Statistica Sinica 1(1991), 51-64

FEEDBACK MODELS FOR DISCRETE AND CONTINUOUS TIME SERIES

Scott L. Zeger and Kung-Yee Liang

Johns Hopkins University

Abstract: In public health research, it is common to follow a cohort of subjects over time, observing a vector of health indicators and a set of covariates at each of many visits. An objective of analysis is to characterize the inter-dependencies, in particular, the feedback of one response upon another while accounting for the covariates. With Gaussian responses, multivariate autoregressive models that incorporate feedback are commonly used. This paper discusses analogous Markov models for multivariate discrete and mixed discrete/continuous response variables. One special case is an extension of seemingly unrelated regressions to discrete and continuous outcomes. A generalized estimating equations approach that requires correct specification of only conditional means and variances is discussed. The methods are illustrated by a study of infectious diseases and vitamin A deficiency in Indonesian children.

Key words and phrases: Generalized linear model, generalized estimating equations, seemingly unrelated regressions, feedback, logistic regression.

1. Introduction

In public health research, it is common to follow a subject over time, observing a vector of health indicators and a set of covariates at each of many visits. If a univariate response is the scientific focus, regression methods for time-dependent data (e.g. Laird and Ware (1982), Liang and Zeger (1986), Zeger (1988)) can be applied to characterize its mean as a function of the predictor variables. With a vector of responses, their inter-dependencies, in particular, the feedback of one variable upon another is typically important. When the vector can reasonably be assumed to follow a Gaussian distribution, multivariate autoregressive and moving average (ARMA) models can incorporate feedback and are in common use. See for example Tiao and Box (1981) for a discussion of ARMA models and Geweke (1982, 1984) for measures of feedback with ARMA processes. Feedback models have received far less attention, however, when the response vector includes both discrete and continuous variables as is common in public health research. The broad objective of this paper is to indicate how generalized estimating equations (Liang and Zeger (1986)) can be used to study feedback in a wide range of problems.

Before turning to specific statistical issues, we briefly consider a motivating example for which multivariate time series models with feedback may be important. It is from a study by Sommer et al. (1984) of pre-school Indonesian children who were medically examined quarterly for eighteen months. A study objective was to assess the role of vitamin A deficiency in children's morbidity. At each visit, it was determined whether a child had xerophthalmia, an ocular condition due to vitamin A deficiency, respiratory infection or diarrheal infection. A number of covariates such as age, weight and height were also determined. An interesting issue is whether there exists a feedback mechanism whereby vitamin A deficient children are more likely to suffer respiratory and diarrheal infections which in turn deplete stores of vitamin A and increase their risk of subsequent infections. This hypothesized relationship is feasible since vitamin A is necessary for maintaining epithelial cells, the first defense against infection. It has public health importance because respiratory and diarrheal infections are among the leading causes of children's mortality in developing countries and also since vitamin A deficiency can be prevented with supplementation programs. To investigate whether such a feedback exists, we must simultaneously model the conditional distribution of respiratory and diarrheal infections as a function of previous vitamin A deficiency as well as the conditional distribution of vitamin A deficiency in terms of previous infections while also adjusting for covariates.

Models for multivariate time series must necessarily originate from models for time-independent multivariate observations. There has been active development in both the statistics and econometrics literature in recent years of models for multivariate discrete response vectors. Goodman (1973), Nerlove and Press (1973), Schmidt and Strauss (1975) and Lee (1981) have formulated log-linear models for multivariate binary or categorical responses in terms of the conditional distributions for each response given the others. More recently, Dale (1986) and McCullagh and Nelder (1989) have parameterized the joint distribution of multivariate categorical data in terms of the marginal distributions of the individual responses as well as the higher order marginals.

There is also substantial literature on independent mixed continuous and categorical responses, much of it focused on the discrimination problem. Olkin and Tate (1961), Krzanowski (1982) and Little and Schluchter (1985) have assumed that conditional on the discrete variables, the continuous responses have a Gaussian distribution and that the marginal distribution of the discrete responses follows a log-linear model. Lauritzen and Wermuth (1989) have recently termed these CG ("Conditionally Gaussian") distributions and have discussed mixed variable analogues of graphical models for contingency tables. Probit-

Gaussian models have been preferred in the econometrics literature. See for example Heckman (1978).

This paper discusses time series analogues of the models above. A flexible class of feedback models can be obtained by specifying only the form of the conditional mean and variance of each response given a subset of the remaining responses at the same time and given the past. Reference to particular joint distributions will be necessary only to establish parameter constraints in some examples. Otherwise, the proposed methods are time series extensions of multivariate quasi-likelihood models (McCullagh and Nelder (1989)). The generalized estimating equations (GEE) approach (Liang and Zeger (1986), Prentice (1988)) will be used for estimating parameters while taking into account dependencies among the responses not explicitly modelled in the conditional means and variances.

