Statistica Sinica 4(1994), 89-106

PARTIAL LIKELIHOOD ANALYSIS OF LOGISTIC REGRESSION AND AUTOREGRESSION

Eric Slud and Benjamin Kedem

University of Maryland

Abstract: A general *logistic autoregressive* model for binary time series that takes into account stochastic time dependent covariates is presented, and its large sample theory is studied via partial likelihood inference in the sense of Cox (1975) and Wong (1986). The maximum partial likelihood estimator is consistent and asymptotically normal under some conditions on the asymptotic behavior of the time dependent covariates. This leads to asymptotic results concerning several goodness of fit and test statistics. Some of these statistics are applied in logistic regression analysis of level-upcrossings of runoff data using rainfall as covariate data.

Key words and phrases: Maximum partial likelihood estimator, asymptotic efficiency, information matrix, empirical measure, goodness of fit, martingale.

1. Introduction

Binary series can occur in many different ways. For example, in a chemical process, a procedure at time t may be under control ($X_t = 1$), or not ($X_t = 0$). In an industrial process, a drill at time t may become dull ($X_t = 1$), or not ($X_t = 0$). When $X_t = 1$, the drill is replaced, and the process continues. It is helpful, however, to think of $\{X_t\}$ as being generated by the upcrossings of a fixed threshold by an underlying process. Then $X_t = 1$ if the process exceeds a given fixed threshold, and is 0 otherwise. As an example of this, $\{X_t\}$ can be the indicator associated with the excitation voltage of a neuron.

In numerous practical situations, one is interested in the prediction of a future value of a stationary or nonstationary univariate binary time series $\{X_t\}$, $t = 0, \pm 1, \pm 2, \ldots$, from past values of $\{X_t\}$, and past (and sometimes also present) values of an auxiliary (column) vector, $\{\boldsymbol{Z}_t\}$.

Two typical cases of interest arise. In the first, X_t is predicted from the past only given the past data which generate the σ-field

$$\mathcal{F}_{t-1} = \sigma(X_{t-1}, X_{t-2}, \ldots, \boldsymbol{Z}_{t-1}, \boldsymbol{Z}_{t-2}, \ldots).$$

This is genuine prediction. This situation is encountered in predicting whether the water level of a river crosses a critical threshold, given past rainfall and runoff data (Yakowitz (1987)). In the second case, X_t is predicted given the σ-field

$$\mathcal{F}_{t-1} = \sigma(X_{t-1}, X_{t-2}, \ldots, \boldsymbol{Z}_t, \boldsymbol{Z}_{t-1}, \ldots).$$

Here the covariate information at time t is known before observing X_t. This is more like a classification problem. This situation is encountered in classifying instantaneous rain rate as being above or below a fixed threshold given the present and past values of cloud microwave temperature at several frequencies (Chiu and Kedem (1990)). In either case, the vector of covariates $\boldsymbol{Z}_t$ may contain components that are functions of X_t. For example, one covariate component could be $X_{t-1}X_{t-7}$. To a biostatistician the components of $\boldsymbol{Z}_t$ are *time dependent covariates.* In econometrics, the components of $\boldsymbol{Z}_t$ derived neither from past values of X_t nor from some related underlying process would be called *exogenous variables.*

The prediction problem is to estimate from past information the one-step conditional probability $p_t = P(X_t = 1 \,|\, \mathcal{F}_{t-1})$. Our approach is via *partial likelihood parametric inference* concerning a *time invariant* vector parameter $\boldsymbol{\beta}$ which parametrizes $p_t = p_t(\boldsymbol{\beta})$. This enables prediction and/or hypothesis testing concerning the strength of dependence between $\{X_t\}$ and the explanatory vector $\{\boldsymbol{Z}_t\}$.

1.1. On partial likelihood

Partial likelihood (PL) was introduced by Cox (1972, 1975), and given more formal definition and theoretical justification in Wong (1986), and Slud (1992, 1993). For survival and counting-process problems, related developments appeared in Andersen and Gill (1982), Arjas and Haara (1984, 1987). The general definition given below follows Slud (1992).

