Biometrika (1985), **72**, 1, pp. 31–8
Printed in Great Britain

The analysis of binary longitudinal data with time-independent covariates

BY SCOTT L. ZEGER, KUNG-YEE LIANG AND STEVEN G. SELF
Department of Biostatistics, Johns Hopkins University, Baltimore Maryland 21205, U.S.A.

SUMMARY

This paper considers extensions of logistic regression to the case where the binary outcome variable is observed repeatedly for each subject. We propose two working models that lead to consistent estimates of the regression parameters and of their variances under mild assumptions about the time dependence within each subject's data. The efficiency of the proposed estimators is examined. An analysis of stress in mothers with infants is presented to illustrate the proposed method.

Some key words: Binary longitudinal data; Logistic regression; Markov chain; Maximum likelihood.

1. INTRODUCTION

Data sets comprised of a short binary time series and a set of time independent covariates for each of many subjects are common in applied sciences. For example, the monthly presence, 1, or absence, 0, of a disease for members of a health plan as well as the members' age, sex and type of medical coverage might be observed over a year. The relationship of an individual's propensity for illness with the covariates is of interest. Given a single binary response for each subject, logistic regression (Cox, 1958, 1970) could be used to assess this relationship. With time series, however, methods that account for time autocorrelation are necessary. In this paper, we propose such methods for the case where the binary series are stationary. As our interest is in propensity, we propose models that lead to estimators that are robust to misspecification of the time dependence.

also E( )

To establish notation, let Y_{it} $(t = 1, \ldots, n_i)$ be the stationary time series and let Z_i be an $s \times 1$ vector of covariates for the ith subject $(i = 1, \ldots, K)$. If $\pi_i \equiv \mathrm{pr}(Y_{it} = 1 \mid Z_i)$, we assume that $\mathrm{logit}(\pi_i) = Z_i'\beta$. Our objective is to estimate the vector of parameters, β.

Cox (1970, Ch. 5) proposed one extension of logistic regression in which the conditional rather than marginal probabilities of a Markov chain are expressed as logistic functions of the covariates. Muenz & Rubinstein (1984) and Korn & Whittemore (1979) have applied this method to the analysis of many binary series. However, unlike the Gaussian case, models for the conditional expectation of a binary variable do not imply equally simple expressions for the marginal expectation i.e. propensity. In addition, the expression for propensity as derived from a specific model for the conditional expectation typically depends in a complicated way on the form of time dependence in the series. Thus, Cox's model leads to propensity estimates which depend strongly on the specification of the time dependence e.g. choice of Markov chain order. When the covariates are categorical, an alternative approach is followed by Grizzle, Starmer & Koch (1969) and Koch et al. (1977) who propose repeated measures models to account for time dependence in dichotomous data.

In this paper, we consider two models in which the marginal rather than conditional probabilities are expressed as logistic functions of the covariates. Our approach is to approximate the actual likelihood with working likelihoods that lead to consistent estimates of β under weak assumptions. We use time series rather than repeated measures models to account for the time dependence. In §2, we introduce the working models, propose estimators of β that are robust to misspecification of the time series structure and give their asymptotic distributions. In §3, we consider the relative efficiencies of these estimators for a few simple situations. Section 4 presents a consistent variance estimator. Section 5 gives an example of the use of the proposed methods. The final section compares our methods with alternatives and discusses possible extensions of this work.

2. WORKING LIKELIHOOD ESTIMATORS OF β

In this section we present two estimators of β. The first, $\hat{\beta}_0$, is derived from a working assumption that repeated observations for a subject are independent of one another. This estimator is consistent given any set of stationary binary processes such that $\operatorname{logit}(\pi_i) = Z_i'\beta$. The second estimator, $\hat{\beta}_1$, is obtained from the working model that each series is a stationary Markov chain of order one. It is consistent when the series satisfy the logit link assumption and have a common first lag autocorrelation, $\rho = \operatorname{corr}(Y_{it}, Y_{i,t-1})$. Both estimators can be derived from a framework such as that of Azzalini (1983) although this approach tends to obscure the connection with logistic regression. We therefore consider each estimator in turn and then afterwards mention the general case.

