INTRODUCTION

ONE

Latent Class Analysis

The Empirical Study of Latent Types, Latent Variables, and Latent Structures

Leo A. Goodman

1. INTRODUCTION

I begin this introductory section on latent class analysis[1] by considering this subject in its simplest context; that is, in the analysis of the cross-classification of two dichotomous variables, say, variables A and B. In this context, we have the simple two-way 2×2 cross-classification table $\{A, B\}$, where the two rows of the 2×2 table correspond to the two classes of the dichotomous variable A, and the two columns of the 2×2 table correspond to the two classes of the dichotomous variable B. We let P_{ij} denote the probability that an observation will fall in the ith row ($i = 1, 2$) and jth column ($j = 1, 2$) of this 2×2 table. In other words, P_{ij} is the probability that an observation will be in the ith class ($i = 1, 2$) on variable A and in the jth class ($j = 1, 2$) on variable B. When variables A and B are statistically independent of each other, we have the simple relationship

$$P_{ij} = P_i^A P_j^B, \tag{1}$$

where P_i^A is the probability that an observation will fall in the ith row of the 2×2 table, and P_j^B is the probability that an observation will fall in the jth column of the 2×2 table. In other words, P_i^A is the probability that an observation will be in the ith class on variable A, and P_j^B is the probability that an observation will be in the jth class on variable B; with

$$P_i^A = P_{i+} = \sum_j P_{ij}, \quad P_j^B = P_{+j} = \sum_i P_{ij}. \tag{2}$$

When variables A and B are not statistically independent of each other, that is, when formula (1) does not hold true, which is often the case in many areas of empirical research (when both variables A and B are of

substantive interest), the researcher analyzing the data in the 2×2 table will usually be interested in measuring the nonindependence between the two variables (A and B); and there are many different measures of this nonindependence. (Even for the simple 2×2 table, there are many such measures.) However, all of these measures of nonindependence (or almost all of them) are *deficient* in an important respect. Although these measures of nonindependence may help the researcher to determine the magnitude of the nonindependence between the two variables (A and B), they *cannot* help the researcher determine whether this nonindependence is *spurious*. In other words, none (or almost none) of the usual measures of the nonindependence between variables A and B can help the researcher to determine whether the observed relationship (the nonindependence) between variables A and B can be explained away by some other variable, say, variable X, where this variable X may be unobserved or unobservable, or latent. Is there a latent variable X that can explain away the observed (manifest) relationship between variables A and B, when we take into account the (unobserved) relationship that this latent variable X may have with variable A and the (unobserved) relationship that the latent variable may have with variable B? The use of latent class models can help the researcher to consider such questions.

The latent variable X introduced previously can be viewed as a possible explanatory variable. It can be used at times to explain away the observed relationship between variables A and B even when this observed relationship between the two observed variables (A and B) is statistically significant. At other times, the explanatory latent variable X can be used to help the researcher to explain more fully (rather than to explain away) the observed relationship between the two observed variables. With some sets of data, an appropriate latent class model might include several latent variables as explanatory variables; and these latent variables might be useful in helping the researcher to explain more fully (or to explain away) the observed relationships among the set of observed variables under consideration. Use of such latent class models can help the researcher in many ways, as we shall see later in this exposition on the use of latent class models and in the chapters that follow in this book on latent class analysis.

The problem of measuring the relationship (the nonindependence) between two (or more) observed dichotomous (or polytomous) variables has a long history. This problem has been considered by many researchers in many fields of inquiry at various times throughout the twentieth century, and it is a topic that was also considered by some eminent

scholars in the nineteenth century. The use of latent class models as a tool to help researchers gain a deeper understanding of the observed relationships among the observed dichotomous (or polytomous) variables has, in contrast, a much shorter history in the twentieth century, but it might be worthwhile to note here that some mathematical models that were used earlier in some nineteenth-century work can now be viewed as special cases of latent class models or of other kinds of latent structures. With respect to these nineteenth-century models, we refer, in particular, to some work by C. S. Peirce, the great philosopher and logician, who was also an able scientist and mathematician. (In addition to the recognition he has received for some of his other work, he is also sometimes referred to as the "founder of pragmatism.") Peirce introduced such a model (i.e., a latent structure) in order to gain further insight into the relationship between two observed dichotomous variables in the context of measuring the success of predictions (Peirce, 1884; Goodman and Kruskal, 1959). We shall return to this example in a later section herein.

