



Recursive Path Models when Both Predictor and Response Variables are Categorical

P. V. Rao, *Department of Pediatrics, University of Florida,
Gainesville, FL 32610, USA. Email: pejavervrao@gmail.com*

Haihong Li, *Vertex Pharmaceuticals, 130, Waverly St, Cambridge,
MA 02139, USA. Email: haihong_li@vrtx.com*

Jeffrey Roth, *Department of Pediatrics, University of Florida,
Gainesville, FL 32610, USA. Email: rothj@peds.ufl.edu*

Received: June 2008 Revised: August 2008

Abstract

Recursive path analysis is a useful tool for inference on a sequence of three or more response variables in which the causal effects of variables, if any, are in one direction. The primary objective in such analysis is to decompose the total effect of each variable into its direct and indirect components. Methods for recursive analysis of a chain of continuous variables are well developed but there is a lack of uniform methodology when the variables are categorical. In this paper we propose an approach for categorical response variables that is based on generalized linear models. The proposed method has the flexibility of allowing the use of common categorical data models such as the Poisson, Probit and logistic regression models, along with definitions of direct and indirect effects in terms of relative risks and odds ratios. The method can be implemented easily using standard statistical software such as the GENMOD procedure of SAS. This proposed method is illustrated using real data.

AMS Subject Classification: 62J12; 62F10.

Key-words: Categorical data; Direct effect; Generalized linear models; Indirect effect; Odds ratio; Path analysis; Relative risk; Total effect.

1. Introduction

Structural equation modeling and recursive path models play a key role in causal analysis of a sequence of outcome variables (Maruyama, 1998, Chapter 1). In the terminology of structural equation modeling, recursive path models refer to the case where all causal effects are in a single direction in that the outcome variable is causally affected by a chain of

more than one predictor variable, each of which may be affected by any of the preceding variables in the chain. For instance, when modeling the effects of mother's health status during pregnancy (Y_1) and child's low birth weight (Y_2) on the likelihood of developmental delay or disability during the first three years of the child's life (Y_3), we have an outcome variable Y_3 which may be affected by the chain of predictor variables Y_1, Y_2 . Clearly, because the effects in the chain Y_1, Y_2, Y_3 are in a single direction – Y_2 cannot affect Y_1 and Y_3 cannot affect Y_2 or Y_1 – a recursive model is appropriate in this case.

The primary objective in analysis of causal effects is to decompose the total effect of each variable into its direct and indirect components (Hoyle, 1995). In the above example Y_1 can have a direct effect on Y_3 and there can be an indirect effect through the intermediate variable Y_2 . When all variables in a chain are continuous variables, structural equation modeling (SEM) methods for recursive path analysis are well developed. As noted by Pearl (2001), SEM methods of path analysis have been mostly restricted to linear analysis. Assuming a hierarchical structural model in which the expected responses are expressed as linear combinations of the predictor variables, these methods utilize the Ordinary Least Squares (OLS) techniques for estimating and testing the direct and indirect effects. For a review of recursive path analysis with continuous predictor variables, see Wright (1934, 1960), Duncan (1966), Land (1969), Goldberger (1972), and Asher (1976, chapter 3).

Path analysis with categorical predictor variables has been applied in social, econometric, psychometric, medical and epidemiological research. However, in spite of a wide variety of applications of recursive path analysis and recent work extending its application to categorical predictor variables (see Section 2), the use of these methods for practical data analysis has suffered due to a lack of simple methodology for estimating and interpreting direct and indirect effects of a chain of categorical outcome variables.

The objective in this paper is to suggest a method of recursive path analysis that simplifies the estimation and interpretation of direct and indirect effects of categorical outcome variables. The proposed method uses generalized linear models on a sequence of categorical response variables that are dependent on a set of independent variables. The method is unified in the sense that it has the flexibility of allowing the use of common categorical data models such as the Poisson, probit and logistic regression models, along with definitions of direct and indirect effects in terms of well understood common risk measures such as relative risks and odds ratios. The method can be implemented easily using standard generalized linear model software such as the GENMOD procedure of SAS.

This paper does not address the underlying theoretical assumptions and/or conditions that are logically necessary to support causal interpretation of statistical conclusions derived from path analytic methods. References cited in Section 2 point to some important conceptual papers in this active area of current research. These authors specify conditions under which the causal effects are identifiable. The aim of the present paper is more pragmatic: to demonstrate to researchers a methodology for decomposing the total effect of a categorical response variable by utilizing some well-known properties of the generalized linear model.

OLS-based recursive path analysis methods are not suitable for categorical response variables for two reasons. First, such methods do not work well because of the problems associated with modeling expected responses as unrestricted linear combinations of the predictor variables. Second, the simple products of regression coefficients used to measure causal relationships between continuous predictors are essentially meaningless in the case of categorical variables because these coefficients have different interpretations in the continuous

and discrete cases. In the continuous case, the regression coefficients are used to measure the changes in the mean response as a function of the changes in the predictor variables, whereas in the categorical case they represent the changes in the probabilities of responses. We extend OLS methods to analysis of categorical responses by modeling the expected responses using generalized linear models (GLM) as described by McCullagh and Nelder (1989, Section 2.2).

The paper is organized as follows. In Section 2, we present a brief survey of earlier work on recursive analysis of categorical variables. After introducing the proposed recursive model for binary categorical variables in Section 3, we define the direct and indirect effects for these models and discuss their properties in Section 4. Section 5 shows how the binary model can be generalized to multinomial responses while Section 6 contains a discussion of how the proposed model can be fitted using the GENMOD procedure of SAS. Section 7 presents an example to illustrate the method using longitudinal data on early childhood diagnosis of developmental delay or disability as a function of health and demographic variables. Two appendices provide mathematical details of the likelihood function and the calculation of standard errors.