Under the assumption of time independence, the standard likelihood analysis of the logistic regression model as discussed by Cox (1970) is appropriate. Let $S_u(\beta)$ be the score equation for β_u $(u = 1, \ldots, s)$ and let $I_0(\beta)$ be the Fisher information matrix for this independence working model. Then the estimator, $\hat{\beta}_0$, is consistent and as $K \to \infty$ asymptotically Gaussian for any set of stationary processes such that $\operatorname{logit}(\pi_i) = Z_i'\beta$ as stated in the following result.

PROPOSITION 2·1. *Let Y_{it} $(t = 1, \ldots, n_i < \infty; i = 1, \ldots, K)$ be stationary binary series such that* $\operatorname{logit} E(Y_{it} \mid Z_i') = Z_i'\beta$. *Then $K^{\frac{1}{2}}(\hat{\beta}_0 - \beta)$ is asymptotically multivariate Gaussian with expectation 0 and covariance matrix $V_0 = I_0^{-1} I_0^* I_0^{-1}$, where I_0 is the working Fisher information matrix and $I_0^* = E(SS')$.*

The proof follows directly from results of Huber (1967) under mild regularity conditions. Note that the expectation necessary to obtain I_0^* is with respect to the actual model and hence I_0^* and V_0 depend upon the actual time dependence. In §4, a consistent estimator of V_0 is presented.

The estimator, $\hat{\beta}_1$, is obtained by assuming that each binary series is a realization of a stationary Markov chain for which

$$\operatorname{logit}(\pi_i) = Z_i'\beta, \quad \operatorname{corr}(Y_{it}, Y_{i,t-1} \mid Z_i) = \rho. \tag{2·1}$$

That is, subjects are assumed to have a common lag one autocorrelation. Under this first-order working model, the log likelihood is

glm

$$l_1(\beta, \rho) = \sum_{i=1}^{K} \left[y_{i1} Z_i'\beta \underbrace{- \log(1 + e^{Z_i\beta})}_{+\log(1-\pi)} + \underbrace{\sum_{t=2}^{n_i} \{y_{it} \log \pi_{it} + (1 - y_{it}) \log(1 - \pi_{it})\}}_{\text{after } i=1} \right], \tag{2·2}$$

for exp family

baseline

where

$$\pi_{it} = E(Y_{it} \mid Y_{i,t-1}, Z_i) = \pi_i + \rho(Y_{i,t-1} - \pi_i).$$

As a referee has pointed out, the two-dimensional parameter space for π and ρ is constrained to satisfy $\max(-\pi/\bar{\pi}, -\bar{\pi}/\pi) \leqslant \rho \leqslant 1$, where $\bar{\pi} = 1 - \pi$. An alternative unconstrained set of parameters is obtained by replacing ρ with the log odds ratio, $\log \psi$, of the first-order transition matrix. We have chosen π and ρ for two reasons: the log likelihood and score equations are much simpler with this parameterization; and in practice, we most often encounter positive values of ρ for which π is unconstrained.

The score equations for β and ρ have the form, for $u = 1, \ldots, s$,

$$\left.\begin{aligned} 0 = S_u(\beta, \rho) = \partial l_1/\partial \beta_u &= \sum_{i=1}^{K} \left\{ Z_{iu}(y_{i1} - \pi_i) + (1-\rho) \sum_{t=2}^{n_i} \frac{(Z_{iu}\pi_i\bar{\pi}_i)(y_{it} - \pi_{it})}{\pi_{it}\bar{\pi}_{it}} \right\}, \\ 0 = S_{s+1}(\beta, \rho) = \frac{\partial l_1}{\partial \rho} &= \sum_{i=1}^{K} \sum_{t=2}^{n_i} \frac{(y_{it} - \pi_{it})(y_{i,t-1} - \pi_i)}{\pi_{it}\bar{\pi}_{it}}. \end{aligned}\right\} \qquad (2{\cdot}3)$$