The main development of latent class models has taken place during the last half of the twentieth-century, and the practical application of these models by researchers in various fields of inquiry has become a realistic possibility only during the last quarter of the twentieth century (after more efficient and more usable statistical methods were developed and more general latent class models were introduced). Although the problem of measuring the relationship (the nonindependence) between two or more observed dichotomous (or polytomous) variables has arisen and has been considered in many fields of inquiry at various times throughout the nineteenth and twentieth centuries, we can expect that researchers in some of these fields of inquiry (and in other fields as well) will find that the introduction and application of latent class models can help them to gain further insight into the observed relationships among these observed variables of interest. The introduction of latent class models can insert a useful perspective into the study of the relationships among these variables.

Thus far in this introductory section on latent class analysis we have focused our attention on the possible use of a latent dichotomous or polytomous variable (or a set of such latent dichotomous or polytomous variables) as an explanatory variable (or as explanatory variables) in the study of the relationships among a set of observed (or manifest) dichotomous or polytomous variables. (In this case, our primary focus is on the set of observed variables and on possible explanations of the observed relationships among these variables.) We can also use the latent class

models in those situations in which the observed dichotomous or polytomous variables may be viewed as indicators or markers for an unobserved latent variable X, where the unobserved variable is, in some sense, being measured (in an indirect way and with measurement error) by the observed variables. (In this case, our primary focus is on the unobserved latent variable; and the observed variables are, in some sense, ascriptive or attributive variables pertaining to the latent variable.) We can also use models of this kind in the study of the relationships among a set of unobserved (or latent) dichotomous or polytomous variables in the situation in which there are observed dichotomous or polytomous variables that can be viewed as indicators or markers for the unobserved (or latent) variables. (In this case, our primary focus is on the set of unobserved latent variables and on the unobserved relationships among these variables.)

2. THE LATENT CLASS MODEL

Now let us consider the latent class model in the situation in which variable A is an observed (or manifest) dichotomous or polytomous variable having I classes ($i = 1, 2, \ldots, I$), variable B is an observed (or manifest) dichotomous or polytomous variable having J classes ($j = 1, 2, \ldots, J$), and variable X is an unobserved (or latent) dichotomous or polytomous variable having T classes ($t = 1, 2, \ldots, T$). Let π_{ijt}^{ABX} denote the joint probability that an observation is in class i on variable A, in class j on variable B, and in class t on variable X; let $\pi_{it}^{\bar{A}X}$ denote the conditional probability that an observation is in class i on variable A, given that the observation is in class t on variable X; let $\pi_{jt}^{\bar{B}X}$ denote the conditional probability that an observation is in class j on variable B, given that the observation is in class t on variable X; and let π_t^X denote the probability that an observation is in class t on variable X. The latent class model in this situation can be expressed simply as follows:

$$\pi_{ijt}^{ABX} = \pi_t^X \pi_{it}^{\bar{A}X} \pi_{jt}^{\bar{B}X}, \quad \text{for } i = 1, \ldots, I;\ j = 1, \ldots, J;\ t = 1, \ldots, T. \tag{3}$$

This model states that variables A and B are conditionally independent of each other, given the class level on variable X. That is,

$$\pi_{ijt}^{\bar{A}\bar{B}X} = \pi_{ijt}^{ABX} / \pi_t^X = \pi_{it}^{\bar{A}X} \pi_{jt}^{\bar{B}X}, \tag{4}$$

where $\pi_{ijt}^{\bar{A}\bar{B}X} = \pi_{ijt}^{ABX} / \pi_t^X$ is the conditional probability that an observation is in class i on variable A and in class j on variable B, given that the observation is in class t on variable X.

We have presented the latent class model above for the situation in which there are only two observed (manifest) variables (say, A and B). This we do for expository purposes in order to consider this subject in its simplest context. However, it should be noted that some special problems arise when latent class models are considered in the situation in which there are only two observed variables that do not arise in the situation in which there are more than two observed variables. However, these problems need not deter us here. For illustrative purposes, next we shall consider some examples in which latent class models are applied in the analysis of cross-classified data in the situation in which there are two observed variables and also in the situation in which there are more than two observed variables.

3. FIRST EXAMPLE: THE ANALYSIS OF THE RELATIONSHIP BETWEEN TWO OBSERVED VARIABLES

To begin, let us consider the cross-classified data presented in Table 1. These data on the relationship between parental socioeconomic status and mental health status were analyzed earlier by various researchers using various methods of analysis. For the purposes of the present exposition, we note here that the data were used earlier to illustrate both the application of association models and the application of correlation models in measuring the observed relationship (the nonindependence) between the two polytomous variables in Table 1 (see, e.g., Goodman, 1979a, 1985; Gilula and Haberman, 1986). These data were also analyzed by using a latent class model (see, e.g., Goodman, 1987), but in the present

Table 1. Cross-Classification of 1660 Subjects

		Mental Health Status			
Parental Socioecon. Status		**Well 1**	**Mild Symptoms 2**	**Mod. Symptoms 3**	**Impaired 4**
High	1	64	94	58	46
	2	57	94	54	40
	3	57	105	65	60
	4	72	141	77	94
	5	36	97	54	78
Low	6	21	71	54	71

Note: This cross-classification of 1660 subjects is according to their parental socioeconomic status and their mental health status, as shown.