The organization of this paper is as follows. Section 2.1 discusses a series of specific models for discrete and continuous multivariate time series; and Section 2.2 gives a general formulation in terms of conditional means and variances. The GEE approach will be briefly reviewed and extended to include parameter constraints in Section 3. The methods will be illustrated with an analysis of the Indonesian children's data in Section 4 followed by discussion.

2. Multivariate Time Series Models

To establish notation, let $y_t = (y_{1t}, \dots, y_{nt})'$ be an $n \times 1$ vector of responses at time t and define y_{-jt} to be the $(n-1) \times 1$ vector with the jth response left out. Let c_{jt} be a subset of y_{-jt} containing the responses to be explicitly used in predicting y_{jt}. Also let x_{jt} be the $p_j \times 1$ vector of covariates associated with y_{jt} and define $P_t = \{y_s, s < t\}$ to be the past outcomes. Finally let $\mu_{jt} = E(y_{jt}|c_{jt}, P_t)$ and $v_{jt} = \text{var}(y_{jt}|c_{jt}, P_t)$. Dependence of these moments on the covariates is obvious and is suppressed in the notation.

2.1. Examples

We now consider four specific examples of multivariate time series models which incorporate feedback. To simplify notation, we will restrict attention to the case where $y_t = (y_{1t}, y_{2t})'$ and to first order time dependence.

(i) *Multivariate autoregressive* (AR) *models*

In a multivariate regression with AR errors, it is assumed that conditioning on P_t, y_t is Gaussian with mean $\mu_t = (\mu_{1t}, \mu_{2t})'$ given by

$$\begin{aligned}\mu_{1t} &= E(y_{1t}|P_t) = x'_{1t}\beta_1 + \gamma_{11}(y_{1t-1} - x'_{1t-1}\beta_1) + \gamma_{12}(y_{2t-1} - x'_{2t-1}\beta_2)\\ \mu_{2t} &= E(y_{2t}|P_t) = x'_{2t}\beta_2 + \gamma_{21}(y_{1t-1} - x'_{1t-1}\beta_1) + \gamma_{22}(y_{2t-1} - x'_{2t-1}\beta_2)\end{aligned} \tag{2.1}$$

with $\mathrm{var}(y_t|P_t) = \Sigma$. Here c_{1t} and c_{2t} are taken to be null, that is, we have not conditioned y_{jt} on the other outcome at time t, only on the past. However, μ_{jt} does depend explicitly on y_{t-1} through the parameters γ allowing for feedback of each series on the other. See Tiao and Box (1981) for a detailed discussion of ARMA models like (2.1). In the econometrics literature, (2.1) is referred to as "seemingly unrelated regression equations" or SURE models (Zellner (1962)).

(ii) *Log-linear Markov models*

Suppose both y_{1t} and y_{2t} are binary (or more generally categorical) variables. A log-linear model for $Pr(y_t|P_t)$ can be obtained by letting $c_{jt} = y_{-jt}$, $j = 1,\ldots,n$ (=2 in this case) and specifying a logistic model for $\mu_{jt} = E(y_{jt}|c_{jt}, P_t)$. With two dichotomous outcomes, we have for example

$$\begin{aligned} \operatorname{logit} E(y_{1t}|y_{2t}, y_{t-1}) &= x_{1t}'\beta_1 + \delta_1 y_{2t} + \gamma_{11} y_{1t-1} + \gamma_{12} y_{2t-1} \\ \operatorname{logit} E(y_{2t}|y_{1t}, y_{t-1}) &= x_{2t}'\beta_2 + \delta_2 y_{1t} + \gamma_{21} y_{1t-1} + \gamma_{22} y_{2t-1} \end{aligned} \qquad (2.2)$$

and $v_{jt} = \mu_{jt}(1-\mu_{jt})$, $j = 1,2$. Liang and Zeger (1989) have shown that (2.2) uniquely determined a log-linear model for $Pr(y_t|P_t)$ subject to the constraint $\delta_1 = \delta_2$. Nerlove and Press (1972) and Lee (1981) discussed the analogous model with $\gamma_{ij} = 0$ for time-independent responses.

In (2.2), γ_{12} represents the feedback of y_{2t-1} on y_{1t} *conditional on* y_{2t}. When there is serial dependence within each series y_{jt}, the feedback of y_{2t-1} on y_{1t}, marginalized with respect to y_{2t}, may be of more scientific interest. We believe this is the case in the Indonesian children's example discussed in Section 4. The next example may be more appropriate in such cases.