Let $\mathcal{F}_k$, $k = 0, 1, 2, \ldots$, be an increasing sequence of σ-fields, and let $X_1, X_2, \ldots$ be a sequence of random variables on some common probability space, such that X_k is $\mathcal{F}_k$-measurable. Let $p_k(x_k; \theta)$ be the probability density given $\mathcal{F}_{k-1}$ for X_k under a probability measure P_θ. The *partial likelihood function* relative to $\theta, \{\mathcal{F}_k\}$, and the data $\{X_k\}$ is given by the product

$$\mathrm{PL}(\theta; \boldsymbol{X}_N) \equiv \mathrm{PL}(\theta; X_1, \ldots, X_N) = \prod_{k=1}^{N} p_k(X_k; \theta). \qquad (1)$$

We think of $\mathcal{F}_{k-1}$ as the σ-field generated by past X_t, $t \leq k-1$, and past covariate information, possibly including also present covariates. In this respect PL generalizes the notions of *both* likelihood and conditional-likelihood. More precisely, PL reduces to the usual *likelihood* when the auxiliary covariate information is absent, and it becomes *conditional-likelihood* when all the auxiliary covariate information is known throughout the period of observation; i.e. is measurable with respect to some initial σ-field $\mathcal{F}_0$, where

$$\mathcal{F}_0 \subset \mathcal{F}_1 \subset \mathcal{F}_2 \subset \cdots \subset \mathcal{F}_N.$$

There is a subtle difference between conditional and partial likelihood based inference which needs a clarification. The key idea behind PL is that it permits *sequential conditional inference*: data are processed as they arrive in time, taking into account all that is known to the observer at the time, including past auxiliary information. Hence, time ordering is an essential element. On the other hand, in conditional likelihood based inference the auxiliary information must be known *throughout the period of observation.* Examples of conditional-likelihood based inference, in the context of binary and categorical data, are treated by Keenan (1982), Muenz and Rubinstein (1985), Kaufmann (1987), Fahrmeir and Kaufmann (1987), Fahrmeir (1992); see also Kedem (1980), and Liang and Zeger (1989). The parameter θ has different interpretations in different contexts and the choice among the two approaches depends on the scientific problems and objectives at hand.

To be more specific, if $\mathcal{G}_0 \subset \mathcal{G}_1 \subset \mathcal{G}_2$ are nested σ-fields generated by the covariates, and $\mathcal{F}_t = \sigma(\mathcal{G}_t, \{X_s, s \leq t\})$, a conditional-likelihood function for a parameter θ given $\mathcal{G}_2$ is

$$p_\theta(X_1, X_2, X_3 | \mathcal{G}_2) = p_\theta(X_1 | \mathcal{G}_2) p_\theta(X_2 | X_1, \mathcal{G}_2) p_\theta(X_3 | X_1, X_2, \mathcal{G}_2).$$

A partial likelihood with respect to $\{\mathcal{F}_t\}$, on the other hand, is given by

$$p_\theta(X_1 | \mathcal{F}_0) p_\theta(X_2 | \mathcal{F}_1) p_\theta(X_3 | \mathcal{F}_2)$$

where $\mathcal{F}_0 \subset \mathcal{F}_1 \subset \mathcal{F}_2$. Thus $p_\theta(X_1 \mid \mathcal{G}_2)$ depends on *future* auxiliary information, as opposed to $p_\theta(X_1 \mid \mathcal{F}_0)$ where only *past* auxiliary information and past data enter.

Among the important properties of PL is that the *score process* obtained as the gradient with respect to θ of the partial likelihood is a (vector) *martingale* with respect to the filtration $\mathcal{F}_k$. Hence, rigorous inference about θ is possible using asymptotic results from martingale theory.