2. Background

Some early work on causal inference for categorical variables is by Goodman (1973a, b). Using a combination of logistic and loglinear models, he proposed methods for estimation and testing of the model parameters for models incorporating a priori knowledge of the causal relationship. See Hagenaars (1990, Chapter 2), Hagenaars (1993, chapter 2) and Hagenaars (1994, 1998, 2002) for a modification of the Goodman approach that leads to a methodology similar to the well known LISREL model for continuous variables (Joreskog and Sorbom, 1989). Vermunt (1996) provided more details on the results of Hagenaars. None of these papers address the question of defining, estimating and interpreting direct and indirect effects of categorical predictors. Wilson and Bielby (1983) developed a methodology for analyzing quasi-linear recursive models – models that are functions of linear combinations of predictor variables – containing both categorical and continuous variables as either exogenous or endogenous variables. For the case where all the variables are discrete, they showed how the direct and indirect effects can be estimated as logarithms of appropriate odds ratios. The methodology described by these authors extended the method of path analysis of recursive models with continuous variables to models with both continuous and discrete variables. However, their methodology can be quite complex, both in terms of computation and interpretation, even in such simple but commonly occurring situations involving binary endogenous variables and categorical or continuous exogenous variables. Also, their approach of modeling probabilities as linear combinations of variables is problematic because, in general, such practices lead to misspecified models (Agresti, 1990, Section 4.2.1).

Johnson (2001, Section 6.4) considered the problem of extending the well known method of “Calculus of Coefficients (COC)” for analyzing recursive chains of continuous variables to the analysis of recursive systems with categorical response variables and/or nonlinear relationships between response and predictor variables. By defining suitable measures of direct and indirect effects, she developed a method of analysis called the “Calculus of Effects”

that is an extension of COC for continuous endogenous variables to categorical variables. Johnson (2001, Section 7.1) did mention the possibility of measuring effects using relative risks and odds ratios, but did not address the problem in the setting of generalized linear models. Eshima et al. (2001) provided a method of analyzing recursive systems of categorical endogenous variables in which the structural relationship between the variables is specified by a logistic regression model without interaction terms. Despite the fact that their method provided a reasonable definition of the direct and indirect effects, the method suffers from the fact that it is restricted to logistic regression models and does not allow for the possibility that there may be exogenous variables that must be included in the model. As we shall see in Section 5, their definitions of the direct and indirect effects under the logistic model can be considered as approximations of the effect measures proposed in this paper.

Recently, there have been several attempts (Robins and Greenland, 1992; Pearl, 1995; Spirtes et al., 2000, Chapter 3) to investigate conditions necessary for causal interpretation of statistical conclusions and extend the linear model causal analysis techniques to nonlinear and nonparametric models. Robins and Greenland (1992) developed a theory of causal analysis assuming the existence of certain counterfactual (potential) random variables, while Spirtes et al. (2000, Chapter 7) and Pearl (1995) used directed acyclic graphs (DAGs) to do the same. The causal analysis techniques developed by these authors are applicable to continuous and categorical outcomes in non-linear and nonparametric models. However, unlike the counterfactual theory of Robins and Greenland, the DAG based theories directly incorporate path-analytic techniques for defining and estimating direct and indirect effects. See Robins (2003), Van der Laan and Petersen (2004), Rubin (2004), Pearl (2000, Chapter 7) for comparisons of assumptions, implications and applicability of the above approaches.

3. The Model for Binary Outcomes

Suppose there is an ordered sequence of m causally ordered binary response (endogenous) variables Y_1, Y_2, \dots, Y_m , each with possible values 0 and 1, and a vector of k independent (exogenous) variables:

$$\mathbf{Z} = (Z_1, Z_2, \dots, Z_k).$$

The assumption of causal ordering of the endogenous variables implies that Y_i may have a causal effect on Y_j if and only if $i < j$.

Let $Y_0 = 0$ and define the vectors of variables antecedent and subsequent to Y_i in the causal chain as

$$\mathbf{Y}_{Ai}^T = (Y_0, Y_1, \dots, Y_{i-1}), \quad i = 1, \dots, m.$$

$$\mathbf{Y}_{Si}^T = (Y_{i+1}, \dots, Y_m), \quad i = 1, \dots, m-1.$$

Then the conditional expectation of Y_i given $\mathbf{Y}_{Ai} = \mathbf{y}_{Ai}$, $\mathbf{Y}_{Si} = \mathbf{y}_{Si}$ and $\mathbf{Z} = \mathbf{z}$ equals the conditional probability of $Y_i = 1$ given $\mathbf{Y}_{Ai} = \mathbf{y}_{Ai}$, $\mathbf{Y}_{Si} = \mathbf{y}_{Si}$ and $\mathbf{Z} = \mathbf{z}$. That is,

$$E(Y_i | \mathbf{y}_{Ai}, \mathbf{y}_{Si}, \mathbf{z}) = P(Y_i = 1 | \mathbf{Y}_{Ai} = \mathbf{y}_{Ai}, \mathbf{Y}_{Si} = \mathbf{y}_{Si}, \mathbf{Z} = \mathbf{z}), \quad i = 1, \dots, m.$$

The assumption that the effects of the Y_j are ordered implies that P_i , the conditional expectation of Y_i is a function of \mathbf{y}_{Ai} and \mathbf{z} alone. That is, the variables Y_i satisfy the structural

equations:

$$P_i = \phi_i(\mathbf{y}_{Ai}, \mathbf{z}), \quad i = 1, \dots, m, \tag{3.1}$$

for some appropriately selected functions ϕ_i .

Two points concerning model (3.1) are worth noting. First, the key assumption in this model is the assumption of causal ordering which implies that for $i < j$, Y_i can cause Y_j but not vice versa. In practical applications, this assumption is often justified on the basis that Y_i and Y_j are characteristics of events occurring in sequence over time with Y_i referring to the earlier event. Second, the causal chain implied by model (3.1) can be reversed using Bayes Theorem:

$$P(\mathbf{Y}_{Ai} = \mathbf{y}_{Ai} | Y_i, \mathbf{Z}) = \frac{P(Y_i | \mathbf{Y}_{Ai} = \mathbf{y}_{Ai}, \mathbf{Z}) P(\mathbf{Y}_{Ai} = \mathbf{y}_{Ai} | \mathbf{Z})}{\sum_{\alpha} P(Y_i | \mathbf{Y}_{Ai} = \mathbf{y}_{\alpha}, \mathbf{Z}) P(\mathbf{Y}_{Ai} = \mathbf{y}_{\alpha} | \mathbf{Z})},$$

where the summation in the denominator is over the 2^{i-1} possible values of \mathbf{y}_{α} .