(iii) *Simultaneous logistic regressions*

As an alternative to (2.2), let each c_{jt} be null leading to the specification

$$\begin{aligned} \operatorname{logit} E(y_{1t}|P_t) &= x_{1t}'\beta_1^* + \gamma_{11}^* y_{1t-1} + \gamma_{12}^* y_{2t-1} \\ \operatorname{logit} E(y_{2t}|P_t) &= x_{2t}'\beta_2^* + \gamma_{21}^* y_{1t-1} + \gamma_{22}^* y_{2t-1} \end{aligned} \qquad (2.3)$$

with $v_{jt} = \mu_{jt}^*(1-\mu_{jt}^*)$, $j = 1,2$. The starred notation in (2.3) is to accentuate that the parameters in (2.2) and (2.3) are distinct. For example, γ_{12}^* represents feedback of y_{2t-1} on y_{1t} averaged over the two possible states of y_{2t}. The model in (2.3) is a Markov analogue of the one discussed by Dale (1986), McCullagh and Nelder (1989) and Liang, Zeger and Qaqish (1989). Note that the conditional dependence between y_{1t} and y_{2t} given P_t is not explicitly modelled in (2.3). To completely specify $Pr(y_t|P_t)$, we must also model all higher order marginal distributions (second order in the case $n = 2$). For example, we might assume that the log odds ratio of y_{1t} and y_{2t} is a linear function of a subset of the explanatory variables with coefficients α. These "marginal" models for multivariate responses are discussed in detail by Liang, Zeger and Qaqish (1989).

Equation (2.3) involves two logistic models to be fit simultaneously along with a model for the covariance of the responses. However, the approach applies to any quasi-likelihood model for $Pr(y_{jt}|P_t)$. That is, we can simultaneously estimate a series of models for discrete and continuous responses in which no response is an explanatory variable in another equation but where the correlation among the outcomes is simultaneously modelled. This class is an extension of SURE models which might be referred to as "seemingly unrelated generalized regression equations" (SUGRE).

(iv) *Models for mixed binary and continuous responses*

Suppose now that y_{1t} is a continuous variable while y_{2t} is binary. Lauritzen and Wermuth (1989) have reviewed properties of the class of "conditionally Gaussian" or CG distributions where y_{1t} given y_{2t} is assumed to be Gaussian and the distribution of y_{2t} follows a log-linear model. One attractive feature of the CG class is that y_{2t} given y_{1t} is then also in the log-linear family. In particular, if y_{2t} is univariate, then assuming $E(y_{2t}|P_t)$ follows a logistic model implies that $E(y_{2t}|y_{1t}, P_t)$ also has a logistic form. Thus, two distinct models for mixed binary-Gaussian time series data are suggested. In the first, we let $c_{1t} = \{y_{2t}\}$ and c_{2t} be null. One such formulation in which all effects are assumed to be additive has the form

$$E(y_{1t}|y_{2t}, P_t) = x'_{1t}\beta_1 + \delta_1 y_{2t} + \gamma_{11}(y_{1t-1} - x'_{1t-1}\beta_1) + \gamma_{12}y_{2t-1} \tag{2.4a}$$

$$\text{logit}E(y_{2t}|P_t) = x'_{2t}\beta_2 + \gamma_{21}(y_{1t-1} - x'_{1t-1}\beta_1) + \gamma_{22}y_{2t-1} \tag{2.4b}$$

where $v_{1t} = \sigma^2$ and $v_{2t} = \mu_{2t}(1 - \mu_{2t})$. Models with interaction can similarly be defined. Here γ_{21} represents the feedback of y_{1t-1} on y_{2t} averaged over the distribution of y_{1t} given P_t. An alternate approach is to condition on y_{1t}, that is let $c_{2t} = \{y_{1t}\}$, and replace (2.4b) by

$$\begin{aligned}&\text{logit}E(y_{2t}|y_{1t}, P_t)\\ &\quad = x'_{2t}\beta^*_2 + \delta_2(y_{1t} - x'_{1t}\beta_1) + \gamma^*_{21}(y_{1t-1} - x'_{1t-1}\beta_1) + \gamma^*_{22}y_{2t-1}.\end{aligned} \tag{2.4c}$$

In this case, δ_2 and δ_1 must satisfy the constraint $\delta_2 = \sigma^{-2}\delta_1$.

2.2. General formulation

In each of the models above, a "link" function h_j of the conditional mean μ_{jt} is assumed to be linear in the covariate x_{jt} and in the elements of c_{jt} and P_t.