The Fisher information matrix, I_1, for this working model can be partitioned

$\frac{\partial l_1^{2}}{\partial\partial}$

$$I_1 = \begin{bmatrix} I_{\beta\beta} & I_{\beta\rho} \\ I_{\rho\beta} & I_{\rho\rho} \end{bmatrix}, \qquad (2{\cdot}4)$$

where the (u, v)th element of $I_{\beta\beta}$ is

$$\sum_{i=1}^{K} Z_{iu} Z_{iv} \pi_i \bar{\pi}_i \left\{ 1 + (n_i - 1)(1-\rho) \frac{\pi_i\bar{\pi}_i(1-3\rho)+\rho}{(1-\rho)^2 \pi_i \bar{\pi}_i + \rho} \right\},$$

and the uth element of $I_{\beta\rho}$ is

$$\sum_{i=1}^{K} (n_i - 1) Z_{iu} \pi_i \bar{\pi}_i \left(\frac{\pi_i}{\pi_i + \rho\bar{\pi}_i} - \frac{\bar{\pi}_i}{\bar{\pi}_i + \rho\pi_i} \right),$$

and where

$$I_{\rho\rho} = \sum_{i=1}^{K} (n_i - 1) \frac{1+\rho}{1-\rho} \left(\frac{\pi_i \bar{\pi}_i}{(1-\rho)^2 \pi_i \bar{\pi}_i + \rho} \right).$$

We define $\hat{\beta}_1$ as the solution of the system of equations (2·3) and have the following result.

PROPOSITION 2·2. *Let Y_{it} $(t = 1, \ldots, n_i < \infty; i = 1, \ldots, K)$ be stationary binary series such that* $\operatorname{logit} E(Y_{it} \mid Z_i) = Z_i'\beta$ *and* $\operatorname{corr}(Y_{it}, Y_{i,t-1}) = \rho$. *Then $K^{\frac{1}{2}}(\hat{\beta}_1 - \beta)$ is as $K \to \infty$ asymptotically Gaussian with expectation 0 and covariance matrix $V_1 = I_1^{-1} I_1^{*} I_1^{-1}$, where I_1 is the working Fisher information matrix given by* (2·4) *and* $I_1^{*} = E(SS')$.

The proof is similar to that for Proposition 2·1. Note again that the expectation used to calculate I_1 and I_1^{*} are with respect to the actual rather than working model. A method for obtaining a consistent estimate of V_1 under assumption (2·1) is described in §4.

The working likelihoods for each subject used to define $\hat{\beta}_0$ and $\hat{\beta}_1$ above are special cases of the 'nth-order likelihoods' considered by Azzalini (1983) and Ogata (1980). In the nth-order case, the true log likelihood for the ith subject, $l_i(\theta)$, is approximated by

$$l_i^{(n)}(\theta) = \sum_{t=1}^{n_i} \log f(y_{it} \mid y_{i,t-1}, \ldots, y_{i,t-n}; \theta).$$

Using an nth-order approximation for each subject's likelihood, we can obtain a general estimator, $\hat{\beta}_n$. However, its consistency requires us to assume that the nth-order conditional probabilities $E(Y_{it} \mid Y_{i,t-1}, \ldots, Y_{i,t-n}; Z_i)$ for different subjects have $2^n - 1$ parameters in common, differing only by the marginal expectations, π_i. This assumption is a strong one hence making lower order approximations more attractive.