In this paper we focus on situations in which the model in (3.1) is a generalized linear model. That is, we assume that for each i , $i = 1, \dots, m$, there exists a linear combination, lp_i , of the covariate values \mathbf{y}_{Ai} and \mathbf{z} such that (3.1) can be expressed as

$$P_i = g^{-1}(lp_i), \quad i = 1, \dots, m, \tag{3.2}$$

where $g(\cdot)$ is an invertible real-valued function defined on $(0, 1)$. McCullagh and Nelder (1989, page 27) refer to lp_i and $g(\cdot)$ as the linear predictor and the link function, respectively. Some of the most commonly used link functions are $g(p) = \log p$ for loglinear regression, $g(p) = \log(\frac{p}{1-p})$ for logistic regression and $g(p) = \Phi^{-1}(p)$, where Φ is the standard normal distribution function, for probit regression.

In this paper we shall restrict our attention to simple linear predictors of the form

$$\begin{aligned} lp_i &= \beta_{i0} + \beta_{i1}Y_1 + \dots + \beta_{i,i-1}Y_{i-1} + \gamma_{i1}Z_1 + \dots + \gamma_{ik}Z_k \\ &= \mathbf{Y}_{Ai}^T \boldsymbol{\beta}_i + \mathbf{Z}^T \boldsymbol{\gamma}_i, \end{aligned} \tag{3.3}$$

where

$$\begin{aligned} \boldsymbol{\beta}_i^T &= (\beta_{i0}, \dots, \beta_{i,i-1}), \\ \boldsymbol{\gamma}_i^T &= (\gamma_{i,1}, \dots, \gamma_{i,k}), \quad i = 1, \dots, m. \end{aligned}$$

Results for more complex linear predictors containing higher order and/or interaction terms, though somewhat more complicated, may be derived in a straightforward manner from the results for the linear predictor (3.3).

4. Definitions of Effects of Binary Outcomes

In a causally ordered sequence each variable can have a direct effect as well as an indirect effect on any subsequent variable, with their sum being the total effect. We begin with the definition of the total effect.

The **total effect** (TE) of Y_i on Y_j , $i < j$, when \mathbf{Y}_{A_i} and \mathbf{Z} are fixed at \mathbf{y}_{A_i} and \mathbf{z} , respectively, is the change, measured on the scale of the link function $g(\cdot)$, in the conditional probability of $Y_j = 1$ resulting from changing Y_i from 0 to 1. Thus the total effect of Y_i on Y_j is a comparison of the conditional probability of $Y_j = 1$ given $Y_i = 1$ with the conditional probability given $Y_i = 0$. The total effect may be expressed as

$$TE_{ij}(\mathbf{y}_{A_i}, \mathbf{z}) = g(P(Y_j = 1 | \mathbf{Y}_{A_i} = \mathbf{y}_{A_i}, Y_i = 1, \mathbf{Z} = \mathbf{z})) - g(P(Y_j = 1 | \mathbf{Y}_{A_i} = \mathbf{y}_{A_i}, Y_i = 0, \mathbf{Z} = \mathbf{z})). \quad (4.1)$$

The total effect of Y_i on Y_j depends only on the values of the antecedent and exogenous variables whereas the direct and indirect effects will also depend on the values of the variables intermediate to Y_i and Y_j . For $1 \leq i+1 \leq j-1 \leq m$, let

$$\mathbf{Y}_{I_{ij}} = (Y_{i+1}, \dots, Y_{j-1})$$

denote the sequence of endogenous variables intermediate to Y_i and Y_j .

The **direct effect** (DE) of Y_i on Y_j when \mathbf{Y}_{A_i} , $\mathbf{Y}_{I_{ij}}$ and \mathbf{Z} are fixed at \mathbf{y}_{A_i} , $\mathbf{y}_{I_{ij}}$ and \mathbf{z} , respectively, is the change, measured on the scale of the link function $g(\cdot)$, in the conditional probability of $Y_j = 1$ resulting from changing Y_i from 0 to 1. That is,

$$DE_{ij}(\mathbf{y}_{A_i}, \mathbf{y}_{I_{ij}}, \mathbf{z}) = g(P(Y_j = 1 | \mathbf{Y}_{A_i} = \mathbf{y}_{A_i}, Y_i = 1, \mathbf{Y}_{I_{ij}} = \mathbf{y}_{I_{ij}}, \mathbf{Z} = \mathbf{z})) - g(P(Y_j = 1 | \mathbf{Y}_{A_i} = \mathbf{y}_{A_i}, Y_i = 0, \mathbf{Y}_{I_{ij}} = \mathbf{y}_{I_{ij}}, \mathbf{Z} = \mathbf{z})). \quad (4.2)$$

The direct effect of an outcome variable Y_i on a subsequent outcome variable Y_j defined in (4.2) is a comparison of an appropriately chosen function of the conditional expectation of Y_j when Y_i is held fixed at the reference value 0 with the same function of the conditional expectation of Y_j when Y_i is fixed at the value of interest 1.

The **indirect effect** (IE) of Y_i on Y_j when \mathbf{Y}_{A_i} , $\mathbf{Y}_{I_{ij}}$ and \mathbf{Z} are fixed at \mathbf{y}_{A_i} , $\mathbf{y}_{I_{ij}}$ and \mathbf{z} , respectively, is the difference between the corresponding total and direct effects. That is,

$$IE_{ij}(\mathbf{y}_{A_i}, \mathbf{y}_{I_{ij}}, \mathbf{z}) = TE_{ij}(\mathbf{y}_{A_i}, \mathbf{z}) - DE_{ij}(\mathbf{y}_{A_i}, \mathbf{y}_{I_{ij}}, \mathbf{z}). \quad (4.3)$$

Three important consequences of the above definitions are worth noting. First, in the frequently used special case where the linear predictor has the form (3.3), the direct effect of Y_i on Y_j is

$$\begin{aligned} DE_{ij}(\mathbf{y}_{A_i}, \mathbf{y}_{I_{ij}}, \mathbf{z}) &= lp_j(\mathbf{y}_{A_i}, 1, \mathbf{y}_{I_{ij}}, \mathbf{z}) - lp_j(\mathbf{y}_{A_i}, 0, \mathbf{y}_{I_{ij}}, \mathbf{z}) \\ &= \beta_{ji}, \end{aligned}$$

a result consistent with that found in path analysis (Duncan 1975, page 31) of a sequence of continuous endogenous variables.

