Path Analysis for Recursive Generalized Linear Model Systems

Eshima, Nobuoki
Oita University, Department of Biostatistics
1-1, Idaiga-oka, Hasama
Yufu 879-5593, Japan
E-mail: eshima@med.oita-u.ac.jp

Tabata, Minoru
Osaka Prefecture University, Department of Mathematical Sciences
1-1,Gakuencho, Naka
Sakai 599-8531, Japan

Ohyama, Tetsuji
Oita University, Department of Biostatistics
1-1, Idaiga-oka, Hasama
Yufu 879-5593, Japan

1 Introduction

Path analysis is usually carried out in causal systems of continuous variables, i.e. Linear Structural Equation Model (LISREL) (Bentler & Weeks, 1980). In LISREL approach, causal relationships among variables concerned are described by a path diagram, and the relationships are translated into linear equations of the variables. 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. Hagenaars (1998) made a discussion of path analysis of categorical variables by using a loglinear model approach. Although the approach is an analogy to LISREL, the discussion of the direct and indirect effects was not made. In path analysis with categorical variables, it is a question how the effects are measured. Eshima et al. (2001) proposed a method of path analysis of categorical variables by using logit models. In this approach, the direct and indirect effects of variables are discussed according to log odds ratios and the average effects are defined for summarizing them; however the interpretation of the average effects was not provided. Kuha & Goldthorpe (2010) proposed a path analysis method according to log odds ratios; however increasing categories in variables makes the path analysis to be complex.
This paper proposes a basic method for path analysis in causal systems with generalized linear models (GLMs). First, the odds ratio in GLMs is discussed and the interpretation in entropy is given. Second, the total, direct and indirect summary effects in GLMs are discussed by using log odds ratios, and to standardize the effects the entropy crenelation coefficient (ECC) or the entropy coefficient of determination in GLMs is employed. A numerical example is given to illustrate the present approach. Finally, discussions and conclusions to this study are provided.

2 An Example

The data for an investigation of factors influencing the primary food choice of alligators are analyzed with a generalized logit model (Agresti, 2002; pp. 268-271). In this example, explanatory variables are $X_{(L)}$: lakes where alligators live, {1. Hancock, 2. Oklawaha, 3. Trafford, 4. George}; and $X_{(S)}$: sizes

of alligators, {1. Small, 2. Large}; and the response variable is Y: primary food choice of alligators, {1. Fish, 2. Invertebrate, 3. Reptile, 4. Bird, 5. Other}. Lake $X_{(L)}$ is a fixed variable, and it can be assumed that $X_{(L)}$ precedes Size of alligators $X_{(S)}$ and that $X_{(L)}$ and $X_{(S)}$ affect Food Y. The path diagram is shown in Figure 2.1.

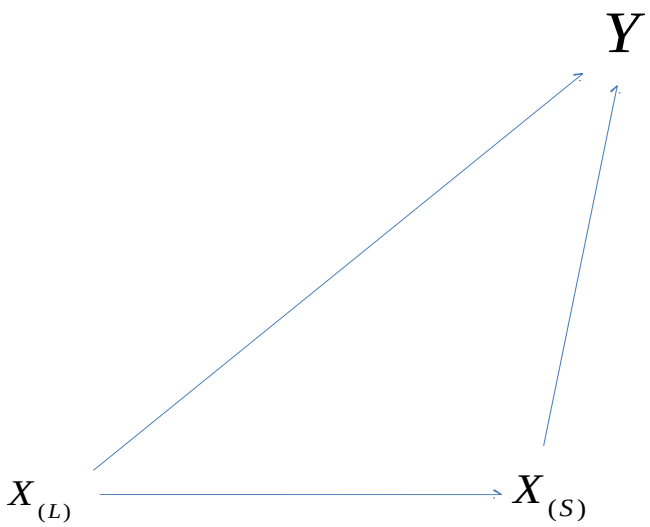

Figure 2.1. Path Diagram of $X_{(L)}$, $X_{(S)}$, and Y

Let us consider the effects of $X_{(L)}$ and $X_{(S)}$ on Y. The following dummy variables are introduced for the categorical variables. Let