2. The Logistic Model

Logistic regression models have been used for years by statisticians, econometricians, and psychometricians. Berkson (1944), Cox (1970), and Nerlove and

Press (1973) are early references. More recently, the following logistic model has drawn attention in connection with time series applications,

$$p_t(\boldsymbol{\beta}) \equiv P_{\boldsymbol{\beta}}(X_t = 1|\mathcal{F}_{t-1}) = \frac{1}{1+\exp[-\boldsymbol{\beta}'\boldsymbol{Z}_{t-1}]} \tag{2}$$

where $\boldsymbol{\beta}$ is a column vector parameter of the same dimension as $\boldsymbol{Z}_{t-1}$. For convenience we assume that the first coordinate of $\boldsymbol{Z}_{t-1}$ is 1 (so that there is an intercept-coefficient), and that $\boldsymbol{Z}_{t-1}$ contains past values of X_t. The model has been considered in various forms by Arjas and Haara (1987), Fahrmeir and Kaufmann (1987), Kaufmann (1987), Slud and Kedem (1988), Liang and Zeger (1989), and Fahrmeir (1992). We can see that when $\boldsymbol{\beta}'\boldsymbol{Z}_t$ includes linear functions of $X_{t-1}, X_{t-2}, \ldots$, the model (2) becomes a form of *binary autoregression.*

Since X_t is binary, (2) implies that

$$p_t(x_t;\boldsymbol{\beta}) = P_{\boldsymbol{\beta}}(X_t = x_t|\mathcal{F}_{t-1}) = [p_t(\boldsymbol{\beta})]^{x_t}[1-p_t(\boldsymbol{\beta})]^{1-x_t}.$$

The corresponding partial likelihood is simply the product

$$\mathrm{PL}(\boldsymbol{\beta}) = \prod_{t=1}^{N} p_t(X_t;\boldsymbol{\beta}) = \prod_{t=1}^{N}[p_t(\boldsymbol{\beta})]^{x_t}[1-p_t(\boldsymbol{\beta})]^{1-x_t}. \tag{3}$$

The maximizer $\hat{\boldsymbol{\beta}}$ of $\mathrm{PL}(\boldsymbol{\beta})$ is called the *Maximum Partial Likelihood Estimator* (MPLE) of $\boldsymbol{\beta}$. Under rather mild regularity conditions on the large-sample behavior of the covariate process $\{\boldsymbol{Z}_t\}$ as $N \to \infty$, $\hat{\boldsymbol{\beta}}$ is a numerically stable estimator of $\boldsymbol{\beta}$, which is consistent and asymptotically normal with easily estimated covariance matrix (Arjas and Haara (1987)). This will be shown in Section 3.

How can time series models of the type (2) arise? To answer this we discuss, next, two important special cases.

2.1. Logistic autoregression

Let $\{Y_t\}$, $t = 0, \pm 1, \pm 2, \ldots$, be an autoregressive process of order p,

$$Y_t = \gamma_0 + \gamma_1 Y_{t-1} + \cdots + \gamma_p Y_{t-p} + \lambda\epsilon_t$$

where λ is a constant, and the ϵ_t are i.i.d. random variables logistically distributed, $\epsilon_t \sim f(x) = e^x/(1+e^x)^2$. Now fix a threshold $r \in (-\infty, \infty)$, and define a binary time series by clipping Y_t at this threshold:

$$X_t \equiv I_{[Y_t \geq r]} = \begin{cases} 1, & \text{if } Y_t \geq r \\ 0, & \text{if } Y_t < r. \end{cases}$$

Then $X_t = I_{[Y_t \geq r]}$ satisfies (2) for each fixed r, with

$$Z_{t-1} = (1, Y_{t-1}, Y_{t-2}, \ldots, Y_{t-p})'$$

and

$$\beta = \frac{1}{\lambda}(\gamma_0 - r, \gamma_1, \ldots, \gamma_p)'.$$

That is,

$$p_t(\beta) \equiv P_\beta(X_t = 1 | \mathcal{F}_{t-1}) = \frac{1}{1 + \exp[-(\gamma_0 - r + \gamma_1 Y_{t-1} + \cdots + \gamma_p Y_{t-p})/\lambda]}. \quad (4)$$

In other words, an $AR(p)$ model with logistic errors implies (2) for all r, with only the intercept component of β depending on r. Conversely, it is easy to see that if (4) holds for $X_t = I_{[Y_t \geq r]}$ for all r, then Y_t is an $AR(p)$ process with logistically distributed noise.