3. EFFICIENCIES OF THE ESTIMATORS

To examine the relative efficiencies of $\hat{\beta}_0$ and $\hat{\beta}_1$, we have considered the case where subjects fall into one of two groups. We let π_i be the propensity of response for a subject in group i, and write

$$\operatorname{logit} \pi_i = \alpha + \beta i \quad (i = 0, 1).$$

Let n be the length of each binary series and λ be the proportion of subjects in group 0. Our interest is in the parameter β. We consider three cases for the actual time dependence where the binary time series for a subject is as follows: a sequence of independent observations; a realization of a stationary Markov process of order 1; and generated using a beta-binomial model. In the first case, $\hat{\beta}_0$ and $\hat{\beta}_1$ are both asymptotically efficient since $\hat{\beta}_0$ corresponds to the true maximum likelihood estimator and, at $\rho = 0$, the score for β in the first-order Markov working likelihood is orthogonal to the score for ρ. Hence additional comparison is unnecessary. In the second case, $\hat{\beta}_1$ is efficient and we consider the asymptotic relative efficiency of $\hat{\beta}_0$. In the third case, neither proposed estimator is efficient and we compare their asymptotic variances.

Table 1. *Asymptotic relative efficiencies of $\hat{\beta}_0$ to $\hat{\beta}_1$ when each series is a Markov chain of order $1, \alpha = \beta = 0$ and $\lambda = 0{\cdot}5$ for varying values of ρ and n*

n	$\rho = 0$	$\rho = 0{\cdot}1$	$\rho = 0{\cdot}3$	$\rho = 0{\cdot}5$	$\rho = 0{\cdot}7$	$\rho = 0{\cdot}9$
5	1·00	0·97	0·77	0·55	0·37	0·24
10	1·00	0·96	0·74	0·48	0·28	0·15

Table 1 lists asymptotic relative efficiencies of β_0 for a range of values of the lag one autocorrelation, ρ, when the binary series are first-order Markov and where $\alpha = \beta = 0$, $n = 5$ or 10, and $\lambda = \frac{1}{2}$. Note that for $\rho \leqslant \frac{1}{2}$, $\hat{\beta}_0$ is reasonably efficient in this two-sample problem. As it does not require the additional assumption that the correlation is the same across individuals, it might be preferred. As the correlation approaches 1 however, $\hat{\beta}_0$ becomes increasingly inefficient and the first-order Markov estimator, $\hat{\beta}_1$, is preferred. The asymptotic relative efficiency $\hat{\beta}_0$ decreases with increasing length of a series as is indicated in Table 1. It has little dependence on α, β or λ.

For the beta-binomial case, each series is a sequence of conditionally independent Bernoulli observations given the subject's propensity which itself is an observation from a beta distribution with expectation π and variance $\pi(1-\pi)\rho$. This variance was chosen

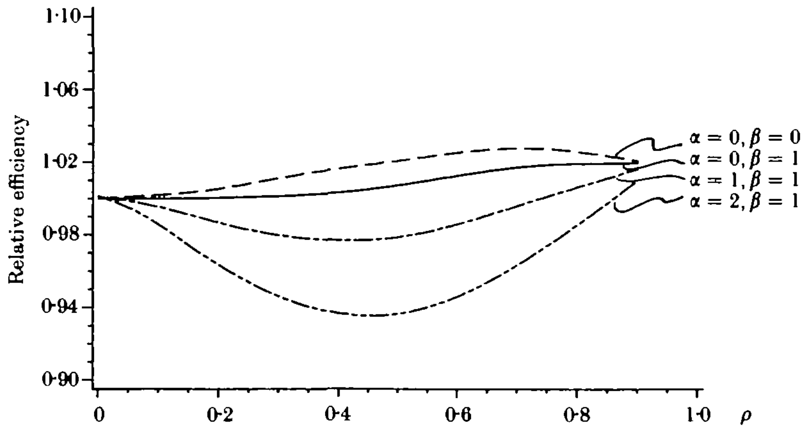

Fig. 1. Asymptotic efficiency of independence model estimator, $\hat{\beta}_0$, relative to first-order Markov estimator, $\hat{\beta}_1$, versus correlation, ρ, for two-sample problem. True model, beta-binomial.

so that the unconditional correlation for any two observations from the same subject is ρ. Figure 1 plots the ratio of the asymptotic variances of $\hat{\beta}_1$ and $\hat{\beta}_0$ against ρ for four cases with changing values of α and β. In general, there is little difference in efficiency between these two estimators. In the first case, $\alpha = \beta = 0$, $\hat{\beta}_0$ is slightly more efficient regardless of the actual first lag correlation. For other values of α and β, $\hat{\beta}_1$ is marginally more efficient than $\hat{\beta}_0$ but the difference is never more than 15%. In summary, the efficiency of $\hat{\beta}_0$ relative to $\hat{\beta}_1$ is satisfactory when the true time dependence is first-order Markov and ρ is less than about 0·4; it is near 1·0 for any value of ρ in the beta-binomial case.