$$X_{(L)i} = \begin{cases} 1 & (X_{(L)} = i) \\ 0 & (X_{(L)} \neq i) \end{cases} \quad (i = 1, 2, 3, 4); \; X_{(S)j} = \begin{cases} 1 & (X_{(S)} = j) \\ 0 & (X_{(S)} \neq j) \end{cases} \quad (j = 1, 2)$$

and

$$Y_k = \begin{cases} 1 & (Y = k) \\ 0 & (Y \neq k) \end{cases} \quad (k = 1, 2, 3, 4, 5) \; .$$

Then, the explanatory variables $X_{(L)}$, $X_{(S)}$ and the response variable Y are identified with the correspondent dummy random vectors $\boldsymbol{X}_{(L)} = \left(X_{(L)1}, X_{(L)2}, X_{(L)3}, X_{(L)4}\right)^T$, $\boldsymbol{X}_{(S)} = \left(X_{(S)1}, X_{(S)2}\right)^T$ and $\boldsymbol{Y} = (Y_1, Y_2, Y_3, Y_4, Y_5)^T$, respectively. In this analysis, the following generalized logit model is assumed:

$$f(y|x) = \frac{\exp\left\{\boldsymbol{y}^T\left(\boldsymbol{\alpha} + \boldsymbol{B}_{(L)}^T x_{(L)} + \boldsymbol{B}_{(S)}^T x_{(S)}\right)\right\}}{\sum_{\boldsymbol{y}} \exp\left\{\boldsymbol{y}^T\left(\boldsymbol{\alpha} + \boldsymbol{B}_{(L)}^T x_{(L)} + \boldsymbol{B}_{(S)}^T x_{(S)}\right)\right\}}$$

where $\boldsymbol{\alpha}$, $\boldsymbol{B}_{(L)}$, and $\boldsymbol{B}_{(S)}$ are 5×1, 5×4 and 5×2 regression parameter matrices, respectively.

The present paper proposes a path analysis method for summarizing the effects of polytomous explanatory variables on response variables, and the analysis is applied to this example.

3 Log odds and entropy

Let $\boldsymbol{X}$ and Y be a $p\times 1$ explanatory variable vector and a response variable respectively, and let $f(y|\boldsymbol{x})$ be the conditional probability or density function of Y given $\boldsymbol{X}=\boldsymbol{x}$. The conditional probability or density function $f(y|\boldsymbol{x})$ is assumed to be the following exponential family distribution:

$$f(y|\boldsymbol{x})=\exp\left\{\frac{y\theta-\boldsymbol{b}(\theta)}{a(\varphi)}+c(y,\varphi)\right\}, \tag{3.1}$$

where θ and φ are parameters, and $a(\varphi)$ (>0), $b(\theta)$ and $c(y,\varphi)$ are specific functions. Let $\boldsymbol{\beta}^T=(\beta_1,\beta_2,...,\beta_p)^T$. Since θ is a function of $\eta=\boldsymbol{\beta}^T x$ through a link function $h(u)$, for simplification the function is denoted by $\theta=\theta(\boldsymbol{\beta}^T\boldsymbol{x})$. Let us consider the following log odds ratio:

$$\mathrm{logOR}(\boldsymbol{x},\boldsymbol{x}_0;y,y_0)=\log\frac{f(y|\boldsymbol{x})f(y_0|\boldsymbol{x}_0)}{f(y_0|\boldsymbol{x})f(y|\boldsymbol{x}_0)}=\frac{1}{a(\varphi)}(y-y_0)\left(\theta(\boldsymbol{\beta}^T\boldsymbol{x})-\theta(\boldsymbol{\beta}^T\boldsymbol{x}_0)\right),$$