The model (4) for $X_t = I_{[Y_t \geq r]}$ with a fixed known r, is a much weaker restriction than the assumption that Y_t is an $AR(p)$ process with logistic errors, since it leaves the conditional law of Y_t given $\mathcal{F}_{t-1}$ and $\{Y_t \geq r\}$ completely unspecified. In Section 6 we shall present an example where the data seem to obey (4) for $X_t = I_{[Y_t \geq r]}$ with some fixed r, but the data seem not to fit an $AR(p)$.

2.2. Discrete-time Cox models with time-dependent covariates and multiple event-times

The logistic regression model (2) can accommodate multiple independent realizations $\{X_t^j, Z_{t-1}^j\}$, $j = 1, \ldots, m$. In this case we may entertain the model

$$p_t^i(\beta) \equiv P_\beta(X_t^i = 1 | \mathcal{F}_{t-1}) = \frac{1}{1 + \exp[-\beta' Z_{t-1}^i]} \quad (5)$$

where $\mathcal{F}_{t-1}$ is now generated by all the variables X_s^i, Z_s^i, $i = 1, \ldots, m, 0 \leq s \leq t-1$. General models of the type (5) have appeared before in the context of survival analysis (see Cox (1975), Andersen and Gill (1982), Arjas and Haara (1987)). In that setting, X_t^i is the indicator of the event that the ith individual under study fails at time t (or before time t, if time is continuous), and the vector Z_t^i of time dependent covariates contains all the observable data relevant to the failure of the ith individual at time $t+1$ (or, in continuous time to failure in the period $(t, t+1]$).

The model (5) can be applied in studying multiple failures for the same individual, such as multiple times to tumor corresponding to an individual (see Gail et al. (1980)). In this case, Z_{t-1}^i may contain differently modified medical measurements according to the number of values $s = 1, \ldots, t-1$, for which

$X_s^i = 1$. The purpose of this is to allow different coefficients β_j to operate when previous failures have occurred. We could condition in (5) also on $\{X_t^j, j \neq i\}$ in addition to $\mathcal{F}_{t-1}$.

The partial likelihood corresponding to (5) is a generalization of (3), and is given by,

$$\mathrm{PL}(\beta) = \prod_{t=1}^{N}\prod_{i=1}^{m} p_t(X_t^i;\beta) = \prod_{t=1}^{N}\prod_{i=1}^{m}[p_t^i(\beta)]^{x_t^i}[1-p_t^i(\beta)]^{1-x_t^i}. \tag{6}$$

The MPLE, $\hat{\beta}$, is obtained by maximizing (6) with respect to β.

The following key fact about the logistic regression model (5) will be used repeatedly. Conditionally given $\mathcal{F}_{t-1}$, the binary variable X_t^i has mean $p_t^i(\beta)$ and variance $p_t^i(\beta)(1-p_t^i(\beta))$, so that for each i and $s \le t$

$$E_\beta\Big[Z_{s-1}^i Z_{t-1}^{i'}(X_s^i - p_s^i(\beta))(X_t^i - p_t^i(\beta))|\mathcal{F}_{s-1}\Big]$$
$$= \begin{cases} 0, & \text{if } s<t \\ Z_{s-1}^i Z_{s-1}^{i'} p_s^i(\beta)(1-p_s^i(\beta)), & \text{if } s=t. \end{cases}$$

3. Large Sample Theory

Because the extra generality achieved by (5) is useful for certain applications when independent binary processes $\{X_t^i\}$ are available, we shall follow the general case presented in Section 2.2, using (6). Throughout the paper E_β and Var_β indicate that expectations are being taken with respect to probability P_β satisfying (5).

