[1] Recent advances in the loglinear analysis of contingency tables (see, e.g., Bishop, Fienberg and Holland, 1975) have been extended by Goodman (1972, 1973, 1979) to include cases with a specified order of priority for variables in a causal model.

One of the salient features of path analysis is that the formulation of causal models forces the researcher to express ideas in explicit form, and allows one to read unambiguously the ideas of others (Wolfle, forthcoming). Goodman's analog to quantitative path analysis also creates diagramatic representations of causal effects, which are estimated with loglinear or logit models. While the method of estimation necessarily varies between quantitative and qualitative path analysis, both systems force a degree of explicitness desirable in social science documents.

To illustrate an educational application of path analysis with categorical data, consideration will be given to the problem of withdrawal from institutions of higher education. Tinto (1975) has suggested (after Durkheim, 1951) that dropouts are more likely to be individuals

---

[1] A reviewer of an earlier version of this paper incorrectly noted that path analysis with categorical data could be achieved through LISREL (Jöreskog and Sörbom, 1978). Au contraire, LISREL (like all regression or correlation procedures) requires the input (or computation) of a variance-covariance matrix; because such matrices presuppose continuous data, LISREL is <u>not</u> appropriate when the data are categorical.

insufficiently integrated into the fabric of the social system. Consequently, students in institutions of higher education are more likely to dropout[2] when they fail to become integrated into the relatively meritocratic reward structure. Specifically, the model proposed here relates a respondent's race and ability (two exogenous variables) to postsecondary grades, hence to the dropout variable. The assumed model is shown in Figure 1.

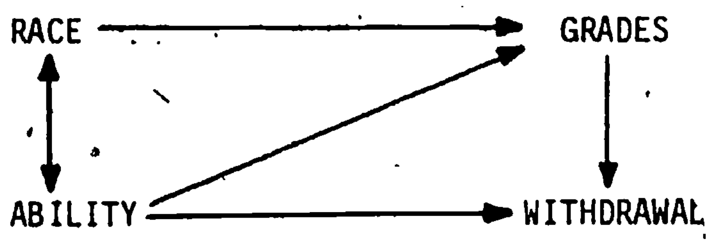

Figure 1. Conceptual Model of College Withdrawal

The model specifies that race and ability are associated, but for reasons unanalyzed in the present model; hence, these variables are exogenous. Both race and ability are seen as additive causes of college grades (a manifest measure of integration into the social system). To explain withdrawal from college, ability and grades are hypothesized to be additive causes of withdrawal, but the model specifies no additive effect of race once ability and grades have been controlled for.

The data used in the estimation of this model were taken from the National Longitudinal Study of the 1972 high school graduating class (see Levinsohn, et al., 1978). The sample included in this analysis represents that portion of the 1972 high school graduating class which went on to attend in the fall of 1972 either a two-year or four-year college or

---

[2] This may be analagous to suicide in Durkheim's (1951) analysis.

university. Since the initial survey in 1972, the respondents have been resurveyed three times, in 1973, 1974, and 1976. At each followup survey respondents were asked if they were in school, and if not, had they graduated. Dropouts were defined as those people who entered a two or four-year college or university in the fall of 1972, who subsequently reported themselves not to be attending school, but did not graduate, and had not re-entered school at the time of the last survey in 1976. Nondropouts were defined as people who had attended school continuously since graduating from high school, or had graduated.

Three variables were considered causal antecedents of college withdrawal. These were respondent's race, ability, and college grades. Race was measured by a dichotomous variable which classified whites into one group, and blacks, Mexican-Americans, Peurto Ricans, and American Indians into a second group. Ability was a trichotomous variable calculated from a scaled ability test administered while the respondents were in their senior year of high school; the respondents were classified according to whether they fell into the lower quartile, the middle two quartiles, or the upper quartile of the test-score distribution. Grades were determined by respondent's self-report, and were categorized into four groups: mostly A's, mostly B's, mostly C's, or mostly D's or F's. The question on grades was repeated in each followup survey. For nondropouts the answer to the 1976 survey was used to classify the respondents. For dropouts the answer to the last survey before dropping out was used (because these respondents would have been routed around the grade question in subsequent surveys). Thus, the grades for dropouts could have been taken from either the 1973, 1974, or 1976 survey.