where $\boldsymbol{x}_0$ and y_0 are baselines of $\boldsymbol{X}$ and Y, respectively. Since

$$\mathrm{logOR}(\boldsymbol{x},\boldsymbol{x}_0;y,y_0)=\{-\log f(y_0|\boldsymbol{x})-(-\log f(y|\boldsymbol{x}))\}-\{-\log f(y_0|\boldsymbol{x}_0)-(-\log f(y|\boldsymbol{x}_0))\}, \tag{3.2}$$

the log odds ratio (3.2) is the decrease of the uncertainty of response Y in explanatory variable vector $\boldsymbol{X}$, and the quantity can be interpreted as the effect of $\boldsymbol{x}$ on y, where $\boldsymbol{x}_0$ and y_0 are baselines of $\boldsymbol{X}$ on Y, respectively. For levels of the factor vector $\boldsymbol{X}=\boldsymbol{x}_1,\boldsymbol{x}_2,\ldots,\boldsymbol{x}_K$, the means of $\boldsymbol{X}$ and Y are defined as follows:

$$\boldsymbol{\mu}_X\equiv\mathrm{E}(\boldsymbol{X})=\frac{\sum_{k=1}^K\boldsymbol{x}_k}{K}\quad\text{and}\quad\mu_Y\equiv\mathrm{E}(Y)=\frac{\sum_{k=1}^K\mathrm{E}(Y|\boldsymbol{X}=\boldsymbol{x}_k)}{K}.$$

If $\boldsymbol{X}$ is random, the above means are the usual expectations. When the baselines are replaced by $\boldsymbol{\mu}_X$ and μ_Y, respectively, the effect of $\boldsymbol{x}$ on y is defied by

$$\mathrm{logOR}(\boldsymbol{x},\boldsymbol{\mu}_X;y,\mu_Y)\equiv\frac{1}{a(\varphi)}(y-\mu_Y)\left(\theta(\boldsymbol{\beta}^T\boldsymbol{x})-\theta(\boldsymbol{\beta}^T\boldsymbol{\mu}_X)\right).$$

The effect of variable $\boldsymbol{X}$ on Y is defined on the basis of the following basic measure of predictive power:

$$\mathrm{E}(\log\mathrm{OR}(X,\boldsymbol{\mu}_X;Y,\mu_Y)|\boldsymbol{X})\equiv\frac{1}{a(\varphi)}\mathrm{E}\big(\mu_Y\left(\boldsymbol{\beta}^T\boldsymbol{X}\right)-\mu_Y)\left(\theta(\boldsymbol{\beta}^T\boldsymbol{X})-\theta(\boldsymbol{\beta}^T\boldsymbol{\mu}_X)\right)\big)=\frac{\mathrm{Cov}(\theta,Y)}{a(\varphi)}. \tag{3.3}$$

The above quantity can be expressed by a symmetric type of the Kullback information between GLM with (3.1) and the null model with $\boldsymbol{\beta}=\boldsymbol{0}$ (Eshima & Tabata, 2007), so we denote (3.3) as $KL\left(\boldsymbol{X},Y\right)$ in this paper. It is useful to standardize the information by ECC or ECD. In GLM with random component (3.1), ECC is

$$\mathrm{ECorr}(X,Y)=\frac{\mathrm{Cov}(Y,\theta)}{\sqrt{\mathrm{Var}(Y)}\sqrt{\mathrm{Var}(\theta)}},$$

and, it can be interpreted as the proportion of the explained entropy of response Y. On the other hand, ECD is given by

$$\mathrm{ECD}(X,Y)=\frac{\mathrm{Cov}(Y,\theta)}{\mathrm{Cov}(Y,\theta)+a(\varphi)}.$$

Another expression of ECD is as follows:

$$\mathrm{ECD}(X,Y)=\frac{\mathrm{Cov}(Y,\theta)/a(\varphi)}{\mathrm{Cov}(Y,\theta)/a(\varphi)+1}=\frac{KL(X,Y)}{KL(X,Y)+1}.$$

The measure is interpreted as the proportion of explained variation of Y in entropy (Eshima & Tabata, 2010). In this paper, ECD is mainly used for assessing the effects of explanatory variables.