Let the dimension of β be d, $\beta \in \mathcal{R}^d$, and let ∇ denote the column gradient operator:

$$\nabla f(\beta) \equiv \left(\frac{\partial f}{\partial \beta_1}, \frac{\partial f}{\partial \beta_2}, \ldots, \frac{\partial f}{\partial \beta_d}\right)'$$

so that the Hessian matrix-valued operator of second order partial derivatives is given by $\nabla\nabla'$. With this notation,

$$\nabla p_t^i(\beta) = Z_{t-1}^i p_t^i(\beta)(1-p_t^i(\beta))$$

and the *score vector* (d-dimensional) is defined by,

$$S_N(\beta) \equiv \nabla \log \mathrm{PL}(\beta) = \sum_{s=1}^{N}\sum_{i=1}^{m} Z_{s-1}^i(X_s^i - p_s^i(\beta)).$$

The *score vector process* $\boldsymbol{S}_t(\boldsymbol{\beta})$, $t = 1, 2, \ldots, N$, is defined by the partial sums,

$$\boldsymbol{S}_t(\boldsymbol{\beta}) = \sum_{s=1}^{t} \sum_{i=1}^{m} \boldsymbol{Z}_{s-1}^i (X_s^i - p_s^i(\boldsymbol{\beta})).$$

The score process, being the sum of martingale differences, is easily seen to be a martingale with respect to the filtration $\mathcal{F}_t$, relative to $P_{\boldsymbol{\beta}}$. That is, $E[\boldsymbol{S}_t(\boldsymbol{\beta}) \mid \mathcal{F}_{t-1}] = \boldsymbol{S}_{t-1}(\boldsymbol{\beta})$. Clearly, $E[\boldsymbol{S}_t(\boldsymbol{\beta})] = \mathbf{0}$. Next we define,

$$\boldsymbol{I}(\boldsymbol{\beta}) \equiv \nabla\nabla'(-\log \mathrm{PL}(\boldsymbol{\beta})) = \sum_{t=1}^{N} \sum_{i=1}^{m} \boldsymbol{Z}_{t-1}^i \boldsymbol{Z}_{t-1}^{i\prime} p_t^i(\boldsymbol{\beta})(1 - p_t^i(\boldsymbol{\beta})).$$

The quantity $\boldsymbol{I}(\boldsymbol{\beta})/mN$ is the *sample information matrix per observation* for estimating $\boldsymbol{\beta}$. It is easily seen that $\boldsymbol{I}(\boldsymbol{\beta})$ is the sum of conditional covariance matrices,

$$\boldsymbol{I}(\boldsymbol{\beta}) = \sum_{s=1}^{N} \mathrm{Var}_{\boldsymbol{\beta}} \left[\sum_{i=1}^{m} \boldsymbol{Z}_{s-1}^i (X_s^i - p_s^i(\boldsymbol{\beta})) \mid \mathcal{F}_{s-1} \right].$$

Since $\boldsymbol{S}_t(\boldsymbol{\beta})$ is a martingale, we refer to $\boldsymbol{I}(\boldsymbol{\beta})$ as the *cumulative conditional variance-covariance matrix* for $\boldsymbol{S}_N(\boldsymbol{\beta})$.

The large sample properties of the MPLE $\hat{\boldsymbol{\beta}}$ are studied with the aid of $\boldsymbol{S}_t(\boldsymbol{\beta})$, and $\boldsymbol{I}(\boldsymbol{\beta})$. This theory was developed by Andersen and Gill (1982), Wong (1986), and Arjas and Haara (1987). We follow the general development given in Slud (1993, Chs. 6, 7). The approach taken in these references for proving consistency and asymptotic normality of MPLE's is based on the martingale Central Limit Theorem for $\boldsymbol{S}_N(\boldsymbol{\beta})/\sqrt{mN}$, the almost sure concavity of the random function PL on $\mathcal{R}^d$, and the stability of the sample information matrix $\boldsymbol{I}(\boldsymbol{\beta})/mN$. There is no need to duplicate formal proofs, and we shall be content with general comments, a precise set of regularity conditions, and statements of some small extensions, helpful in time series applications.