The conditional variance is also assumed to be a known function g of the mean. The general formulation has the following assumptions:

$$\begin{aligned} h_j(\mu_{jt}) &= x'_{jt}\beta_j + \xi_j(c_{jt}, P_t; \delta_j, \gamma_j) \\ v_{jt} &= g_j(\mu_{jt}) \cdot \phi_j, \quad j = 1, \dots, n, \end{aligned} \tag{2.5}$$

where h_j, g_j and ξ_j are known functions; β_j, δ_j, γ_j are parameters characterizing the dependence of y_{jt} on x_{jt}, c_{jt} and P_t respectively and ϕ_j are nuisance scale parameters. Note that ξ_j may also depend on previous values of the covariate, x_{jt-1} as in Example (i). Here the ξ_j's depend on $\beta = (\beta'_1, \dots, \beta'_n)'$. This requires slightly special treatment in model fitting as is described in Zeger and Qaqish (1988). With a linear model, it is preferable in our opinion to use the residual $(y_{it-1} - x'_{it-1}\beta)$ instead of y_{it-1} as a predictor as in (2.1). With binary responses, the prior responses y_{it-1} themselves are used in (2.2) to (2.4). In non-linear models, the definition of a residual is ambiguous. Furthermore, with discrete responses, the occurrence of the event may be of greater importance than a relative measure which includes how likely the event was. Finally, with responses of low prevalence, the results will be similar whether or not a residual is used.

The choice of c_{jt}'s is largely determined by the scientific objectives of the study. This choice however can place constraints on the forms of ξ_j which are consistent with the existence of a joint distribution $Pr(y_t|P_t)$. When the c_{jt}'s are null sets as in Examples (i) and (iii), no constraints are necessary except possibly to guarantee ergodicity of the process. More generally constraints are unnecessary if the c_{jt}'s are nested as in Example (iv), Equations 2.4a-b. In this case, only the conditional moments in (2.5) and not the entire joint distribution $Pr(y_t|P_t)$ need be specified. However, when the c_{jt}'s are not nested as in Examples (ii) and (iv), Equations 2.4a and 2.4c, the entire joint distribution $Pr(y_t|P_t)$ is necessary to determine the constraints on the ξ_j's. These are given by the Hammersley-Clifford Theorem (Besag (1974)).

3. Estimation

This section discusses an extension of the generalized estimating equation (GEE) approach to estimating the parameters $\theta' = \{(\beta'_j, \delta'_j, \gamma'_j),\ j = 1, \dots, n\}$ in Equation (2.5) when θ may be on a lower dimensional subspace due to constraints. See Liang and Zeger (1986) and Prentice (1988) for additional discussion of the unconstrained case. Let $S_{jt} = v_{jt}^{-1/2}(y_{jt} - \mu_{jt})$ and $S_t = (S_{1t}, \dots, S_{nt})'$. We will suppose $\mathrm{Cov}(S_t|P_t) = R(\mu_t, \alpha)$ where α is an $a \times 1$ vector of unknown parameters. Note in examples where the c_{jt}'s are nested, $R(\mu_t, \alpha) = I$ since the μ_{jt}'s are conditional expectations. The GEE approach is a multivariate analogue of quasi-likelihood (McCullagh and Nelder (1989)). Define

$$U_{\theta t} = \frac{\partial \mu_t'}{\partial \theta} V_t^{-1}(y_t - \mu_t),$$

where V_t is the working covariance matrix for y_t. In the cases such as Example (ii) above when θ is linearly constrained, there exists a vector θ^* of unconstrained coefficients given by $\theta^* = L(\phi)\theta$. In Example (ii), L is known; in Example (iv), L depends on the unknown scale parameters. By the chain rule, the estimating equations for θ^* at time t are simply $U_{\theta^* t} = L(\phi)U_{\theta t}$. When L depends on an unknown ϕ, we use $\hat{\phi}$ defined below.

To complete the specification, we require estimating equations for α and ϕ. Moment estimators are adequate with larger data sets common in public health. Let $Z_t = (S_{1t}S_{2t}, \ldots, S_{1t}S_{nt}, S_{2t}S_{3t}, \ldots, S_{n-1t}S_{nt})'$ be the vector of cross-products and denote $E(Z_t|P_t) = \eta_t(\alpha)$. Following Prentice (1988), the equation for α is

$$U_{\alpha t} = \frac{\partial \eta_t'}{\partial \alpha} W_t^{-1}(Z_t - \eta_t)$$

where W_t is a weighting matrix that in most situations can be chosen as the identity matrix. The scale coefficients $\hat{\phi}_j$ can be obtained from the moment equations

$$U_{\phi j} = \sum_t \left(S_{jt}^2/g(\mu_{jt}) - \phi_j\right) = 0, \quad j = 1, \ldots, n.$$

Assembling the components and assuming $t = 1, \ldots, T$, we have

$$U = \begin{pmatrix} \sum_t U_{\theta^* t} \\ \sum_t U_{\alpha t} \\ \sum_t U_{\phi t} \end{pmatrix} = \begin{pmatrix} L(\hat{\phi}) \sum_t \frac{\partial \mu_t'}{\partial \theta} V_t^{-1}(y_t - \mu_t) \\ \sum_t \frac{\partial \eta_t'}{\partial \alpha} W_t^{-1}(Z_t - \eta_t) \\ \sum_t \left(S_t^2/g(\mu_t) - \phi\right) \end{pmatrix} = 0. \qquad (3.1)$$

Let $\hat{\theta}$ be the solution of (3.1). Then under the usual Craḿer regularity conditions, $\sqrt{T}(\hat{\theta} - \theta)$ converges to a multivariate mean zero Gaussian vector with variance matrix $V_{\hat{\theta}} = (\frac{\partial U}{\partial \theta})^{-1} \mathrm{Var}(U) (\frac{\partial U}{\partial \theta})^{-1}$ which can be consistently estimated by substituting $\sum_t U_t U_t'$ for Var(U).