4 Effect Analysis in Recursive GLM systems

For simplicity of the discussion, the path system of variables X_i $(i = 1, 2, 3)$ shown in Figure 4.1 is considered, and by using a similar discussion as the previous section we propose a method for effect decomposition. In the recursive system, X_i precedes X_{i+1} $(i = 1, 2)$, and we discuss the effects of X_1 and X_2 on X_3. Let μ_i be the expectations of X_i $(i = 1, 2, 3)$. Then, for a GLM with the conditional density or probability function of X_3 given $(X_1, X_2) = (x_1, x_2)$ (4.1), the total effect of $(X_1, X_2) = (x_1, x_2)$ on $X_3 = x_3$ can be defined by using the following odds ratio:
$\frac{f(x_3|x_1,x_2)\ f(\mu_3|\mu_1,\mu_2)}{f(\mu_3|\mu_1,x_2)f(x_3|x_1,\mu_2)} = \frac{(x_3-\mu_3)\left(\theta(\boldsymbol{\beta}^T(x_1,x_2))-\theta(\boldsymbol{\beta}^T(\mu_1,\mu_2))\right)}{a(\varphi)}$.

By taking the expectation of the above effect, we have the total effect of (X_1, X_2) on X_3:

$\frac{Cov(\theta(\boldsymbol{\beta}^T(X_1,X_2)),X_3)}{a(\varphi)}$.

Let $\mu_2(x_1)$ and $\mu_3(x_1)$ be the conditional expectations X_2 and X_3 given $X_1 = x_1$, respectively. The total effect of $X_1 = x_1$ on $X_3 = x_3$ is defined by

$$\frac{f(x_3|x_1,x_2)\ f(\mu_3|\mu_1,\mu_2)}{f(\mu_3|\mu_1,x_2)f(x_3|x_1,\mu_2)} - \frac{f(x_3|x_1,x_2)\ f(x_3|\mu_1(x_1),\mu_2(x_1))}{f(x_3|\mu_1(x_1),x_2)f(x_3|x_1,\mu_2(x_1))}$$
$$= \frac{(x_3-\mu_3)\left(\theta(\boldsymbol{\beta}^T(x_1,x_2))-\theta(\boldsymbol{\beta}^T(\mu_1,\mu_2))\right)}{a(\varphi)} - \frac{(x_3-\mu_3(x_1))\left(\theta(\boldsymbol{\beta}^T(x_1,x_2))-\theta(\boldsymbol{\beta}^T(x_1,\mu_2(x_1)))\right)}{a(\varphi)}.$$

The second term of the above equation is defined as the total effect of $X_2 = x_2$ on $X_3 = x_3$ given $X_1 = x_1$, because X_1 precedes X_2. By taking the expectation of the right hand side of the above equation, the total effect of X_1 on X_3 is given by

$\frac{Cov(\theta(\boldsymbol{\beta}^T(X_1,X_2)),X_3)}{a(\varphi)} - \frac{Cov(\theta(\boldsymbol{\beta}^T(X_1,X_2)),X_3|X_1)}{a(\varphi)}$.

The second term implies the total effect of X_2 on X_3, i.e.

$\frac{Cov(\theta(\boldsymbol{\beta}^T(X_1,X_2)),X_3|X_1)}{a(\varphi)}$.

The direct effect of $X_1 = x_1$ on $X_3 = x_3$ can be discussed based on the following odds ratio given $X_2 = x_2$:

$\frac{f(x_3|x_1,x_2)\ f(\mu_3|x_1,\mu_2)}{f(\mu_3|x_1,x_2)f(x_3|x_1,\mu_2)} = \frac{(x_3-\mu_3(x_2))\left(\theta(\boldsymbol{\beta}^T(x_1,x_2))-\theta(\boldsymbol{\beta}^T(\mu_1,x_2))\right)}{a(\varphi)}$.