Second, the definitions are consistent with the practice of comparing risks using log relative risk in log-linear models and log odds ratio in logistic models. For example, for log-linear models the link function is $g(p) = \log p$ and the expressions for direct and total effects reduce to

$$DE_{ij}(\mathbf{y}_{A_i}, \mathbf{y}_{I_{ij}}, \mathbf{z}) = \log \left\{ \frac{P(Y_j = 1 | \mathbf{Y}_{A_i} = \mathbf{y}_{A_i}, Y_i = 1, \mathbf{Y}_{I_{ij}} = \mathbf{y}_{I_{ij}}, \mathbf{Z} = \mathbf{z})}{P(Y_j = 1 | \mathbf{Y}_{A_i} = \mathbf{y}_{A_i}, Y_i = 0, \mathbf{Y}_{I_{ij}} = \mathbf{y}_{I_{ij}}, \mathbf{Z} = \mathbf{z})} \right\},$$

$$TE_{ij}(\mathbf{y}_{Ai}, \mathbf{z}) = \log \left\{ \frac{P(Y_j = 1 | \mathbf{Y}_{Ai} = \mathbf{y}_{Ai}, Y_i = 1, \mathbf{Z} = \mathbf{z})}{P(Y_j = 1 | \mathbf{Y}_{Ai} = \mathbf{y}_{Ai}, Y_i = 0, \mathbf{Z} = \mathbf{z})} \right\}.$$

Thus, for log-linear models, the direct and total effects are defined as the logs of the direct and total conditional relative risks of $(Y_j = 1)$ given $(Y_i = 1)$ relative to $(Y_j = 1)$ given $(Y_i = 0)$. Consequently, the corresponding conditional relative risks are obtained by exponentiating the effects.

Third, because the direct and indirect effects of Y_i on Y_j are functions of $\mathbf{y}_{Ai}, \mathbf{y}_{Iij}$ and \mathbf{z} , one may desire a single overall measure of each of these effects. Such measures can be obtained by calculating appropriately weighted averages over the possible values of $(\mathbf{y}_{Ai}, \mathbf{y}_{Iij}, \mathbf{z})$. Thus, for example, if $w(\mathbf{y}_{Ai}, \mathbf{y}_{Iij}, \mathbf{z})$ are nonnegative weights such that

$$\sum w(\mathbf{y}_{Ai}, \mathbf{y}_{Iij}, \mathbf{z}) = 1,$$

where the summation is over all possible values of $(\mathbf{y}_{Ai}, \mathbf{y}_{Iij}, \mathbf{z})$, the average or adjusted direct effect may be defined as

$$DE_{ij} = \sum w(\mathbf{y}_{Ai}, \mathbf{y}_{Iij}, \mathbf{z}) DE_{ij}(\mathbf{y}_{Ai}, \mathbf{y}_{Iij}, \mathbf{z}). \tag{4.4}$$

See Section 7 for an example.

5. Extension to Multinomial Outcomes

The binary model in Section 3 and the corresponding definitions of the direct and indirect effects can be extended to multinomial outcomes in a straightforward manner. Suppose for $i = 1, \dots, m$, the i th response variable, Y_i , is multinomial with L_i levels and let \mathbf{Y}_i be an $(L_i - 1)$ dimensional vector of dummy variables in which the j th element equals 1 if the response level is j and 0 otherwise. We generalize the model (3.2) by assuming

$$\mathbf{P}_i = g^{-1}(\mathbf{l}\mathbf{p}_i), \quad i = 1, \dots, m, \tag{5.1}$$

where \mathbf{P}_i is the $(L_i - 1)$ dimensional vector of probabilities for \mathbf{Y}_i , $\mathbf{l}\mathbf{p}_i$ is a $(L_i - 1)$ dimensional vector of linear combinations:

$$\mathbf{l}\mathbf{p}_i = \boldsymbol{\beta}_{i0} + \boldsymbol{\beta}_{i1} \mathbf{Y}_1 + \dots + \boldsymbol{\beta}_{i,i-1} \mathbf{Y}_{i-1} + \boldsymbol{\gamma}_i \mathbf{Z},$$

such that

- $\boldsymbol{\beta}_{i0}$ is a $(L_i - 1)$ dimensional vector of intercepts.
- For $j \geq 1$, $\boldsymbol{\beta}_{ij}$ is a $(L_i - 1) \times (L_j - 1)$ matrix of coefficients.
- $\boldsymbol{\gamma}_i \mathbf{Z}$ is the term for exogenous variables,

and $g^{-1}(\mathbf{l}\mathbf{p}_i)$ is the $L_i - 1$ dimensional vector resulting from applying g^{-1} on the vector $\mathbf{l}\mathbf{p}_i$ componentwise. The effect of Y_i on Y_j may be defined as in Section 4 except that the single effect in the binary case will be replaced by an $(L_i - 1) \times (L_j - 1)$ effect matrix. For example, the (k, l) element of the matrix of direct effects of Y_1 on Y_m when Y_2, \dots, Y_{m-1} are fixed at y_2, \dots, y_{m-1} is the change, measured on the scale of the link function $g(\cdot)$, in the

conditional probability of $Y_j = l$ resulting from changing Y_i from L_1 to k . In symbols,

$$DE(Y_m, Y_1)_{kl} = g(P(Y_m = l | Y_1 = k; Y_2 = y_2, \dots, Y_{m-1} = y_{m-1})) \\ - g(Pr\{Y_m = l | Y_1 = L_1; Y_2 = y_2, \dots, Y_{m-1} = y_{m-1}\}).$$

As in the case of binary outcomes, the matrix of direct effects of Y_1 on Y_m under model (5.1) is

$$DE(Y_1, Y_m) = \boldsymbol{\beta}_{m1}.$$

Estimation and interpretation of the effects in the binary case have direct analogues in the multinomial case. However, in the following sections, we concentrate on the binary case.

