THE UTILITY OF SYSTEMS OF SIMULTANEOUS LOGISTIC RESPONSE EQUATIONS

by

Stephen S. Brier

Technical Report No. 284

4 March 1977

Department of Applied Statistics

School of Statistics

University of Minnesota

St. Paul, Minnesota 55101

ABSTRACT

Structural equation models have been used in the past to provide causal interpretations of the relationships between continuous variables. Recently attempts have been made to extend these techniques to the analysis of categorical data. Some crucial differences between the two situations are pointed out and a number of important problems that arise are discussed. Using a logistic response model there is no identification problem but there is no way to estimate reciprocal effects. It is shown that the usual models for categorical data analysis are not the proper models for estimating reciprocal effects.

Key Words

Structural equations, logistic models, contingency tables, identification.

1. Introduction

Econometricians have spent considerable time developing multiple-equation models to describe systems in which variables have simultaneous effects on each other. This general area of research falls under the title of structural equation models. The books by Fisher (1966) and Goldberger (1964) present a basic reference for the subject. More recently, sociologists such as Duncan (1975) have developed these techniques in a sociological framework. Until quite recently all of the models were formulated to describe continuous data, i.e. normal-theory models. Nerlove and Press (1976) develop what they refer to as multivariate log-linear models for dealing with categorical response variables that may depend on continuous variables, and they present model specification and estimation techniques for a number of interesting situations. The work of Schmidt and Strauss (1975) goes in the same direction as Nerlove and Press. Goodman (1973) also presents multiple-equation logit models for the analysis of categorical data when there is a natural ordering of causal effects and presents structural models and path diagrams for understanding these relationships.

Although some analogies to the normal-theory simultaneous equation models have been pointed out,we will show that in the natural logistic model formulations we explore there are a number of points of difference between the categorical case and the traditional continuous-variable simultaneous-model situation that have not been explicitly noted before. We show first that there is not an identification problem in the discrete multiple-logit-equation case, and secondly that there is no way to separate reciprocal effects of variables.

2. Two Endogenous and Two Exogenous Variables

To illustrate the relevant points we consider an example with four variables,

two of which are exogenous or determined outside of the system and two of which are endogenous. The following pair of equations describes the traditional simultaneous equations model:

$$X_1 = \alpha_1 + \beta_{12}X_2 + \beta_{13}X_3 + \beta_{14}X_4 + \varepsilon_1 \ , \tag{1}$$

$$X_2 = \alpha_2 + \beta_{21}X_1 + \beta_{23}X_3 + \beta_{24}X_4 + \varepsilon_2 \ , \tag{2}$$

where we assume $\underset{\sim}{\varepsilon}' \equiv (\varepsilon_1, \varepsilon_2)$ to have a bivariate normal distribution with mean 0 and arbitrary covariance matrix. In this model X_1 and X_2 are the endogenous variables.

Without any further restrictions on parameters, equations (1) and (2) are not specified. To see what this means look at what is usually called the reduced form of the model given above. We are assuming that the variables X_1, X_2, ε_1 and ε_2 are normally distributed and, since they are exogenous, that X_3 and X_4 are fixed. Thus the distribution of the observations is given by the conditional distribution of $\underset{\sim}{X}^{(1)}$ given $\underset{\sim}{X}^{(2)}$ where we define $\underset{\sim}{X}^{(1)} \equiv (X_1, X_2)$ and $X^{(2)} \equiv (X_3, X_4)$. Thus our basic assumption is that

$$\underset{\sim}{X}^{(1)} \mid \underset{\sim}{X}^{(2)} = \underset{\sim}{x}^{(2)} \sim N(\underset{\sim}{\gamma} + \underset{\sim}{x}^{(2)}\underset{\sim}{\delta}, \Sigma) \tag{3}$$

where $\underset{\sim}{\gamma} = (\gamma_1, \gamma_2)$, $\underset{\sim}{\delta} = \begin{pmatrix} \delta_{11} & \delta_{12} \\ \delta_{21} & \delta_{22} \end{pmatrix}$ and $\Sigma = \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{pmatrix}$.

Thus $\underset{\sim}{\gamma}$ and $\underset{\sim}{\delta}$ are the regression parameters and Σ is the residual covariance matrix.