These data are shown in Table 1, which cross-classifies 4392 respondents according to their distribution among the variables already defined. Note that these numbers are frequency counts, and that Table 1 is merely a 2 x 3 x 4 x 2 contingency table.

To determine whether the causal model specified above is congruent with the observed data, the observed frequencies in the table are compared with the frequencies estimated by the model, and an assessment is made of the goodness of fit using the likelihood ratio chi-square statistic. The expected frequencies are estimated using the loglinear approach (see, e.g., Bishop, Fienberg and Holland, 1975; Fienberg, 1977).

Goodman (1973, 1979) suggests that a model such as the one illustrated in Figure 1 can be estimated by taking the marginal distribution of race and ability as given, and testing for the links between grades and race, and grades and ability. The second part of the model assumes the relationship between race, ability and grades, and tests the links between withdrawal and grades, and withdrawal and ability.

Specifically, for the model with grades as the response variable, and race and ability as explanatory, one compares the [RA][RG][AG] model to the completely saturated model; if the [RA][RG][AG] model fits just as well as the saturated model, one concludes the effects of race and ability on grades are additive (i.e., no interaction among race, ability and grades). The results of this test are shown in the upper panel of Table 2. The letters define the variables as shown in Table 1. The [RA][RG][AG] model has a likelihood ratio chi-square statistic of 13.63 with 6 degrees of freedom. Compared to the saturated model (χ^2 = 0.0, with df = 0), one sees that the [RA][RG][AG] model fits the data reasonably well; the expected frequencies under the [RA][RG][AG] model do not differ significantly at the .01 level from the observed frequencies. One would conclude at this level of probability that the effects of race and ability on grades are additive. Comparisons of the models [RA][AG] and [RA][RG] to [RA][RG][AG] (note that the [RA] marginal effect is never removed from the model because it is a given association as far as grades is concerned) reveals that the removal of the effect of either race or ability on grades seriously erodes the model's ability to reproduce the observed frequencies. That is, both race and ability independently contribute to the explanation of the distribution of grades.

The model with college withdrawal as the response variable, and race, ability and grades as explanatory, specifies the most appropriate model to be [RAG][AD][GD]. That is, the association of race, ability and grades is taken as given, and withdrawal depends upon the additive effects of ability and grades. To determine if, in fact, the effects of race, ability and grades on withdrawal are additive, one compares the [RAG][RD][AD][GD] model to the interactive model [RAG][RAD][RGD][AGD]. These results are shown in the lower panel of Table 2, in lines (4) and (5). The likelihood ratio chi-square statistic may be partitioned, and the presence of interactive associations may be determined by testing the difference between the chi-squares. The partitioned chi-square is 23.92 with 11 degrees of freedom; this value is not significant at the .01 level of probability, and one concludes the effects of race, ability and grades on withdrawal are additive.

To determine if the effect of race on withdrawal is negligible once ability and grades are accounted for, one compares the [RAG][AD][GD] model to the [RAG][RD][AD][GD] model. These results are shown in the lower panel of Table 2 in lines (5) and (6), and reveal that the [RAG][AD][GD] model fits the data very well indeed compared to the model with an [RD] marginal effect. Thus, the hypothesis of no race effect on withdrawal is confirmed. Finally, lines (7) and (8), compared with line (6) show that both ability and grades have important additive effects on withdrawal; dropping either marginal effect seriously erodes the model's ability to reproduce the observed frequencies.

In conclusion, withdrawal from colleges and universities depends upon the additive effects of ability and grades, but not race. If we take grades to be a manifest measure of integration into a meritocratic reward structure, we see that those people insufficiently integrated into the system are those most likely to withdraw from the system. Moreover, those people with less ability are also those more likely to withdraw. These effects of ability and grades are independent of race.