This asymptotic distribution can be derived in the same fashion as in Liang and Zeger (1986) and is therefore omitted.

4. An Example

Data from the Indonesian children's study (Sommer, et al. (1984, 1987)) illustrate the use of GEE for feedback models. Approximately 3,000 pre-school children were medically examined quarterly for up to seven visits. The possible responses include presence/absence at the medical visit of diarrheal infection D_t; respiratory infection R_t; xerophthalmia Z_t, an ocular sign of vitamin A deficiency; as well as weight for height W_t, a surrogate for recent nutritional status. The covariates include: age in months A_t (centered at 36 months); sex S; seasonal sine $\sin_t$ and cosine $\cos_t$ and height for age H_t as a percent of NCHS standard (centered at 90%). The first scientific questions of interest are: (i) are children who are vitamin A deficient as indicated by xerophthalmia more susceptible to infectious diseases; and (ii) do respiratory and diarrheal diseases increase the risk of future vitamin A deficiency.

Based upon extensive preliminary analysis of these data (Sommer, Katz and Tarwotjo (1984, 1987)) the following multivariate regression model was chosen:

$$\begin{aligned}
\operatorname{logit} Pr(D_t = 1|Z_t, P_t) &= \beta_{10} + \beta_{11}Z_t + \beta_{12}A_t + \beta_{13}H_t + \beta_{14}\sin_t + \beta_{15}\cos_t \\
&\quad + \beta_{16}S + \gamma_{11}D_{t-1} + \gamma_{12}R_{t-1} + \gamma_{13}Z_{t-1} + \gamma_{14}Z_t \cdot Z_{t-1} \\
\operatorname{logit} Pr(R_t = 1|Z_t, P_t) &= \beta_{20} + \beta_{21}Z_t + \beta_{22}A_t + \beta_{23}H_t + \beta_{24}\sin_t + \beta_{25}\cos_t \\
&\quad + \beta_{26}S + \gamma_{21}D_{t-1} + \gamma_{22}R_{t-1} + \gamma_{23}Z_{t-1} + \gamma_{24}Z_t \cdot Z_{t-1} \\
\operatorname{logit} Pr(Z_t = 1|P_t) &= \beta_{30} + \beta_{31}A_t + \beta_{32}H_t + \beta_{33}S + \beta_{34}\sin_t + \beta_{35}\cos_t \\
&\quad + \gamma_{31}D_{t-1} + \gamma_{32}R_{t-1} + \gamma_{33}Z_{t-1} + \gamma_{34}Z_{t-1} \cdot D_{t-1} + \gamma_{35}Z_{t-1}R_{t-1}.
\end{aligned} \tag{4.1}$$

Note we have chosen to model the conditional distributions $Pr(D_t|Z_t)$, $Pr(R_t|Z_t)$ and the marginal distribution $Pr(Z_t)$ supressing for the moment conditioning on the covariates X_t and the past P_t. Vitamin A deficiency may affect a child's propensity for both infections. If in studying the impact of vitamin A deficiency on respiratory infection R_t, we condition on the presence of diarrheal infection D_t, some of the vitamin A impact on R_t will be incorrectly attributed to D_t. Similarly it is more appropriate to estimate $Pr(D_t|Z_t)$ than $Pr(D_t|R_t, Z_t)$. However, in the other direction, we focus on $Pr(Z_t)$ rather than $Pr(Z_t|R_t, D_t)$. Xerophthalmia signs of vitamin A deficiency take some time to develop since longer-term supplies of vitamin A can be stored in the liver. The effect of infections might be to reduce stores but resulting ocular signs are unlikely to occur immediately. Hence we estimate $Pr(Z_t|D_{t-1}, R_{t-1})$ rather than $Pr(Z_t|D_t, R_t)$. Sommer (1982) gives a detailed discussion of xerophthalmia and its association with infections.

Ignoring covariates and past outcomes, our model for (D_t, R_t, X_t) is distinct from more traditional log-linear models (Bishop, Fienberg and Holland (1975)). The parameters in a log-linear model would have interpretation in terms of the conditional probabilities $Pr(D_t|R_t, Z_t)$, $Pr(R_t|D_t, Z_t)$, and $Pr(Z_t|D_t, R_t)$.