6. Statistical Inference

The likelihood function for fitting the generalized linear model (3.2) is shown in Appendix A. Since $\boldsymbol{\beta}_i$ and $\boldsymbol{\gamma}_i$ appear only in the i th regression equation, their point estimates can be obtained by fitting the model for P_i using standard statistical software such as the GENMOD procedure in SAS. The effect estimates are calculated by replacing the regression coefficients with their estimated values in the functional forms of the effects. For instance, in the log-linear model, the total effect of Y_1 on Y_m at $\mathbf{Z} = \mathbf{z}$ is

$$TE_{1m}(\mathbf{z}) = \log \left\{ \frac{\{P(Y_m = 1 | Y_1 = 1, \mathbf{Z} = \mathbf{z})\}}{\{P(Y_m = 1 | Y_1 = 0, \mathbf{Z} = \mathbf{z})\}} \right\} \\ = \log \left\{ \frac{E\{P(Y_m = 1 | Y_1 = 1, Y_2, \dots, Y_{m-1}, \mathbf{Z} = \mathbf{z})\}}{E\{P(Y_m = 1 | Y_1 = 0, Y_2, \dots, Y_{m-1}, \mathbf{Z} = \mathbf{z})\}} \right\}, \quad (6.1)$$

where the expectations in the numerator and denominator are with respect to the conditional distribution of (Y_2, \dots, Y_{m-1}) given $Y_1 = 1, \mathbf{Z} = \mathbf{z}$ and $Y_1 = 0, \mathbf{Z} = \mathbf{z}$, respectively. For $m = 3$ and simple linear predictor of the form (3.3), the expression in (6.1) reduces to:

$$TE_{13}(\mathbf{z}) = \log \left[\{P(Y_3 = 1 | Y_1 = 1, Y_2 = 1, \mathbf{Z} = \mathbf{z})P(Y_2 = 1 | Y_1 = 1, \mathbf{Z} = \mathbf{z}) \right. \\ \left. + P(Y_3 = 1 | Y_1 = 1, Y_2 = 0, \mathbf{Z} = \mathbf{z})P(Y_2 = 0 | Y_1 = 1, \mathbf{Z} = \mathbf{z})\} / \right. \\ \left. \{P(Y_3 = 1 | Y_1 = 0, Y_2 = 1, \mathbf{Z} = \mathbf{z})P(Y_2 = 1 | Y_1 = 0, \mathbf{Z} = \mathbf{z}) \right. \\ \left. + P(Y_3 = 1 | Y_1 = 0, Y_2 = 0, \mathbf{Z} = \mathbf{z})P(Y_2 = 0 | Y_1 = 0, \mathbf{Z} = \mathbf{z})\} \right] \\ = \beta_{31} + \log \left\{ \frac{1 + \exp\{\beta_{20} + \beta_{21} + \mathbf{z}^T \boldsymbol{\gamma}_2\} (\exp\{\beta_{32}\} - 1)}{1 + \exp\{\beta_{20} + \mathbf{z}^T \boldsymbol{\gamma}_2\} (\exp\{\beta_{32}\} - 1)} \right\}. \quad (6.2)$$

If $\hat{\beta}_{20}, \hat{\beta}_{21}, \hat{\beta}_{32}$ and $\hat{\boldsymbol{\gamma}}_2$ are the estimates of the regression parameters, the estimated total effect is

$$\widehat{TE}_{13}(\mathbf{z}) = \hat{\beta}_{31} + \log \left\{ \frac{1 + \exp\{\hat{\beta}_{20} + \hat{\beta}_{21} + \mathbf{z}^T \hat{\boldsymbol{\gamma}}_2\} (\exp\{\hat{\beta}_{32}\} - 1)}{1 + \exp\{\hat{\beta}_{20} + \mathbf{z}^T \hat{\boldsymbol{\gamma}}_2\} (\exp\{\hat{\beta}_{32}\} - 1)} \right\}. \quad (6.3)$$

Again, the fact that the parameters in the i th regression equation do not appear in any other equations implies that the information matrix is block diagonal. Thus the likelihood estimates of regression coefficients appearing in different regression equations are uncorrelated. We can then use the functional forms of effects and the delta method to obtain the standard errors of the effect estimates.

When the number of variables is small, the total, direct, and indirect effects are easily expressed as functions of regression coefficients by calculating the required conditional probabilities directly. However, when the number of endogenous variables is large, the expression for the total effect can be quite complicated because the conditional expectations in (6.1) are sums of 2^{m-2} terms and each term needs to be evaluated separately.

One way around this difficulty is to approximate the expected values of the conditional probabilities in (6.1) with conditional probabilities evaluated at the expected values of the random variables that are being integrated. Thus an approximation to the total effect (6.2) may be computed as

$$TE_{13}(\mathbf{z}) = \log \left\{ \frac{E\{P(Y_3 = 1|Y_1 = 1, Y_2, \mathbf{Z} = \mathbf{z})\}}{E\{P(Y_3 = 1|Y_1 = 0, Y_2, \mathbf{Z} = \mathbf{z})\}} \right\} \\ \approx \log \left\{ \frac{\{P(Y_3 = 1|Y_1 = 1, Y_2 = E(Y_2|Y_1 = 1, Z = z), \mathbf{Z} = \mathbf{z})\}}{\{P(Y_3 = 1|Y_1 = 0, Y_2 = E(Y_2|Y_1 = 0, Z = z), \mathbf{Z} = \mathbf{z})\}} \right\}. \quad (6.4)$$

The approximate total effect calculated as in (6.4) can be interpreted as the total effect wherein the intermediate variables are fixed at their mean values. Indeed, this approximate total effect is the same as the total effect proposed by Eshima, Tabata, and Zhi (2001) for the logistic model. Another approximation is to use the sample conditional proportions instead of the model-based conditional expectations when evaluating $E(Y_i|Y_1, Z = z)$.

7. An Example

In this section, a subset of maternal and child health data for the State of Florida linked from various sources by the Maternal and Child Health and Education Research and Data Center (MCHERDC) at the University of Florida will be used to illustrate the calculation and interpretation of the effects defined earlier. For the purpose of this example, let us suppose that we are interested in studying the causal effects of (1) diabetes and/or hypertension (DH) during mother’s pregnancy and (2) low birth weight of the child (LBW) on the incidence of developmental delay or disability (DDD) in the first three years of the child’s life. Suppose further that we want to study these causal effects after adjusting for two exogenous variables: mother’s prenatal smoking (yes/no) and the child’s gender. A detailed analysis in which all relevant endogenous and exogenous variables are taken into consideration is being prepared for publication elsewhere. Since DH can have causal effects on LBW and DDD, and LBW can have a causal effect on DDD, the sequence : DH, LBW, DDD forms a recursive causal path. Consequently, the required causal analyses can be done using models of the form (3.2) with $m = 3$ and $k = 2$. Define

$$Y_1 = \begin{cases} 1 & \text{if there is diabetes or hypertension during pregnancy,} \\ 0 & \text{otherwise;} \end{cases}$$

$$\begin{aligned}
 Y_2 &= \begin{cases} 1 & \text{if birth weight is lower than 2500 grams,} \\ 0 & \text{otherwise;} \end{cases} \\
 Y_3 &= \begin{cases} 1 & \text{if DDD is present,} \\ 0 & \text{otherwise;} \end{cases} \\
 Z_1 &= \begin{cases} 1 & \text{if the mother smoked during pregnancy,} \\ 0 & \text{otherwise;} \end{cases} \\
 Z_2 &= \begin{cases} 1 & \text{if the child is male,} \\ 0 & \text{otherwise.} \end{cases}
 \end{aligned}$$