There are no problems in the estimation of the parameters in (3). The problem arises when we try to relate the parameters in the simultaneous model, (1) and (2), to those in (3). Note that there are nine parameters in (3) and eleven parameters in (1) and (2). Thus not all of the parameters of the

simultaneous system are uniquely determined from the distribution of the endogenous variables. Specifically β_{12} and β_{21} cannot be determined uniquely and so it makes no sense at all to try to estimate them from observed data. The only way to identify the parameters in an equation is to assume a priori that one of the exogenous variables does not affect that endogenous variable, e.g. in the model below all parameters are specified and hence estimable:

$$X_1 = \alpha_1 + \beta_{12}X_2 + \beta_{13}X_3 + \varepsilon_1 , \tag{4}$$

$$X_2 = \alpha_2 + \beta_{21}X_1 + \beta_{24}X_4 + \varepsilon_2 . \tag{5}$$

Note that (4) results from setting $\beta_{14} = 0$ in (1) and (5) results from setting $\beta_{23} = 0$ in (2). There are now only nine parameters in the simultaneous model, the same number as in the conditional distribution.

The key feature of this simultaneous model is that β_{12} and β_{21} are distinct parameters in the sense that the posterior distribution for each parameter can be computed using only the prior distribution for that parameter and the likelihood function for the data. These parameters are not regression parameters in any conditional distribution. Thus for the above model we can separate out the effect of X_2 on X_1 (β_{12}) from the effect of X_1 on X_2 (β_{21}). We can do this because we made two a priori assumptions about causal effects in (4) and (5).

Now we consider the situation where all variables are dichotomous, each corresponding to the occurrence or non-occurrence of an event. This corresponds to viewing the observations as forming a 2×2×2×2 contingency table. Each variable may now take on only the values 0 or 1. One of the standard ways of modeling a dichotomous response variable is to use a logistic model as described by Cox (1970) and Fienberg (1977). Cox points out a number of reasons for choosing the logistic transform as a natural parameterization. One point to note

is that the logistic transform yields the natural parameterization in the exponential family (see Lehman (1959). In the logistic formulation a direct analogy to equations (1) and (2) is:

$$\log\left(\frac{p^{+}_{1|2}}{p^{-}_{1|2}}\right) = \alpha_1 + \beta_{12}X_2 + \beta_{13}X_3 + \beta_{14}X_4 \qquad (6)$$

$$\log\left(\frac{p^{+}_{2|1}}{p^{-}_{2|1}}\right) = \alpha_2 + \beta_{21}X_1 + \beta_{23}X_3 + \beta_{24}X_4 \qquad (7)$$

where $p^{+}_{i|j} \equiv \text{Prob}\ (X_i = 1 \mid X_j, X_3, X_4)$ and $p^{-}_{i|j} \equiv \text{Prob}\ (X_i = 0 \mid X_j, X_3, X_4)$.

There are two essential differences between this pair of equations and (1) and (2). First, there are no error terms in (6) and (7). Second, the quantities on the left hand sides of the equations are not random variables but functions of parameters (probabilities). The latter is because the logistic model involves the probability structure and not the individual observations themselves.

In order to talk about identification problems we need to talk about the analogue of the reduced form, or joint structure of the endogenous variables given the exogenous ones, namely a log-linear model (e.g. see Bishop, Fienberg, and Holland (1975)) fit to the four dimensional table. As does the logistic model, the log-linear model treats the logarithms of the probabilities in a contingency table as the "natural" parameters. Let us define $p_{ijk\ell} \equiv$ Prob $(X_1 = i, X_2 = j, X_3 = k, X_4 = \ell)$, e.g. $p_{1001} = \text{Prob}\ (X_1=1, X_2=0, X_3=0, X_4=1)$. The general or saturated log-linear model for our 2×2×2×2 table would be:

$$\begin{aligned}\log p_{ijk\ell} = u &+ u_{1(i)} + u_{2(j)} + u_{3(k)} + u_{4(\ell)} + u_{12(ij)} + u_{13(ik)} + u_{14(i\ell)} \\ &+ u_{23(jk)} + u_{24(j\ell)} + u_{34(k\ell)} + u_{123(ijk)} + u_{124(ik\ell)} \\ &+ u_{234(jk\ell)} + u_{1234(ijk\ell)} \qquad (8)\end{aligned}$$

where each of the u-terms satisfies the constraints of summing to zero over any of its indices, e.g. $\sum_{i=0}^{1} u_{13(ik)} = \sum_{k=0}^{1} u_{13(ik)} = 0$. Thus there are 15 independent parameters in (8) which correspond to the 15 independent probabilities in the 2×2×2×2 table. For a further description of the log-linear model see Bishop, Fienberg, and Holland (1975).

