PATH ANALYSIS WITH LOGISTIC REGRESSION MODELS

Nobuoki Eshima*, Minoru Tabata** and Geng Zhi***

*) *Oita Medical University, Oita 879-5593, Japan*
**) *Kobe University, Kobe 657-8501, Japan*
***) *Peking University, Beijing 100871, P. R. China*

1. Introduction

Path analysis is usually performed for continuous variables by using linear regression equations (Asher 1976), and the basic idea is applied to the analysis of causal systems of continuous variables, LISREL model (Jöreskog and Sörbom, 1988). In comparison with path analysis of continuous variables, that of categorical variables is complex, because the causal system under consideration cannot be described by linear regression equations. Goodman (1973a, b, 1974) considered path analysis of binary variables by using logistic regression models (Cox, 1970), and discussed the effects by logit parameters. Hagenaars (1998) made a general discussion of path analysis of recursive causal systems of categorical variables by using the directed loglinear model approach that is combined Goodman's approach and Graphical modeling. Although the approach is an analogy to LISREL approach, the discussion of the direct, indirect and total effects was not made.

In this paper, we provide a method of path analysis of categorical variables. Path analysis is discussed in structural logistic regression models. We give definitions of the direct, indirect and total effects, and the effects are explained in terms of log odds ratios. A numerical example is also given to illustrate the present approach.

2. Effects of explanatory variables in logistic regression models without interactive terms

Let X_i (i =1,2,…,k) be categorical variables having categories {1,2,…,I_i}. Assume that the structural relationship between X_k and X_i (i =1,2,…,k-1) in Fig.1 is expressed by a logistic regression model without interactive terms. Let $\boldsymbol{X}_{pa(k)} = (X_1, X_2, \ldots, X_{k-1})'$; and let $p(x_k \mid \boldsymbol{x}_{pa(k)})$ be the conditional probability of $X_k = x_k$ given $\boldsymbol{X}_{pa(k)} = \boldsymbol{x}_{pa(k)}$. Then, the logit model is given as follows:

$$\text{(2.1)} \qquad \mathrm{Log}[p(x_k \mid \boldsymbol{x}_{pa(k)}) / \{1 - p(x_k \mid \boldsymbol{x}_{pa(k)})\}] = \alpha_{x_k} + \sum_{i=1}^{k-1} \beta_{ix_ix_k}.$$

In considering structural relationships among variables concerned, logistic regression

models are referred to as structural logistic regression models in the present paper. In the above model, we introduce the following dummy variables:

$$X_{ij} = 1 \text{ (if } X_i = j) \text{ and } X_{ij} = 0 \text{ (if } X_i \neq j) \quad (j = 1,2,\ldots,I_i;\ i = 1,2,\ldots,k).$$

Random dummy vector $\boldsymbol{X}_i = (X_{i1}, X_{i2},\ldots, X_{iI_i})'$ and the corresponding categorical variables are identified. Then, model (2.1) is rewritten as follows:

$$\text{(2.1)} \qquad \text{Log}[p(x_k \mid \boldsymbol{x}_{pa(k)}) / \{1 - p(x_k \mid \boldsymbol{x}_{pa(k)})\}] = x_k' \boldsymbol{\alpha}_k + \sum_{i=1}^{k-1} x_k' \boldsymbol{B}_i x_i,$$

where $\boldsymbol{\alpha}_k$ and $\boldsymbol{B}_i$ are a vector and a matrix corresponding to parameters α_{xk} and β_{ixixk}, respectively. The log odds ratio of $X_k = x_k$ over $X_k = x_k^*$ is given by

$$\text{(2.2)} \qquad \log \text{OR}(x_k, x_k^* ; \boldsymbol{x}_{pa(k)}, \boldsymbol{x}_{pa(k)}^*) = \sum_{i=1}^{k-1} \text{tr}\, \boldsymbol{B}_i (x_i - x_i^*)(x_k - x_k^*)'.$$

When we formally substitute the baselines x_k^* and $\boldsymbol{x}_{pa(k)}^*$ for the expectations μ_k^* and $\boldsymbol{\mu}_{pa(k)}$ respectively, we have

$$\text{(2.2)} \qquad \log \text{OR}(x_k, \mu_k ; \boldsymbol{x}_{pa(k)}, \boldsymbol{\mu}_{pa(k)}) = \sum_{i=1}^{k-1} \text{tr}\, \boldsymbol{B}_i (x_i - \mu_i)(x_k - \mu_k)'.$$

The above quantity can be viewed as the log odds ratio. First, the total effect of $\boldsymbol{x}_{pa(k)}$ on x_k is defined by

$$\text{(2.3)} \qquad e_T(\boldsymbol{x}_{pa(k)} \rightarrow x_k) = \log \text{OR}(x_k, \mu_k ; \boldsymbol{x}_{pa(k)}, \boldsymbol{\mu}_{pa(k)}).$$

Secondly, the direct effect of x_i on $\boldsymbol{x}_k$ is defined by

$$e_d(x_i \rightarrow x_k) = \text{tr}\, \boldsymbol{B}_i (x_i - \mu_i)(x_k - \mu_k)'.$$

This quantity can be regarded as the partial log odds ratio with respect to x_i and x_k given other variables, and is denoted by $\log \text{OR}(x_k, \mu_k; x_i, \mu_i \mid \boldsymbol{\mu}_{pa(k)}, x_{i+1}, x_{i+2},\ldots, x_{k-1})$. Thirdly, the total effect of x_i on x_k is defined by

$$\text{(2.4)} \quad e_T(x_i \rightarrow x_k) = \log \text{OR}\{x_k, \mu_k ; (x_i, \mu_{i+1}(x_i),\ldots, \mu_{k-1}(x_i)), (x_i, \mu_{i+1},\ldots, \mu_{k-1}) \mid \boldsymbol{\mu}_{pa(i)}\}$$
$$= \boldsymbol{B}_i (x_i - \mu_i)(x_k - \mu_k)' + \sum_{j=i+1}^{k-1} \text{tr}\, \boldsymbol{B}_j (\mu_j(x_i) - \mu_j)(x_k - \mu_k)'.$$