As in (3.2), let P_i be the conditional probability of $(Y_i = 1)$ for $i = 1, 2, 3$. Under loglinear models for the P_i , the structural equations for estimating the regression parameters can be expressed as

$$\begin{aligned}
 \log P_1 &= \beta_{10} + \gamma_{11}z_1 + \gamma_{12}z_2, \\
 \log P_2 &= \beta_{20} + \beta_{21}y_1 + \gamma_{21}z_1 + \gamma_{22}z_2, \\
 \log P_3 &= \beta_{30} + \beta_{31}y_1 + \beta_{32}y_2 + \gamma_{31}z_1 + \gamma_{32}z_2.
 \end{aligned} \tag{7.1}$$

Assuming Poisson distribution for the event $(Y_i = 1)$, $i = 2, 3$, we used the GENMOD procedure of SAS for fitting loglinear models for P_2 and P_3 . The fitted models are

$$\begin{aligned}
 \log P_2 &= -2.6060 + 0.7919y_1 + 0.4173z_1 - 0.1929z_2, \\
 \log P_3 &= -0.5694 + 0.0147y_1 + 0.1167y_2 - 0.0952z_1 + 0.0219z_2.
 \end{aligned}$$

As noted earlier, the direct effects of Y_1 (DH) and Y_2 (LBW) on Y_3 (DDD) are β_{31} and β_{32} , respectively. Thus the estimated direct effect of DH on DDD is 0.0147 and that of LBW on DDD is 0.1167. Also since the direct effects in loglinear models are the logs of the relative risks, the relative risk of DDD for a child whose mother had DH relative to a child whose mother did not have DH is $\exp\{0.0147\} = 1.0148$. In other words, as the result of the direct effect of DH on DDD, a child with a DH mother is 1.0148 times as likely to have DDD as a child with a mother without DH. The corresponding relative risk that measures the direct effect of LBW on DDD is $\exp\{0.1167\} = 1.1238$.

Table 1 shows the standard errors of the regression coefficients. These standard errors can be used to construct confidence intervals for the direct effects. For example, a 95% confidence interval for the direct effect of DH on DDD is (0.0002, 0.0292). A 95% confidence interval for the corresponding relative risk is $(\exp\{0.0002\}, \exp\{0.0292\}) = (1.0002, 1.0296)$. Equation (6.3) can be used to calculate the estimated total effect for each combination of the values of z_1 and z_2 . The corresponding indirect effects are then obtained by subtraction. The estimated values of the direct, total and indirect effects of DH on DDD are given in Table 2. This table also shows the sample proportion in the four smoking-gender categories.

Since the total and indirect effects vary with population strata, one would want to summarize the strata-specific pairs of indirect and total effects using a single pair of values. This can be done by taking a weighted average of the strata-specific effects. The weights can be

Table 1. Parameter estimates and standard errors using loglinear models.

effect	estimate	SE
β_{20}	-2.6060	0.0078
β_{21}	0.7919	0.0147
γ_{21}	0.4173	0.0123
γ_{22}	-0.1929	0.0102
β_{30}	-0.5694	0.0030
β_{31}	0.0147	0.0074
β_{32}	0.1167	0.0068
γ_{31}	-0.0952	0.0054
γ_{32}	0.0219	0.0038

Table 2. The direct, indirect and total effects (in log scale) of pregnancy diabetes or hypertension (DH) on early developmental delay or disability (DDD) in different smoking by gender categories.

smoking	gender	effects			sample proportion
		direct	indirect	total	
Yes	Male	0.0147	0.0137	0.0283	0.0832
Yes	Female	0.0147	0.0164	0.0311	0.0756
No	Male	0.0147	0.0090	0.0237	0.4368
No	Female	0.0147	0.0109	0.0256	0.4044

taken as the sample proportion, which reflects the true population mix; or any weights that represent the population mix in an idealized population. For example, we can use equal weights to adjust for the effects of smoking and gender, or we can set the proportion of smoking at a certain percentage to estimate the reduction in the DDD rate if the rate of prenatal smoking is decreased by a certain amount through smoking cessation programs. Using the sample proportions as the weights for combining strata-specific effects we get the values 0.0107 and 0.0254 as the indirect and total effects of DH on DDD. The standard errors of the estimated total and indirect effects can be computed using the estimated covariance matrix of the regression coefficients along with the delta method applied to the function of regression coefficients in (6.3). The details of the computation are given in Appendix B. The resulting standard errors of the estimated total and indirect effects are 0.0074 and 0.0007, respectively. Therefore, the 95% confidence interval for the indirect and total effects of DH on DDD are (0.0093, 0.0121), and (0.0109, 0.0399), respectively. We have already seen that the 95% confidence interval for the direct effect of DH on DDD is (0.0002, 0.0292). On the scale of relative risks, the direct, indirect, and total effects of DH on DDD are 1.0148, 1.0108, and 1.0257, respectively, with the corresponding confidence intervals (1.0002, 1.0296), (1.0094, 1.0121), and (1.0110, 1.0407).

Thus, we may conclude that (1) mother’s prenatal hypertension or diabetes has significant ($p < .05$) direct and indirect (due to low birth weight) effects on child’s DDD, and (2)

the children of mothers with prenatal hypertension or diabetes are 3% more likely to have a DDD problem than children of mothers without prenatal hypertension or diabetes.