4. Variance estimators

The variances given in Propositions 2·1 and 2·2 may be estimated using the information provided in the replication across subjects. We separately estimate the matrices I and I^*. As I depends on the parameters but not on the Y's, a consistent estimate is obtained by using the estimators $\hat{\beta}_0$ or $\hat{\beta}_1$ and $\hat{\rho}$. Further I^* is estimated by 'borrowing strength' across subjects. We average the cross-products of the score equations over subjects as given by

$$\hat{I}^*_{uv} = \frac{1}{K}\sum_{i=1}^{K}\left(\frac{\partial l_i}{\partial \beta_u}\frac{\partial l_i}{\partial \beta_v}\right).$$

Note that $\hat{I}^*$ is a consistent estimator of I^* regardless of the true structure of the time dependence.

5. Example

As an illustration, we have analysed a binary longitudinal data set collected in a study of stress in mothers with infants. Each binary series indicates whether a mother was stressed, 1, or not, 0, relative to normal on 28 consecutive days. The binary stress variable is based upon self-ratings. One hundred and sixty seven women participated in the study. The covariates of interest are: Z_1 denoting marital status with 0, single, and 1, married; Z_2 denoting infant's general health status with 0, poor to fair, and 1, fair to good; and Z_3 denoting employment status with 0, unemployed, and 1, employed. We fitted a model including the three covariates as well as interaction terms for infants' health with marital status and with employment as given by

$$\operatorname{logit} \pi_i = \beta_0 + \beta_1 Z_1 + \beta_2 Z_2 + \beta_3 Z_3 + \beta_{12} Z_1 Z_2 + \beta_{23} Z_2 Z_3.$$

Table 2 presents the coefficients and t statistics obtained from the independence and first-order Markov working models. For comparison, we also give naive t statistics $t_{N,u} = \hat{\beta}_u/(I_{uu}^{-1})^{\frac{1}{2}}$ which would be used if we believed the working models were correct.

Table 2. *Estimated coefficients and t statistics obtained using the independence and first-order Markov working models; 'naive' t statistics, t_N, working models being assumed correct*

Regressor	Independent model $\hat{\beta}$	t	t_N	First-order Markov model $\hat{\beta}$	t	t_N
Intercept	−2·2	16	26	−2·1	16	25
Marital status, MS	0·4	2·9	4·6	0·4	2·3	3·7
Child's health, CH	0·8	2·1	3·5	0·7	1·9	2·7
Employment status, ES	−0·2	−0·9	−1·8	−0·2	−0·8	−1·4
MS × CH	−1·4	−2·6	−3·0	−1·2	−2·0	−2·2
ES × CH	0·6	0·9	1·7	0·6	0·9	1·4
ρ	0·0	—	—	0·2	6·9	8·8

The estimates and the corresponding t statistics are similar for the two working models. Marital status, child's health and the interaction of these two variables appear to be related to a mother's propensity to experience excess stress. Employment does not appear to be a factor. It is interesting that the propensity for stress is greater in married than in single women with healthy infants. However, if the infant has poor health, marriage is associated with less stress as indicated by the interaction term.

Note that the naive t statistics are larger than those obtained with the consistent variance estimates by a factor of 1 to 2. The consistent estimate has probably been inflated by correlation within a series. The lag one autocorrelation is estimated to be 0·18 with the first-order Markov working model. As this is insufficient to inflate the variance of β by the observed amount if the true model is first-order Markov, a more complex correlation structure is likely to be present.