The identification problem in the continuous case involves the relationship between the structural equations (1) and (2) and the reduced form (3). In the dichotomous case identification pertains to the relationship between the conditional logistic equations (6) and (7) and the log-linear model for the overall table. The first question that one might ask is: given (6) and (7), is there a unique structure for the table consistent with them? The answer to this question is no. To see this consider the following unsaturated model:

$$\log p_{ijk\ell} = u + u_{1(i)} + u_{2(j)} + u_{3(k)} + u_{4(\ell)} + u_{12(ij)} + u_{13(ik)} + u_{14(i\ell)} + u_{23(jk)} + u_{24(j\ell)} + u_{34(k\ell)} \qquad (9)$$

The $u_{34(k\ell)}$ parameter corresponds to the exogenous variables X_3, X_4. We can allow this parameter to take on any value without affecting the parameters of (6) and (7). This is reasonable since the "structural equations" of (6) and (7) are models for conditional probabilities and should not depend on the parameters of the marginal distribution of (X_3, X_4).

Now we ask the opposite question. Does the model for the joint table, (9), determine a unique set of parameters in equations (6) and (7)? This question is directly analogous to the question of whether or not the parameters of the simultaneous equations are identified in the continuous case. The answer to this question is yes. To see this we write out the relationships explicitly:

$$\begin{array}{llll} \alpha_1 = 2u_1 & \beta_{12} = 2u_{12} & \beta_{13} = 2u_{13} & \beta_{14} = 2u_{14} \\ \alpha_2 = 2u_2 & \beta_{21} = 2u_{12} & \beta_{23} = 2u_{23} & \beta_{24} = 2u_{24} \end{array} \qquad (10)$$

Thus the parameters of the structural equations of (6) and (7) are fully identified even without making any a priori assumptions about certain effects being zero.

Looking at the equations in (10) we see that $\beta_{12} = \beta_{21}$. Thus the effect of X_1 on X_2 is exactly the same as the effect of X_2 on X_1. Goodman (1973) has noted this point indirectly by placing a single double-headed arrow in his path diagram for the pair of equations, (6) and (7).

What has happened here is comparable to the use of a correlation coefficient to determine the simultaneous effects of two endogenous variables in a system--namely the concept of simultaneity does not apply here. It is important to note that this equality of coefficients is not an artifact of any improper specification. Even if we had a priori assumed that $\beta_{14} = 0$ in (6) and $\beta_{23} = 0$ in (7) we would have $\beta_{12} = \beta_{21}$. In fact we prove below a general result that says the reciprocal effects can never be separated in systems of logistic models.

3. Systems of Simultaneous Logistic Models

Let $\underset{\sim}{Z}$ be a p×1 vector of exogenous variables and let $\underset{\sim}{\gamma}$, $\underset{\sim}{\delta}$ be p×1 vectors of coefficients. We allow the components of $\underset{\sim}{Z}$ to be any mixture of discrete and continuous variables. Denote by X_1 and X_2 the dichotomous endogenous variables. Our model here is:

$$\begin{aligned} \log\left(\frac{p^{+}_{1|2}}{p^{-}_{1|2}}\right) &= \alpha_1 + \beta_{12}X_2 + \underset{\sim}{\gamma}'\underset{\sim}{Z} \\ \log\left(\frac{p^{+}_{2|1}}{p^{-}_{2|1}}\right) &= \alpha_2 + \beta_{21}X_1 + \underset{\sim}{\delta}'\underset{\sim}{Z} \end{aligned} \qquad (11)$$

From the remarks made above it is essential to make our inferences conditional on the exogenous variables $\underset{\sim}{Z}$. Hence we consider $\underset{\sim}{Z} = \underset{\sim}{\eta}$ to be fixed and look at the following 2×2 table of probabilities:

		X_2: 1	X_2: 0
X_1	1	$P_{11}(\underset{\sim}{\eta})$	$P_{10}(\underset{\sim}{\eta})$
	0	$P_{01}(\underset{\sim}{\eta})$	$P_{00}(\underset{\sim}{\eta})$