Appendix

A. The Likelihood Function

To evaluate the conditional probability of Y_i given \mathbf{Y}_{A_i} and Z , we let $\{C_{i1}, \dots, C_{iJ_i}\}$ denote the partition determined by the different combinations of the observed values of \mathbf{Y}_{A_i} and Z . Let N_{ij} denote the total number of observations in C_{ij} and y_{ij} be the observed number of 1's for Y_i in C_{ij} . Let Y_{ij} denote the random variable corresponding to the realization y_{ij} . The likelihood function is

$$L(\boldsymbol{\beta}, \boldsymbol{\gamma}) = \prod_{i=1}^m \prod_{j=1}^{J_i} \Pr\{Y_{ij} = y_{ij}\}. \quad (\text{A.1})$$

Any reasonable assumption about the distribution of Y_{ij} will make $\Pr\{Y_{ij} = y_{ij}\}$ a function of $P_{ij} = \Pr\{Y_i = 1 | C_{ij}\}$. Denote $\Pr\{Y_{ij} = y_{ij}\} = h(P_{ij})$. Since $P_{ij} = g^{-1}(\mathbf{y}_{A_i,j}^T \boldsymbol{\beta}_i + \mathbf{z}_{ij}^T \boldsymbol{\gamma}_i)$, the log likelihood is

$$l(\boldsymbol{\beta}, \boldsymbol{\gamma}) = \sum_{i=1}^m \sum_{j=1}^{J_i} \log(h \circ g^{-1}(\mathbf{y}_{A_i,j}^T \boldsymbol{\beta}_i + \mathbf{z}_{ij}^T \boldsymbol{\gamma}_i)), \quad (\text{A.2})$$

where $\mathbf{y}_{A_i,j}$ and \mathbf{z}_{ij} are the observed values of \mathbf{Y}_{A_i} and \mathbf{Z} taking values that determine the ij th category C_{ij} . Since each parameter vector $(\boldsymbol{\beta}_i, \boldsymbol{\gamma}_i)$ appears only in the i -th summand in (A.2), it can be estimated from the i -th model equation only. The estimates can be obtained using standard statistical packages such as the GENMOD procedure in SAS.

In the special case when Y_i follows the binomial distribution and $g(p) = \log(p/(1-p))$ is the logistic link function, the probability of $(Y_{ij} = y_{ij})$ is proportional to

$$P_{ij}^{y_{ij}} [1 - P_{ij}]^{N_{ij} - y_{ij}},$$

and the log likelihood function is

$$l(\boldsymbol{\beta}, \boldsymbol{\gamma}) = \sum_{i=1}^m \sum_{j=1}^{J_i} \left\{ y_{ij} [\mathbf{y}_{A_i,j}^T \boldsymbol{\beta}_i + \mathbf{z}_{ij}^T \boldsymbol{\gamma}_i] - N_{ij} \log(1 + \exp(\mathbf{y}_{A_i,j}^T \boldsymbol{\beta}_i + \mathbf{z}_{ij}^T \boldsymbol{\gamma}_i)) \right\}.$$

If Y_i has a Poisson distribution and the link function $g(p) = \log p$ is log-linear, the probability of $(Y_{ij} = y_{ij})$ is proportional to

$$\exp(-N_{ij} P_{ij}) \cdot (N_{ij} P_{ij})^{y_{ij}},$$

and the log likelihood can be expressed as

$$l(\boldsymbol{\beta}, \boldsymbol{\gamma}) = \sum_{i=1}^m \sum_{j=1}^{J_i} \left\{ y_{ij} [\mathbf{y}_{A_i,j}^T \boldsymbol{\beta}_i + \mathbf{z}_{ij}^T \boldsymbol{\gamma}_i] - \exp(\mathbf{y}_{A_i,j}^T \boldsymbol{\beta}_i + \mathbf{z}_{ij}^T \boldsymbol{\gamma}_i) \right\}.$$

B. Calculating Standard Errors

From the SAS output, the covariance matrix of $(\beta_{20}, \beta_{21}, \gamma_{21}, \gamma_{22}, \beta_{31}, \beta_{32})^T$ is

$$\Sigma = \begin{pmatrix} 0.0000612 & -0.000031 & -0.000034 & -0.000049 & 0 & 0 \\ -0.000031 & 0.0002172 & 2.6978 \times 10^{-6} & -9.172 \times 10^{-7} & 0 & 0 \\ -0.000034 & 2.6978 \times 10^{-6} & 0.0001517 & -4.699 \times 10^{-7} & 0 & 0 \\ -0.000049 & -9.172 \times 10^{-7} & -4.699 \times 10^{-7} & 0.0001048 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0.0000551 & -4.371 \times 10^{-6} \\ 0 & 0 & 0 & 0 & -4.371 \times 10^{-6} & 0.0000459 \end{pmatrix},$$

and the partial derivatives of (6.2) with respect to $(\beta_{20}, \beta_{21}, \gamma_{21}, \gamma_{22}, \beta_{31}, \beta_{32})^T$ is

$$\nabla = (A, B, C, D, E, F)^T,$$

where

$$A = \frac{\exp(\beta_{20} + \mathbf{z}^T \boldsymbol{\gamma}_2)(\exp(\beta_{21}) - 1)(\exp(\beta_{32}) - 1)}{[1 + \exp(\beta_{20} + \beta_{21} + \mathbf{z}^T \boldsymbol{\gamma}_2)(\exp(\beta_{32}) - 1)][1 + \exp(\beta_{20} + \mathbf{z}^T \boldsymbol{\gamma}_2)(\exp(\beta_{32}) - 1)]}$$

$$B = \frac{\exp(\beta_{20} + \beta_{21} + \mathbf{z}^T \boldsymbol{\gamma}_2)(\exp(\beta_{32}) - 1)}{1 + \exp(\beta_{20} + \beta_{21} + \mathbf{z}^T \boldsymbol{\gamma}_2)(\exp(\beta_{32}) - 1)}$$

$$C = \frac{\exp(\beta_{20} + \mathbf{z}^T \boldsymbol{\gamma}_2)(\exp(\beta_{21}) - 1)(\exp(\beta_{32}) - 1)z_1}{[1 + \exp(\beta_{20} + \beta_{21} + \mathbf{z}^T \boldsymbol{\gamma}_2)(\exp(\beta_{32}) - 1)][1 + \exp(\beta_{20} + \mathbf{z}^T \boldsymbol{\gamma}_2)(\exp(\beta_{32}) - 1)]}$$