Source: Srole et al. (1962).

exposition we shall more fully examine how the analysis of latent classes in the present context can change in a dramatic way our view of the observed relationship between the two polytomous variables in Table 1 and in other such tables. We shall also include here some simplifications and improvements of some of the results presented in the earlier literature on the analysis of cross-classified data of the kind presented in Table 1 and of the kind presented in the other examples considered herein.

In Table 1, the row categories pertain to parental socioeconomic status (from high to low), and these categories have been numbered (six row categories numbered from 1 to 6); and the column categories pertain to mental health status (from well to impaired), and these categories have also been numbered (four column categories numbered from 1 to 4). These numbers have no special numerical meaning except to indicate which row is being referred to (and possibly where the row appears in the ordering of the rows, if the rows are considered to be ordered) and which column is being referred to (and possibly where the column appears in the ordering of the columns, if the columns are considered to be ordered). With the earlier analysis of Table 1 using correlation models, it was possible to find a set of meaningful numerical scores (different from the integers from 1 to 6) for the row categories and a set of meaningful numerical scores (different from the integers from 1 to 4) for the column categories in Table 1; and the correlation calculated between the meaningful scores for the row categories and the meaningful scores for the column categories for the cross-classified data in Table 1 was also meaningful. The correlation turned out to be small in magnitude, 0.16, but it was statistically significant. Also, with the earlier analysis of Table 1 using the association models, a somewhat similar kind of result was obtained; however, with these models, an index of intrinsic association (rather than an index of correlation) turns out to be meaningful, and the row scores and column scores that are obtained with these association models also turn out to be meaningful (but these scores differ in their meaning from the corresponding scores obtained with the correlation models).

The association models and the correlation models view the two-way 6×4 table in a symmetrical way; they consider the association (or the correlation) between the row variable and the column variable, treating the row and column variables symmetrically. However, these models can also be interpreted in an asymmetrical way in a situation in which we might be interested in the possible dependence of, say, the column variable on the row variable. Here we might be interested in, for example, the possible dependence of mental health status on parental socioeconomic

status. With the application of the association models in this context, we can consider the odds of being, say, in mental health status 1 rather than 2, or the odds of being in mental health status 2 rather than 3, or the odds of being in mental health status 3 rather than 4; and we can use the association models to describe how these odds change in a systematic way as we consider these odds for those with different parental socioeconomic status – how these odds change as we move from considering those whose parental socioeconomic status is at the high level to those whose parental socioeconomic status is at a lower level. A somewhat similar kind of result can be obtained when the correlation models are applied to Table 1.

Both the association models and correlation models could be viewed as somewhat improved or somewhat more sophisticated forms of an ordinary regression analysis or an ordinary correlation analysis, or logit analysis or loglinear analysis. They describe, in one way or another, how the two observed variables, the row variable and the column variable, *appear* to be related to each other, or they describe how one of the variables, say, the column variable, *appears* to be related to the other variable or to be affected by the other variable. For the data in Table 1, the apparent relationship or the apparent effect is statistically significant – nevertheless, I continue to refer here to the apparent (or manifest) relationship or to the apparent (or manifest) effect. All of the methods just mentioned (regression analysis, correlation analysis, logit analysis, loglinear analysis, association model analysis, and correlation model analysis) are concerned with apparent (or manifest) relationships or apparent (or manifest) effects. With the introduction of latent class models, we can examine whether these statistically significant apparent relationships and apparent effects might actually be spurious.

Let us now suppose that there are, say, two kinds of families: One kind I shall call simply the "favorably endowed," and the other kind I shall call the "not favorably endowed." These can be viewed as latent categories or latent classes or latent types of families; and we can consider the latent variable E for "endowment" (favorable endowment or not favorable endowment) as a latent dichotomous variable. Further suppose that this latent variable E affects what parental socioeconomic status is attained, and also that it is this latent variable E that affects what the mental health status is of the individual. If this is the case, the relationships among variable S (parental socioeconomic status), M (mental health status), and E (endowment status) can be described by Figure 1(a), where variables S and M are conditionally independent of each other, given the level of variable E (i.e., given the endowment status of the family). In this