Causal analyses with quantitative variables have become a useful means of understanding educational phenomena. Further extensions of causal modeling techniques to include categorical variables will expand the range of theoretical questions educational researchers are able to address. For those interested in learning more about the loglinear analyses of contingency tables, the texts of Fienberg (1977) and Bishop, Fienberg and Holland (1975) are recommended. The papers by Goodman (1973, 1979) explicate models which assume underlying causal priorities among the variables. These references will serve to explain how to analyze multivariate contingency tables. Hopefully, this paper has exemplified to the educational community the utility of these causal methods. Theoretically interesting questions which depended upon the analysis of categorical data were heretofore unanswerable. These new methods open new avenues for the development and testing of theory in education.

TABLE 1

Observed Cross-Classification of 4392 High School Graduates in 1972 Who Attended Two or Four Year Colleges, by College Grades, Ability, and Race

(R) Race	(A) Ability	(G) Grades	(D) Dropout Yes	(D) Dropout No
White	Low	A	15	9
		B	22	14
		C	85	56
		D or F	12	1
	Medium	A	103	234
		B	127	333
		C	270	368
		D or F	43	9
	High	A	140	806
		B	106	486
		C	157	331
		D or F	27	6
Other	Low	A	12	13
		B	15	27
		C	75	68
		D or F	13	4
	Medium	A	17	29
		B	27	56
		C	65	95
		D or F	14	2
	High	A	5	35
		B	3	18
		C	14	25
		D or F	0	0

TABLE 2

Likelihood Ratio Chi-Square Values for Some Loglinear Models Applied to the Data in Table 1

Model	Fitted Marginals	Degrees of Freedom	Likelihood Ratio Chi-Square
GRADES			
(1)	[RA][RG][AG]	6	13.63
(2)	[RA][AG]	9	30.01
(3)	[RA][RG]	12	427.82
WITHDRAWAL			
(4)	[RAG][RAD][RGD][AGD]	6	4.56
(5)	[RAG][RD][AD][GD]	17	28.48
(6)	[RAG][AD][GD]	18	29.48
(7)	[RAG][AD]	21	240.85
(8)	[RAG][GD]	20	157.86

REFERENCES

Bishop, Yvonne M. M., Stephen E. Fienberg, and Paul W. Holland
1975 Discrete Multivariate Analysis: Theory and Practice. Cambridge: The MIT Press.

Duncan, Otis Dudley
1966 "Path analysis: sociological examples." American Journal of Sociology 72: 1-16.

1975 Introduction to Structural Equation Models. New York: Academic Press.

Durkheim, Emile
1951 Suicide: A Study in Sociology. Translated by John A. Spaulding and George Simpson. New York: The Free Press.

Fienberg, Stephen E.
1977 The Analysis of Cross-Classified Categorical Data. Campbridge: The MIT Press.

Goodman, Leo A.
1972 "A general model for the analysis of surveys." American Journal of Sociology 77: 1035-1086.

1973 "Causal analysis of data from panel studies and other kinds of surveys." American Journal of Sociology 78: 1135-1191.

1979 "A brief guide to the causal analysis of data from surveys." American Journal of Sociology 84: 1078-1095.

Jöreskog, Karl G., and Dag Sörbom
1978 LISREL: Analysis of Linear Structural Relationships by the Method of Maximum Likelihood, User's Guide. Chicago: National Educational Resources, Inc.

Levinsohn, Jay R., Louise B. Henderson, John A. Riccobono, and R. Paul Moore
1978 National Longitudinal Study: Base Year, First, Second and Third Follow-Up Data File Users Manual. Washington, D.C.: National Center for Education Statistics.

Tinto, Vincent
1975 "Dropout from higher education: a theoretical synthesis of recent research." Review of Educational Research 45: 89-125.

Wolfle, Lee M.
forth-coming "Strategies of path analysis." American Educational Research Journal.

Wright, Sewell
1921 "Correlation and causation," Journal of Agricultural Research 20: 557-585.