If we consider $\underset{\sim}{Z} = \underset{\sim}{\eta}$ to be fixed we can write the equations in (11) as:

$$\log\left(\frac{P^{+}_{1|2}(\underset{\sim}{\eta})}{P^{-}_{1|2}(\underset{\sim}{\eta})}\right) = \alpha_1^* + \beta_{12}X_2$$

$$\log\left(\frac{P^{+}_{2|1}(\underset{\sim}{\eta})}{P^{-}_{2|1}(\underset{\sim}{\eta})}\right) = \alpha_2^* + \beta_{21}X_1 \qquad (12)$$

where $\alpha_1^* \equiv \alpha_1 + \underset{\sim}{\gamma}'\underset{\sim}{\eta}$ and $\alpha_2^* \equiv \alpha_2 + \underset{\sim}{\delta}'\underset{\sim}{\eta}$. We have thus reduced the problem to that of the two logistic models in a 2×2 table. From the structure of the simple 2×2 table we know that β_{12} and β_{21} are the same parameter, namely the logarithm of the cross product ratio in the 2×2 table. Note that we specified this ratio to be independent of $\underset{\sim}{Z}$. This result as proved here is for dichotomous variables but it generalizes immediately to situations where X_1 and X_2 are discrete variables with more than two possible categories. This is because we can model a two-dimensional contingency table by a series of logistic models. For a discussion of this see Fienberg (1977) or Nerlove and Press (1976).

It is important to note again that the crucial difference between the

simultaneous logistic models and the normal-theory analogues is the conditioning implied in the logistic equations. We condition on different variables in different equations, e.g. in (6) we condition on X_2 and in (7) we condition on X_1. There is no such conditioning in equations (4) and (5).

4. Another Look at the 2^4 Table

It is interesting to note that we can obtain different reciprocal effects, but only by conditioning on different exogenous variables in different equations. We again consider the 2^4 table and its structure given by equation (9). We consider conditional logistic models for the conditional probabilities of variable 1 given 2 and 3 and for variable 2 given 1 and 4. Thus:

$$\log \frac{\tilde{p}_1^+}{\tilde{p}_1^-} = w + w_{2(j)} + w_{3(k)} + w_{23(jk)}$$

$$\log \frac{\tilde{p}_2^+}{\tilde{p}_2^-} = v + v_{1(i)} + v_{4(\ell)} + v_{14(i\ell)} \qquad (13)$$

where $\tilde{p}_1^+ \equiv \text{Prob}\ (X_1 = 1 | X_2, X_3)$ and $\tilde{p}_2^+ \equiv \text{Prob}\ (X_2 = 1 | X_1, X_4)$. $\tilde{p}_1^-$ and $\tilde{p}_2^-$ are defined analogously. The equations in (13) follow directly from (9), and the structure of (9) uniquely determines the parameters in these logistic models. What we have done in the first equation is in effect collapsed the table over variable 4 while in the second equation we have collapsed over variable 3. Note that there are now interaction effects in the two logistic models. With this formulation we see that $w_{2(1)}$ and $v_{1(1)}$ are not equal any longer so we can get different reciprocal effects. However they are still not distinct parameters as defined above.

Another interesting question is whether or not the pair of equations in

(13) uniquely determines the overall table. In this case they do not, but for an interesting discussion of when the overall probabilities can be determined from pairs of logistic models see Goodman (1973). This section has been a digression since we have taken the point of view that we may condition on different subsets of variables in different equations. While leading to some interesting points this is not a sensible approach since one should condition on the entire set of exogenous variables.

5. Other Simultaneous-Equation Models

The only technique we know that allows simultaneous estimation of reciprocal effects involves the concept of latent variables. Muthén (1976) develops models to analyze a problem presented by Duncan (1975b) in which there are two dichotomous endogenous variables, X_1 and X_2. The basic idea is to estimate the parameters in the following pair of simultaneous equations:

$$\begin{aligned} \eta_1 &= \beta_{12}\eta_2 + \underset{\sim}{\gamma}_1'\underset{\sim}{Y} + \varepsilon_1 \\ \eta_2 &= \beta_{21}\eta_1 + \underset{\sim}{\gamma}_2'\underset{\sim}{Y} + \varepsilon_2 \end{aligned} \tag{14}$$