Time series models for these conditional distributions are discussed by Liang and Zeger (1989). For the Indonesian data, we believe the conditional probabilities are of less scientific interest than $Pr(D_t|Z_t)$, $Pr(R_t|Z_t)$ and $Pr(Z_t)$ for the reasons above. A more detailed discussion of "marginal" model alternatives to log-linear models can be found in McCullagh and Nelder (1989) and Liang, Zeger and Qaqish (1989).

The results of fitting Model (4.1) using generalized estimation equations (GEE) are presented in Table 1. The working correlation assumption is that only D_t and R_t are associated since Z_t appears explicitly in the D_t and R_t regressions. Hence the GEE reduces to fitting one model for the bivariate response (D_t, R_t) and a second ordinary logistic regression (OLR) with Z_t as a response. Table 1 also lists the results of separate OLR's for D_t and R_t.

First note that the GEE and OLR results are nearly identical. The estimated correlation between D_t and R_t was 0.020, too small to create a substantial advantage for GEE. Note, however, a correlation of 0.020 corresponds to an odds ratio of 1.36 at the mean propensities for D_t and R_t of .043 and .091 respectively. That is, a typical child with diarrheal disease at a visit is roughly 1.36 times as likely to have respiratory infection as well relative to the child without diarrhea. This is one distinction of models for multivariate binary as opposed to Gaussian data: efficiency gains by accounting for even scientifically important associations are more limited.

These data are suggestive that a feedback system is operating. First both infections are more common among children with xerophthalmia. The xerophthalmia odds ratios for diarrhea and respiratory infections are estimated to be 4.5 and 1.6 respectively; both are highly significant. Given Z_t, previous xerophthalmia (Z_{t-1}) and the interaction $Z_t \cdot Z_{t-1}$ are not important in either regression. Among the other covariates, age, sex, height for age and seasonality are important predictors of both diarrheal and respiratory infection. Turning to xerophthalmia as an outcome, the strongest predictor is previous xerophthalmia with an odds ratio of 33 (95% confidence interval: 26 to 41). However among children without xerophthalmia at the previous visit ($Z_{t-1} = 0$) those with diarrhea at ($D_{t-1} = 1$) are 1.63 times more likely to become xerophthalmic (95% interval: 1.03 to 2.58) at visit t. Hence xerophthalmia increases the risk of diarrheal disease which further increases the risk of future xerophthalmia in a positive feedback cycle.

A second analysis using the responses weight for height and respiratory disease has been performed to illustrate a feedback model with continuous and discrete outcomes. Weight for height was not included along with the three discrete variables in (4.1) because the relationship between diarrhea and weight for height may reflect short-term effects of dehydration rather than a nutritional status-disease relationship. The questions here are whether children with respiratory disease tend to lose weight for height in the near future and whether they are then at increased risk of additional respiratory disease. Based upon preliminary analyses, we have chosen regression models of the form:

$$\begin{aligned}\operatorname{logit} Pr(R_t = 1|W_t, P_t) &= \beta_{10} + \beta_{11}A_t + \beta_{12}H_t + \beta_{13}\sin_t + \beta_{14}\cos_t \\ &\quad + \gamma_{11}W_t + \gamma_{12}R_{t-1} + \gamma_{13}W_{t-1} + \gamma_{14}W_{t-1}R_{t-1} \\ E(W_t|R_t, P_t) &= \beta_{20} + \beta_{21}A_t + \beta_{22}H_t + \beta_{23}\sin_t + \beta_{24}\cos_t + \gamma_{21}R_t \\ &\quad + \gamma_{22}R_{t-1} + \gamma_{23}R_tR_{t-1} + \gamma_{24}(W_{t-1} - E(W_{t-1}|R_{t-1}, P_{t-1})).\end{aligned} \tag{4.2}$$

Note we have modelled $E(R_t|W_t, P_t)$ and $E(W_t|R_t, P_t)$ as described in Equations (2.4a,c). This dictates the constraint $\gamma_{11} = \sigma^{-2}\gamma_{21}$ where σ^2 is the residual variance from the linear regression.

The results are presented in Table 2. In the linear regression for weight for height, there is strong first-lag autocorrelation of 0.64 indicating that poor nutritional status persists over time. Using the group with no respiratory infection at the current or previous visit ($(R_{t-1}, R_t) = (0, 0)$) as the reference, there is a decrease in expected weight for height of 1.9% in children who had infection at the same visit but not the last (0,1) and of 2.4% if they had respiratory disease at the previous and current visits (1,1). Height for age and season are also strongly associated with weight for height.

In the logistic regression for respiratory disease, there is again a strong negative association with current weight for height reflecting the constraint. That is, children who currently have poorer weight for height are at higher risk of having respiratory infection. Having respiratory infection last visit is associated with a 1.73 times increase in risk for the current visit. Among children without infection last visit and with the same current weight for height, previous weight for height is not a predictor of initiating an infection ($\hat{\beta}$/s.e.=.81). But for the subgroup who did have infection last time, those that lost ground in weight for height (higher at last visit) had a substantially increased risk of infection. For example a loss of 10% in weight for height is associated with a 44% increase in risk of current infection. Among the covariates, age, sex, height for age and season are all strong predictors of respiratory infection.