Our regularity conditions, chosen for simplicity rather than utmost generality, are:

(A.1) The covariate-vectors $\boldsymbol{Z}_t^i$ almost surely lie in a nonrandom compact subset Γ of $\mathcal{R}^d$.

(A.2) The probability measure P governing $\{X_t^i, \boldsymbol{Z}_t^i\}$, $i = 1, \ldots, m$, $t = 0, \ldots, N$, obeys (5) with $\boldsymbol{\beta} = \boldsymbol{\beta}_0$.

(A.3) There is a probability measure ν on $\mathcal{R}^d$ for which $\int_{\mathcal{R}^d} \boldsymbol{z}\boldsymbol{z}'\nu(d\boldsymbol{z})$ is positive definite, such that under (5) with $\boldsymbol{\beta} = \boldsymbol{\beta}_0$, for Borel sets $A \subset \mathcal{R}^d$

$$\frac{1}{mN} \sum_{t=1}^{N} \sum_{i=1}^{m} I_{[\boldsymbol{Z}_{t-1}^i \in A]} \xrightarrow{P} \nu(A), \quad mN \to \infty.$$

Assumption (A.1) is somewhat restrictive, but it can be replaced by a more general assumption uniformly bounding the moment generating function of $\boldsymbol{Z}_t^i$ as remarked in Andersen and Gill (1982). The logistic autoregression model (4) satisfies the weakened assumption but not (A.1).

The motivation for Assumption (A.3) comes from the Birkhoff Ergodic Theorem in the setting of independent ergodic stationary processes $\{\boldsymbol{Z}_t^i, t \geq 0\}$, $i = 1, \ldots, m$. Assumption (A.3) says that the empirical measure of the set $\{\boldsymbol{Z}_s^i : 0 \leq s < N, 1 \leq i \leq m\}$ converges weakly almost surely to a nonrandom measure ν. This implies that for every continuous function $g : \mathcal{R}^d \to \mathcal{R}$, (necessarily bounded on the compact support Γ of $\boldsymbol{Z}_t^i$)

$$\frac{1}{mN}\sum_{t=1}^{N}\sum_{i=1}^{m} g(\boldsymbol{Z}_{t-1}^i) \xrightarrow{P} \int_{\mathcal{R}^d} g(\boldsymbol{z})\nu(d\boldsymbol{z})$$

as $mN \to \infty$. From this convergence, it is easy to see that $\boldsymbol{I}(\boldsymbol{\beta})/mN$ has a limit, say $\Lambda(\boldsymbol{\beta})$, given by

$$\Lambda(\boldsymbol{\beta}) = \int_{\mathcal{R}^d} \frac{e^{\boldsymbol{\beta}'\boldsymbol{z}}}{\left(1 + e^{\boldsymbol{\beta}'\boldsymbol{z}}\right)^2} \boldsymbol{z}\boldsymbol{z}'\nu(d\boldsymbol{z}). \tag{7}$$

The matrix $\Lambda(\boldsymbol{\beta})$ at $\boldsymbol{\beta} = \boldsymbol{\beta}_0$ is called the *information matrix per observation* for estimating $\boldsymbol{\beta}$. By (A.3), $\Lambda(\boldsymbol{\beta})$ is positive definite for every $\boldsymbol{\beta}$, and hence also nonsingular. Now, $\nabla \log \mathrm{PL}(\hat{\boldsymbol{\beta}}) = \boldsymbol{0}$. Thus by Taylor series expansion of $\nabla \log \mathrm{PL}(\hat{\boldsymbol{\beta}})$ to one term about $\boldsymbol{\beta}_0$, we obtain the useful approximation up to terms asymptotically negligible in probability,