By taking the conditional expectation of above effect we have

$\frac{Cov(\theta(\boldsymbol{\beta}^T(X_1,X_2)),X_3|X_2)}{a(\varphi)}$.

Standardizing the above effects based on ECD we defined the total effect of X_1 and X_2 on X_3, the total, direct and indirect effects of X_1 on X_3, and the total effect of X_2 on X_3 as follows:

$e_T\left((X_1, X_2) \longrightarrow X_3\right) = \frac{Cov(\theta(\boldsymbol{\beta}^T(X_1,X_2)),X_3)}{Cov(\theta(\boldsymbol{\beta}^T(X_1,X_2)),X_3)+a(\varphi)}$,
$e_T\left(X_1 \longrightarrow X_3\right) = \frac{Cov(\theta(\boldsymbol{\beta}^T(X_1,X_2)),X_3)-Cov(\theta(\boldsymbol{\beta}^T(X_1,X_2)),X_3|X_1)}{Cov(\theta(\boldsymbol{\beta}^T(X_1,X_2)),X_3)+a(\varphi)}$,
$e_D\left(X_1 \longrightarrow X_3\right) = \frac{Cov(\theta(\boldsymbol{\beta}^T(X_1,X_2)),X_3|X_2)}{Cov(\theta(\boldsymbol{\beta}^T(X_1,X_2)),X_3)+a(\varphi)}$, $e_I\left(X_1 \longrightarrow X_3\right) = e_T\left(X_1 \longrightarrow X_3\right) - e_D\left(X_1 \longrightarrow X_3\right)$,
$e_T\left(X_2 \longrightarrow X_3\right) = \frac{Cov(\theta(\boldsymbol{\beta}^T(X_1,X_2)),X_3|X_1)}{Cov(\theta(\boldsymbol{\beta}^T(X_1,X_2)),X_3)+a(\varphi)}$.

In the present path analysis, the total and direct effects are positive. The approach can be extended to a general recursive GLM system with more than three variables.

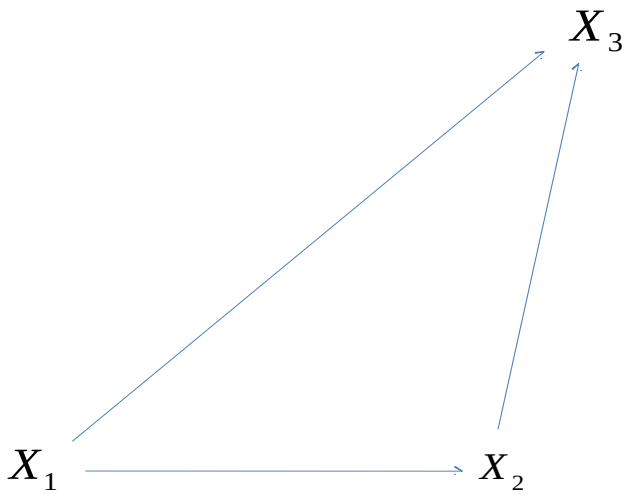

Figure 4.1. Path Diagram of X_1, X_2, and X_3

Analysis of the Example. According to the above estimates we have
$\mathrm{Cov}(\theta, Y) = 0.329$ $(SE = 0.082)$, $\mathrm{Cov}\left(\widehat{\theta, Y}|X_{(L)}\right) = 0.101$ (0.045)
and $\mathrm{Cov}\left(\widehat{\theta, Y}|X_{(S)}\right) = 0.260$ (0.072).

From the above estimates we have the effects of Lake $(X_{(L)})$ and Size $(X_{(S)})$ on Food (Y).