5. Discussion

The issue on how to deal with the problem of feedback has been studied extensively in the literature of economics. Although this problem is also common, as illustrated in Section 1, in biomedical and public health studies, much less attention has been paid in the past. One important feature of feedback problems in public health research is that responses include both discrete and continuous variables and are observed over shorter time periods for many subjects. In this paper, a general Markov model that incorporates feedback was proposed. It can be estimated using the method of generalized estimating equations. There are a few important properties of the suggested approach that deserve mentioning.

(1) In many problems, the entire joint distribution of the vector of responses need not be specified. Only the means and variances are modelled. Thus this approach can be viewed as a multivariate time series analogue of the quasi-likelihood method.

(2) The method has the flexibility to allow the investigators to decide, on the basis of scientific interest, which subset c_{jt} of y_{jt} should be included in the model as predictors.

(3) We have modelled the distributions of observable quantities rather than of latent variables.

In the Indonesian example, we have modelled the conditional distributions of infections given xerophthalmia. It might be more natural to condition on vitamin A level and to treat xerophthalmia as a surrogate for vitamin A. Such an analysis would be more realistic but also much more difficult to perform.

Acknowledgements

The authors thank Alfred Sommer for making the Indonesian data available, Joanne Katz for many helpful discussions about feedback models, M. Rezaul Karim for assistance in computing and the editors and referees for insightful comments on an earlier draft. This work originated and was partly motivated by our participation in the Multicenter AIDS Cohort Study, NIH grant N01-AI-72676. Support from NIH grants R29-AI25529-02 and R29-GM39262-02 is gratefully acknowledged. This work was conducted while Scott L. Zeger was a visitor at the Medical Research Council, Cambridge, England made possible by a travel grant from the Burroughs Wellcome Foundation.

Table 1. Regression results for model (4.1) fitted by generalized estimation equations (GEE) and as three separate ordinary logistic regressions (OLR). Because the diarrhea and respiratory infection regressions condition on current xerophthalmia, the xerophthalmia regression by GEE and OLS are identical.

	Response									
	Diarrhea (t)				Respiratory (t)				Xerophthalmia (t)	
	OLS		GEE		OLS		GEE		OLS/GEE	
Explanatory variable	$\hat{\beta}$	$\hat{\beta}/\hat{s.e.}$	$\hat{\beta}$	$\hat{\beta}/\hat{s.e.}$	$\hat{\beta}$	$\hat{\beta}/\hat{s.e.}$	$\hat{\beta}$	$\hat{\beta}/\hat{s.e.}$	$\hat{\beta}$	$\hat{\beta}/\hat{s.e.}$
Intercept	−3.1	−50	−3.1	−50	−2.3	−51	−2.3	−51	−3.5	−47
Age	−.030	−14	−.030	−14	−.028	−16	−.028	−16	.0043	1.8
Sex	−.15	−2.1	−.15	−2.1	−.33	−5.5	−.33	−5.5	−6.2	−6.5
Height/Age	−.050	−6.6	−.050	−6.7	−.048	−8.1	−.049	−7.6	−.073	−8.2
Seasonal Sine	−.64	−11	−.64	−11	.18	4.2	.18	4.1	−.033	−.50
Seasonal Cosine	.35	6.3	.35	6.6	.014	.34	.014	.33	−.047	−.70
Diarrhea ($t-1$)	.97	7.8	.96	7.8	.24	2.0	.24	2.0	.49	2.14
Respiratory ($t-1$)	.048	.40	.050	.40	.64	7.7	.64	7.6	.17	.95
Xerophthalmia (t)	1.5	10	1.5	10	.51	3.1	.51	3.1	-	-
Xerophthalmia ($t-1$)	−.30	−1.0	−.30	−1.0	.17	.88	.17	.89	3.5	33
Diarrhea ($t-1$)* xerophthalmia ($t-1$)	-	-	-	-	-	-	-	-	−.64	−1.9
Respiratory ($t-1$)* xerophthalmia ($t-1$)	-	-	-	-	-	-	-	-	−.36	−1.2
Xerophthalmia (t)* xerophthalmia ($t-1$)	−.27	−.73	−.27	−.73	−.23	−.76	−.23	−.77	-	-

Table 2. Regression results for a feedback model of respiratory disease and weight for height for the Indonesian children's data as described in Section 4

	Response variable			
	Respiratory(t)		Weight/Height(t)	
Explanatory variable	$\hat{\beta}$	$\hat{\beta}/\hat{s.e.}$	$\hat{\beta}$	$\hat{\beta}/\hat{s.e.}$
Age	−.0240	−13	.00772	1.1
Sex	−.281	−4.1	.211	.83
Height/Age	−.0455	−6.2	−.142	−4.0
Seasonal Sine	.24	4.9	.272	4.0
Seasonal Cosine	−.0194	−.42	−.582	−8.6
Respiratory(t)	—	—	−1.89	−8.6
Respiratory($t-1$)	.549	5.3	−.507	−2.2
Weight/Height(t)	−.0291	−8.6	—	—
Weight/Height($t-1$)	.00318	.81	.637	—
Weight/Height($t-1$)* Respiratory($t-1$)	.0339	3.4	—	—

References

Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems (with discussion). *J. Roy. Statist. Soc. Ser. B* **36**, 192-236.