The first term is the direct effect of x_i on $\boldsymbol{x}_k$, so the indirect effect is defined by the second term:

$$\text{(2.4)} \quad e_{ind}(x_i \rightarrow x_k) = \log \text{OR}\{x_k, \mu_k ; (x_i, \mu_{i+1}(x_i),\ldots, \mu_{k-1}(x_i)), (x_i, \mu_{i+1},\ldots, \mu_{k-1}) \mid \boldsymbol{\mu}_{pa(i)}\}$$
$$= \sum_{j=i+1}^{k-1} \text{tr}\, \boldsymbol{B}_j (\mu_j(x_i) - \mu_j)(x_k - \mu_k)'.$$

In the above consideration, the direct, indirect and total effects can be interpreted through log odds ratios. With respect to the indirect effect, we have the following theorem:

Theorem 1. If X_i and X_j are independent, then

$$e_{ind}(x_i \rightarrow x_k) = 0.$$

Proof. If X_i and X_j are independent, we get $\mu_j(x_i) = \mu_j$. From (2.4) the theorem follows. □

Lastly, the average effects are defined in order to summarize the effects defined above. The expectation of (2.4) is

$$E\{e_T(X_i \rightarrow X_k)\} = \text{tr}\, \boldsymbol{\beta}_i\{\text{Cov}(X_i, X_k)\} + \sum_{j=i+1}^{k-1} \text{tr}\, \boldsymbol{\beta}_j\{\text{Cov}(\mu_j(X_i), \mu_k(X_i))\}.$$

This is the average total effect. The first and second terms are the average direct and indirect effects, respectively.

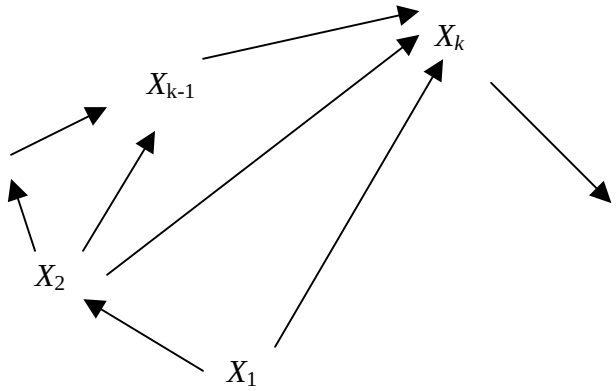

Figure 1. Path Diagram of a Fully Recursive Causal System of Variables

3. Numerical example

Table 1. Data of Primary Food Choice of Alligators by Lake and Size

Lake	Size	Food				
		Fish	Invertebrate	Reptile	Bird	Other
Hancock	small	23	4	2	2	8
	large	7	0	1	3	5
Ocklawaha	small	5	11	1	0	3
	large	13	8	6	1	0
Trafford	small	5	11	2	1	5
	large	8	7	6	3	5
George	small	16	19	1	2	3
	large	17	1	0	1	3

Table 1 shows the data for an investigation of factors influencing the primary food choice of alligators (see Agresti, 1990, pp. 307-310). In this example, X_1 = Lake: lakes that alligators live; X_2 = Size: sizes of alligators; and X_3 = primary foods of alligators. The structural relationships among variables are shown in Fig. 2. In order to show the present approach, the structural relationship between *Food* and (*Lake*,*Size*) is considered. The estimated average effects of *Lake* and *Size* on *Food* are shown in Table 2. The details are omitted for want of space.

Table 3. Average Effects of Size and Lake on Food

	Direct	Indirect	Total
Size	0.085	--------	0.085
Lake	0.221	-0.015	0.206

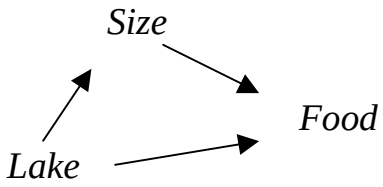

Figure 2. Path Diagram of *Lake*, *Size* and *Food*

4. Conclusion

This paper has provided a method of path analysis of categorical variable. The total, direct and indirect effects have been defined. The effects defined in this paper can be interpreted in terms of (i) log odds ratios; (ii) changes of uncertainty of a response variable; and (iii) the inner product of explanatory and response variables. This is an advantage of the present approach. It is important to extend the present approach to a method for treating more general causal systems of categorical variables, however it needs a further discussion following the present study.

References

Agresti, A. (1990) *Categorical Data Analysis*, New York: John Wiley & Sons, Inc.

Asher, H. B. (1976) *Causal Modelling*, Beverly Hills: Sage Publications, Inc.

Cox, D. R. (1970) *The Analysis of Binary Data*, London: Chapman and Hall, Inc.

Goodman, L. A. (1973a) *American Journal of Sociology*, **78**, 1135-1191.

Goodman, L. A. (1973b) *Bimetrika*, **60**, 179-192.

Goodman, L. A. (1974) *Biometrika*, **61**, 215-231.

Hagenaars, J. A. (1998) Sociological Methods & Research, **26**, 436-489.

Jöreskog, K. G. and Sörbom, D. (1989) *LISREL 7 User's Reference Guide*, Chicago: SPSS, Inc.