$e_T\left((X_{(L)}, X_{(S)}) \longrightarrow Y\right) = \frac{0.329}{0.329+1} = 0.248,$
$e_T\left(X_{(L)} \longrightarrow Y\right) = \frac{0.329-0.101}{0.329+1} = 0.172,$
$e_D\left(X_{(L)} \longrightarrow Y\right) = \frac{0.260}{0.329+1} = 0.196,$ $e_I\left(X_{(L)} \longrightarrow Y\right) = \frac{0.329-0.101-0.260}{0.329+1} = -0.02\,4.$

$e_T\left(X_{(S)} \longrightarrow Y\right) = \frac{0.101}{0.329+1} = 0.076.$

5 Discussion

In the usual path analysis of continuous variable systems, the regression coefficients and the correlation coefficients of factors and response variables are decomposed into components, and path analysis is easily carried out; however for categorical variable systems such a technique cannot be applied. In path analysis of categorical variables using logit and loglinear models (Goodman, 1973a, b; Hagenaars, 1998), the effects in causal systems were assessed with odds ratios; however the discussion of the direct and indirect effects did not made. In the present paper, a path analysis approach with ECD is proposed for measuring the factor effects in GLMs. As shown in an example, the results of path analysis provide summaries of the effects of explanatory variables based on log odds ratios. A similar approach can also be made with ECC. By using the present approach, the direct and indirect effects of factors on response variables can be calculated in all GLMs. For LISREL the ordinary path analysis method may be better than the present approach; however the present path analysis approach will have potential for a wide applicability in practical analyses of recursive GLM causal systems, especially for categorical variables.

Acknowledgement: This research was supported by Grant-in-aid for Scientific Research 22500260, Ministry of Education, Culture, Sports, Science and Technology of Japan.

References

[1] Agresti, A. (2002). *Categorical Data Analysis*, Second Edition, John Wiley & Sons, Inc.: New York.

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

[3] Bentler, P.M. & Weeks, D.B. (1980) Linear structural equations with latent variables, Psychometrika, 45, 289-308.

[4] Eshima, N. & Tabata, M. (2007). Entropy correlation coefficient for measuring predictive power of generalized linear models, *Statistics & Probability Letters*; **77**, 588-593.

[5] Eshima, N & Tabata, M (2010) Entropy coefficient od determination for generalized linear models, *Computational Statistics and Data Analysis*, 54, 1381-1389.

[6] Eshima, N., Tabata, M. & Geng, Z. (2001). Path analysis with logistic regression models: effect analysis of fully recursive causal systems of categorical variables, *Journal of the Japan Statistical Society*; **31**: 1-14.

[7] Goodman, L. A. (1973a). Causal analysis of data from panel studies and other kinds of surveys, *American Journal of Sociology*; 78, 1135-1191.

[8] Goodman, L. A. (1973b). The analysis of multidimensional contingency tables when some variables are posterior to others: a modified path analysis approach, *Biometrika*; **60**: 179-192.

[9] Hagenaars, J. A. (1998). Categorical causal modeling : latent class analysis and directed loglinear models with latent variables, *Sociological Methods & Research*; **26**: 436-489.

[10] Kuha, J. & Goldthorpe, J. H. (2010) Path analysis for discrete variables: the role of education in social mobility, *J. R. Statist. Soc. A*, 1-19.

[11] McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear models*, 2nd Ed. Chapman and Hall: London.

[12] Nelder, J. A. and Wedderburn, R. W. M. (1972). Generalized linear model, *Journal of the Royal Statistical Society* A; **135**: 370-384.

RÉSUMÉ (ABSTRACT)

The objective of the present paper is to propose a path analysis method for causal systems with generalized linear models (GLMs). First, the usual path analysis method for linear equation models is reviewed. Second, a brief introduction of the entropy correlation coefficient (ECC) and the entropy coefficient of determination (ECD) is given, and the effects of factors in GLMs are discussed by using ECC and ECD. Third, a path analysis method for recursive GLM systems is proposed. A numerical illustration is also given to demonstrate the present approach.