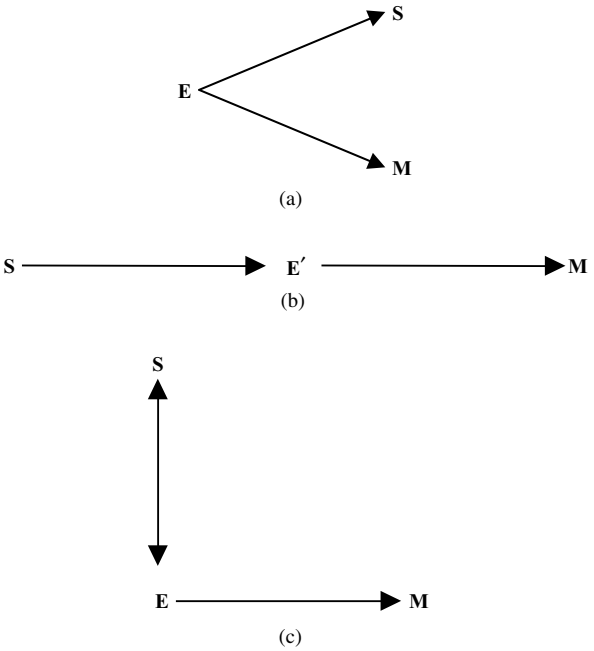

Figure 1. Three different views of the relationship between variables S and M and variable E or variable E': (a) Explanatory latent variable E viewed as antecedent to variables S and M; (b) explanatory latent variable E' viewed as intervening between variables S and M; and (c) explanatory latent variable E viewed as coincident or reciprocal with S and antecedent to M.

case, the statistically significant relationship observed between parental socioeconomic status and mental health status is spurious.

Next let us consider a somewhat different situation. Let us now suppose that there are, say, two kinds of individuals (rather than two kinds of families): One kind I shall call the "favorably endowed," and the other kind I shall call the "not favorably endowed." These can be viewed as latent categories or latent classes or latent types of individuals; and we can consider the latent variable E' for "endowment" (favorable endowment or not favorable endowment) as a latent dichotomous variable. Further suppose that it is this latent variable E' that affects what the mental health status is of the individual, and also that it is parental socioeconomic status that affects what the endowment status E' is (favorable or not favorable) of the individual. If this is the case, the relationships among variables S, M, and E' can then be described by Figure 1(b), with variable S affecting variable E', and variable E' affecting variable M. In this case too, variables S and M are again conditionally independent of each other, given the level of variable E'.

In Figure 1(a), the latent variable E is an antecedent variable, and in Figure 1(b), the latent variable E' is an intervening variable. In addition to Figures 1(a) and 1(b), we might also consider Figure 1(c). Here we again have a somewhat different situation, that is, different from the situations described by Figures 1(a) and 1(b). Figure 1(c) can be used to describe the situation in which variable S (parental socioeconomic status) and latent variable E (endowment status) reciprocally affect each other (or variables S and E are coincident with each other), and it is variable E that affects what the mental health status is of the individual. [In other contexts, where the column variable M might be viewed as prior to the row variable S, we could consider Figure 1(a), and we could also consider the corresponding figures obtained when the symbols S and M are interchanged in Figure 1(b) and also in Figure 1(c).] Each of these figures is congruent with the situation in which variables S and M are conditionally independent of each other, given the level of the latent variable (E or E').[2] Using latent class models, we can examine whether this conditional independence is congruent with the data in Table 1.

It may be worthwhile to note here that if variable E (or E') is viewed as antecedent to variables S and M, as in Figure 1(a), then we could conclude from Figure 1(a) that the observed relationship between S and M has been explained away by variable E (or E'); however, we could also conclude from Figure 1(a) that the observed relationship between S and M has been explained (rather than explained away) by variable E (or E') – that is, by the relationship between variable E (or E') and S and the relationship between variable E (or E') and M. If variable E (or E') is viewed as intervening between variables S and M, as in Figure 1(b), or if variable E (or E') is viewed as coincident or reciprocal with variable S and as antecedent to variable M, as in Figure 1(c), then we could also conclude from Figure 1(b) or Figure 1(c) that the observed relationship between S and M can be explained (rather than explained away) by variable E (or E').

For the cross-classified data in the 6×4 table (Table 1) considered here, we present in Table 2 the chi-square values and the corresponding

Table 2. Models Applied to the Cross-Classified Data in Table 1

Model	No. of Latent Classes	Degrees of Freedom	Chi-Square	
			Goodness of Fit	Likelihood Ratio
Independence H_0	1	15	45.99	47.42
Latent class H_1	2	8	2.74	2.75