6. DISCUSSION

In specifying a model for time dependence in Gaussian longitudinal data, the concern is to increase the efficiency of estimation of the regression parameters. If the regression model is correct, we are guaranteed consistency even when the time dependence is misspecified. This is not necessarily true in the binary case. Inconsistent regression parameter estimates can for example occur when each series is incorrectly assumed to a Markov chain of order p but the data are actually generated as a Markov chain of order greater than p. This problem arises because the marginal expectation of a Markov variate depends on the form of the time dependence as well as on the form of the regression model. Hence caution must be exercized to choose working models like the independence and first-order Markov models that lead to consistent estimators under weak assumptions about the time structure, even at the cost of efficiency.

Missing data are often a concern in longitudinal studies. The independence and first-order Markov working models introduced above lead to simple handling of missing binary observations. For the independence working model, the missing data are left out of the likelihood and the resulting estimator remains consistent given only the logit link assumption. In the first-order Markov case, the observation after a missing value is treated as if it were the first observation for a new subject. This is not the working maximum likelihood approach, the EM algorithm. However, this method yields a consistent estimator under the weak assumptions of a logit link and common first lag autocorrelation. The full EM approach will give a consistent estimator only if each binary series is actually first-order Markov.

Often some of the covariates in longitudinal studies also depend on time. Then no difficulty will be imposed if the independent working model is adopted. However, the first-order Markov working model is not immediately applicable.

An alternative approach for consistent estimation of β is to use any estimating equation of the form

$$0 = \sum_{i=1}^{K} V^{-1}(\rho, \beta)(Y_i - \pi_i 1),$$

where V_i is an $n_i \times n_i$ covariance matrix depending on the unknown parameters β and ρ; $Y_i' = (Y_{i1}, \ldots, Y_{in_i})$ and 1 is an $n_i \times 1$ vector of 1's. The form of V_i may be taken to approximate the actual covariance structure of Y_i, however consistency of the resulting estimator of β does not depend on the adequacy of this approximation. The estimating equations above must be augmented by another series of equations for the estimation of ρ. Provided estimates of ρ converge to some fixed quantity, large-sample properties and estimates of variance may be developed as above. Two advantages to this approach for generalizing the independence working likelihood is that extensions to allow for time dependent covariates are straightforward and accommodation of missing data is simple. Problems with this approach include choosing a reasonable form for V_i and a method for the estimation of ρ which results in a reasonably efficient estimator of β.

Acknowledgement

We are grateful to Richard Royall for suggesting the variance estimate presented in §4.

References

Azzalini, A. (1983). Maximum likelihood estimation of order n for stationary stochastic processes. *Biometrika* **70**, 381–8.

Cox, D. R. (1958). The regression analysis of binary sequences (with discussion). *J. R. Statist. Soc.* B **20**, 215–42.

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

Grizzle, J. E., Starmer, C. F. & Koch, G. G. (1969). Analysis of categorical data by linear models. *Biometrics* **25**, 489–504.

Huber, P. (1967). The behavior of maximum likelihood estimates under nonstandard conditions. *Proc. 5th Berkeley Symp.* **1**, 221–34.

Koch, G. G., Landis, J. R., Freeman, J. L., Freeman, D. H. & Lehnen, R. G. (1977). A general methodology for the analysis of experiments with repeated measurement of categorical data. *Biometrics* **33**, 133–58.

Korn, E. L. & Whittemore, A. S. (1979). Methods for analyzing panel studies of acute health effects of air pollution. *Biometrics* **35**, 795–802.

MUENZ, L. R. & RUBINSTEIN, L. V. (1984). Modelling the covariate dependence of binary sequences. *Biometrics*. To appear.
OGATA, Y. (1980). Maximum likelihood estimates of incorrect Markov models for time series and the derivation of AIC. *J. Appl. Prob.* **17**, 59–72.

[*Received April* 1984. *Revised July* 1984]