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

Dale, J. R. (1986). Global cross-ratio models for bivariate, discrete ordered responses. *Biometrics* **42**, 909-917.

Geweke, J. F. (1982). Measurement of linear dependence and feedback between multiple time series. *J. Amer. Statist. Assoc.* **77**, 304-313.

Geweke, J. F. (1984). Measures of conditional linear dependence and feedback between time series. *J. Amer. Statist. Assoc.* **79**, 907-915.

Goodman, L. A. (1973). The analysis of multidimensional contingency tables when some variables are posterior to others: a modified path analysis approach. *Biometrika* **46**, 179-192.

Heckman, J. J. (1978). Dummy endogenous variables in a simultaneous equation system. *Econometrica* **60**, 931-959.

Krzanowski, W. J. (1982). Mixtures of continuous and categorical variables in discriminant analysis: a hypothesis testing approach. *Biometrics* **38**, 991-1002.

Laird, N. M. and Ware, J. M. (1982). Random effects models for longitudinal data. *Biometrics* **38**, 963-974.

Lauritzen, S. L. and Wermuth, N. (1989). Graphical models for associations between variables, some of which are qualitative and some quantitative. *Ann. Statist.* **17**, 31-57.

Lee, L.-F. (1981). Fully recursive probability models and multivariate log-linear probability, models for the analysis of qualitative data. *J. Econometrics* **16**, 51-69.

Liang, K.-Y. and Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika* **73**, 13-22.

Liang, K.-Y., and Zeger, S. L. (1989). A class of logistic regression models for multivariate binary time series. *J. Amer. Statist. Assoc.* **84**, 447-451.

Liang, K.-Y., Zeger, S. L. and Qaqish, B. (1989). Multivariate regression models using generalized estimating equations. Technical Report No. 700, Department of Biostatistics, Johns Hopkins University, Baltimore, Maryland.

Little, R. J. A. and Schluchter, M. D. (1985). Maximum likelihood estimation for mixed continuous and categorical data with missing values. *Biometrika* **72**, 497-512.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*, 2nd edition. Chapman and Hall, London.

McCullagh, P. (1989). Models for discrete multivariate responses. Technical Report #253, Department of Statistics, University of Chicago.

Nerlove, M. and Press, S. J. (1973). Univariate and multivariate log linear and logistic models. RAND Corporation Report R-1306, Santa Monica, California.

Olkin, I. and Tate, R. F. (1961). Multivariate correlation models with mixed discrete and continuous variables. *Ann. Math. Statist.* **32**, 448-465.

Prentice, R. L. (1988). Correlated binary regression with covariates specific to each binary observation. *Biometrics* 44, 1033-1048.

Schmidt, P. and Strauss, R. P. (1975). Estimation of models with jointly dependent qualitative variables: a simultaneous logit approach. *Econometrica* **43**, 745-754.

Sommer, A. (1982). *Nutritional Blindness*, Chapter 8. Oxford University Press, Oxford, England.

Sommer, A., Katz, J. and Tarwotjo, I. (1984). Increased risk of respiratory infection and diarrhea in children with pre-existing mild vitamin A deficiency. *Amer. J. Clinical Nutrition* **40**, 1090-1095.

Sommer, A., Katz, J. and Tarwotjo, I. (1987). Increased risk of xerophthalmia following diarrhea and respiratory disease. *Amer. J. Clinical Nutrition* **45**, 977-980.

Tiao, G. C. and Box, G. E. P. (1981). Modeling multiple time series with applications. *J. Amer. Statist. Assoc.* **76**, 802-816.

Zeger, S. L. (1988). A regression model for time series of counts. *Biometrika* **75**, 621-629.

Zeger, S. L. and Qaqish, B. (1988). Markov regression models for time series: a quasi-likelihood approach. *Biometrics* **44**, 1019-1031.

Zellner, A. (1962). An efficient method for estimating seemingly unrelated regressions and tests for aggregation bias. *J. Amer. Statist. Assoc.* **57**, 348-368.

Department of Biostatistics, The Johns Hopkins University, Baltimore, Maryland 21205, U.S.A.

(Received September 1989; accepted June 1990)