$$D = \frac{\exp(\beta_{20} + \mathbf{z}^T \boldsymbol{\gamma}_2)(\exp(\beta_{21}) - 1)(\exp(\beta_{32}) - 1)z_2}{[1 + \exp(\beta_{20} + \beta_{21} + \mathbf{z}^T \boldsymbol{\gamma}_2)(\exp(\beta_{32}) - 1)][1 + \exp(\beta_{20} + \mathbf{z}^T \boldsymbol{\gamma}_2)(\exp(\beta_{32}) - 1)]}$$

$$E = 1$$

$$F = \frac{\exp(\beta_{20} + \mathbf{z}^T \boldsymbol{\gamma}_2)(\exp(\beta_{21}) - 1)\exp(\beta_{32})}{[1 + \exp(\beta_{20} + \beta_{21} + \mathbf{z}^T \boldsymbol{\gamma}_2)(\exp(\beta_{32}) - 1)][1 + \exp(\beta_{20} + \mathbf{z}^T \boldsymbol{\gamma}_2)(\exp(\beta_{32}) - 1)]}$$

We calculate the variance $\nabla^T \Sigma \nabla$, evaluated at the estimated parameter values, for each set of values for the exogenous variables. Using the sample proportions in Table 2, the standard error for the TE is 0.0074. The standard error for the IE can be obtained similarly, with value 0.0007. Therefore the 95% confidence interval for the DE is (0.0002, 0.0292), the 95% confidence interval for the IE is (0.0093, 0.0121), and the 95% confidence interval for the TE is (0.0109, 0.0399).

References

Agresti, A., 1990. *Categorical Data Analysis*. John Wiley, New York.
 Asher, H. B., 1976. *Causal Modelling*. Sage Publications, Thousand Oaks.
 Duncan, O. D., 1966. Path analysis: sociology examples. *American Journal of Sociology*, 72, 1–16.
 Duncan, O. D., 1975. *Introduction to Structural Equation Models*. Academic Press, New York.
 Eshima, N., Tabata, M., Zhi, G., 2001. Path analysis with logistic regression models: effect analysis of fully recursive causal systems of categorical variables. *Journal of the Japanese Statistical Society*, 31, 1–14.
 Goldberger, A. S., 1972. Structural equation models in the social sciences. *Econometrika*, 40, 979–1001.
 Goodman, L. A., 1973a. The analysis of multidimensional contingency tables when some variables are posterior to others: a modified path analysis approach. *Biometrika*, 60, 179–192.
 Goodman, L. A., 1973b. Causal analysis of data from panel studies and other kinds of surveys. *American Journal of Sociology*, 78, 1135–1191.

- Hagenaars, J. A., 1990. *Categorical Longitudinal Data*. Sage Publications, New York.
- Hagenaars, J. A., 1993. *Loglinear Models with Latent Variables*. Sage Publications, New York.
- Hagenaars, J. A., 1994. Latent variables in log-linear models. In: Eye, A. V., Clogg, C. C. (editors), *Latent Variables Analysis*, pp. 329–352, Thousand Oaks. Sage Publications.
- Hagenaars, J. A., 1998. Categorical causal modeling: latent class analysis and directed log-linear models with latent variables. *Sociological Methods and Research*, 26, 436–486.
- Hagenaars, J. A., 2002. Directed loglinear modeling with latent variables: causal models for categorical data with nonsystematic and systematic measurement errors. In: Hagenaars, J. A., McCutcheon, A. L. (editors), *Applied Latent Class Analysis*, pp. 234–286, Cambridge. Cambridge University Press.
- Hoyle, R. H., 1995. Structural Equation Modeling Approach: Basic Concepts and Fundamental Issues. In: Hoyle, R. H. (editor), *Structural Equation Modeling, Concepts, Issues and Applications*, pp. 1–15. Sage Publications, Thousand Oaks.
- Johnson, P. L., 2001. *Nonlinear Path Models with Continuous or Dichotomous Variables*. PhD thesis, University of Florida.
- Joreskog, K. C., Sorbom, D., 1989. *LISREL 7 User's Reference Guide*. SPSS Inc., Chicago.
- Land, K. C., 1969. Principles of path analysis. *Sociological Methodology*, 1, 3–37.
- Maruyama, J. M., 1998. *Basics of Structural Equation Modeling*. Sage Publications, Thousand Oaks.
- McCullagh, P., Nelder, J. A., 1989. *Generalized Linear Models, 2nd Ed.* Chapman & Hall/CRC, Boca Raton.
- Pearl, J., 1995. Causal diagrams for empirical research. *Biometrika*, 82, 669–710.
- Pearl, J., 2000. *Causality: Models, Reasoning and Inference*. Cambridge University Press, New York.
- Pearl, J., 2001. Direct and indirect effects. In: Kaufmann, M. (editor), *Seventeenth Conference on Uncertainty in Artificial Intelligence*, pp. 411–420, San Francisco, CA.
- Robins, J. M., 2003. Semantics of causal dag models and the identification of direct and indirect effects. In: Green, P., Hjort, N. L., Richardson, S. (editors), *Highly Structured Stochastic Systems*, pp. 70–81. Oxford University Press, New York.
- Robins, J. M., Greenland, S., 1992. Identifiability and exchangeability for direct and indirect effects. *Epidemiology*, 3, 143–155.
- Rubin, D. B., 2004. Direct and indirect causal effects via potential outcomes. *Scandinavian Journal of Statistics*, 31, 161–171.
- Spirites, P. C., Glymour, C. N., Scheines, R., 2000. *Causation, Prediction and Search*. MIT Press, Boston MA, 2nd edition.
- Van der Laan, M. J., Petersen, M. L., 2004. Estimation of direct and indirect causal effects in longitudinal studies. In: U. C. Berkeley, *Division of Biostatistics Working Paper 155*, pp. 1–27. The Berkeley Electronic Press, Berkeley.
- Vermunt, J. K., 1996. Causal log-linear modeling with latent variables and missing data. In Engel, U., Reincke, J. (editors), *Analysis of Change: Advanced Techniques in Panel Data Analysis*.
- Wilson, T. P., Bielby, W. T., 1983. Recursive models for categorical data. *Social Science Research*, 12, 109–130.
- Wright, S., 1934. The method of path coefficients. *Annals of Mathematical Statistics*, 5, 161–215.
- Wright, S., 1960. Path coefficients and path regressions: Alternative or complementary concepts? *Biometrics*, 16, 189–202.