Here η_1 and η_2 are latent (unobservable) variables causing the dichotomous response of X_1 and X_2. $\underset{\sim}{Y}$ represents the vector of exogenous variables and ε_1, ε_2 are error terms. Since the η's are unobservable variables some further assumptions must be made. Essentially the two basic assumptions made by Muthén are that (1) the probabilities associated with X_1 and X_2 satisfy a probit model, i.e.

$$\text{Prob}\,(X_1 = 1 \mid \eta_i) = \int_{-\infty}^{\eta_i - \tau_i} \phi(z)\, dz \qquad i = 1,2 \tag{15}$$

where $\phi(z)$ is the standard normal density, and (2) the distribution of $\underset{\sim}{\eta} \equiv (\eta_1, \eta_2)$

conditional on $\underset{\sim}{Y}$ is bivariate normal. It then follows that

$$\text{Prob}\ (X_i = 1 \mid \underset{\sim}{Y}) = \int_{-\infty}^{g_i + h_i' \underset{\sim}{Y}} \phi(z)dz \quad . \tag{16}$$

With this representation, if certain a priori assumptions are made about some parameters in $\underset{\sim}{\gamma}_1$ and $\underset{\sim}{\gamma}_2$ of (14), β_{12} and β_{21} are identified and are in fact distinct. Thus if the idea of latent variables is meaningful this method can be useful in formulating structural models.

6. Summary

The main point of this article is that simultaneous logistic models are not the proper way of handling discrete variables if one is really interested in "simultaneous" effects. This is certainly not meant to imply that the logistic model is to be discarded as an aid in causal interpretation of discrete data. As mentioned before, Goodman (1973) develops an interesting way of looking at causal chains of dichotomous variables using the logistic model. He models a recursive system by conditioning on different sets of variables as he goes up the chain and gets estimates of probabilities that could not be derived by just looking at an overall contingency table. The models of Nerlove and Press (1976) use the logistic model to analyze data involving mixtures of discrete and continuous variables for predictors. Their model essentially corresponds to a reduced form structure.

Thus, while a very useful formulation, the logistic model cannot yield the analogue of a structural equation model because it is essentially a conditional probability. That this point has not always been realized is seen in a comment by Schmidt and Strauss (1975) to the effect that the "discontinuous

analogue to simultaneous equation bias may be reasonably thought to exist" when estimating a single logistic equation. This is clearly not the case. In the continuous case if you don't use a simultaneous approach (e.g. two stage least squares) then you are estimating a regression coefficient which is the wrong parameter. This is why the usual least squares estimates are asymptotically biased. As noted by Nerlove and Press, in the logistic model the estimator from a single equation is consistent although, in general, not efficient. Hopefully then this paper sheds some light on why there is a need to develop new models, such as those of Muthén, to investigate simultaneous relationships.

References

Bishop, Y.M.M., Fienberg, S.E., and Holland, P.W. (1975). Discrete Multivariate Analysis: Theory and Practice. Cambridge: MIT Press.

Cox, D.R. (1970). The Analysis of Binary Data. London: Methuen.

Duncan, O.D. (1975). Structural Equations Models. New York: Academic Press

Duncan, O.D. (1975b). Personal Communication.

Fienberg, S.E. (1977). The Analysis of Cross-Classified Categorical Data. Cambridge: MIT Press.

Fisher, F.M. (1966). The Identification Problem in Econometrics. New York: McGraw-Hill.

Goldberger, A.S. (1964). Econometric Theory. New York: Wiley.

Goodman, L.A. (1973). The Analysis of Multidimensional Contingency Tables When some Variables are Posterior to Others: A Modified Path Analysis Approach. Biometrika 60, 179-192.

Lehman, E.L. (1959). Testing Statistical Hypotheses. New York: Wiley.

Muthén, B. (1976). Structural Equation Models with Dichotomous Dependent Variables: A Sociological Analysis Problem Formulated by O.D. Duncan. Research Report 76-19. Department of Statistics, University of Uppsala, Sweden.

Nerlove, M. and Press, S.J. (1976). Multivariate Log-Linear Probability Models for the Analysis of Qualitative Data. Discussion Paper No. 1, Center for Statistics and Probability: Northwestern University.

Schmidt, P. and Strauss, R.P. (1975). Estimation of Models with Jointly Dependent Qualitative Variables: A Simultaneous Logit Approach. Econometrica 43, 745-755.