$$\begin{aligned}\sqrt{mN}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) &\approx \left(-\frac{1}{mN}\nabla\nabla' \log \mathrm{PL}(\boldsymbol{\beta}_0)\right)^{-1} \frac{1}{\sqrt{mN}} \nabla \log \mathrm{PL}(\boldsymbol{\beta}_0) \\ &= \left(\frac{1}{mN}\boldsymbol{I}(\boldsymbol{\beta}_0)\right)^{-1} \frac{1}{\sqrt{mN}} \boldsymbol{S}_N(\boldsymbol{\beta}_0) \\ &\approx \Lambda(\boldsymbol{\beta}_0)^{-1} \frac{1}{\sqrt{mN}} \boldsymbol{S}_N(\boldsymbol{\beta}_0).\end{aligned} \tag{8}$$

Note that $\boldsymbol{I}(\boldsymbol{\beta}_0)$ behaves asymptotically as $mN\Lambda(\boldsymbol{\beta}_0)$, so that $(\boldsymbol{I}(\boldsymbol{\beta}_0))^{-1}\boldsymbol{S}_N(\boldsymbol{\beta}_0)$ converges in probability to 0. Thus, $\boldsymbol{S}_N(\boldsymbol{\beta}_0)/\sqrt{mN}$ is a martingale, with asymptotic covariance matrix $\Lambda(\boldsymbol{\beta}_0)$. Hence, by appealing to Slutsky's theorem and the CLT for martingales, we have

Theorem 3.1. *Under assumptions* (A.1)-(A.3), *the MPLE* $\hat{\boldsymbol{\beta}}$ *is almost surely unique for all sufficiently large* mN, *and as* $mN \to \infty$,

(i) $\hat{\boldsymbol{\beta}} \xrightarrow{P} \boldsymbol{\beta}_0$.

(ii) $$\sqrt{mN}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \xrightarrow{D} \mathcal{N}(\mathbf{0}, \boldsymbol{\Lambda}^{-1}(\boldsymbol{\beta}_0)).$$

(iii) $$\sqrt{mN}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) - \frac{1}{\sqrt{mN}}\boldsymbol{\Lambda}^{-1}(\boldsymbol{\beta}_0)\boldsymbol{S}_N(\boldsymbol{\beta}_0) \xrightarrow{P} 0.$$

The large sample behavior of $\hat{\boldsymbol{\beta}}$ described here is based on the assumed stability of $\boldsymbol{I}(\boldsymbol{\beta})/mN$, the *information per observation*. It is convenient that a single theoretical result encompasses all three cases where m and/or N become large. However, in general $\boldsymbol{I}(\boldsymbol{\beta})/mN$ could converge in probability to a random limit as in the *non ergodic* cases treated in Basawa and Scott (1983).

We turn next to the question of goodness of fit. In any attempt to fit (5), it is important to examine the logistic regression *residuals* $X_t^i - p_t^i(\hat{\boldsymbol{\beta}})$ in order to judge the quality of fit. This, however, can be done in many different ways.

One useful way to test goodness of fit is to *classify* the responses X_t^i according to mutually exclusive events defined in terms of the covariates $\boldsymbol{Z}_{t-1}^i$, and then check for each category the deviation of the number of positive responses from its conditional expected value (Schoenfeld (1980)). More precisely, let $C_1, \ldots, C_k$, constitute a partition of $\mathcal{R}^d$. For $j = 1, \ldots, k$, define,

$$M_j \equiv \sum_{i=1}^{m}\sum_{t=1}^{N} I_{[\boldsymbol{Z}_{t-1}^i \in C_j]} X_t^i$$

and

$$e_j(\boldsymbol{\beta}) \equiv \sum_{i=1}^{m}\sum_{t=1}^{N} I_{[\boldsymbol{Z}_{t-1}^i \in C_j]} p_t^i(\boldsymbol{\beta}).$$

Put $\boldsymbol{M} \equiv (M_1, \ldots, M_k)'$, $\boldsymbol{e}(\boldsymbol{\beta}) \equiv (e_1(\boldsymbol{\beta}), \ldots, e_k(\boldsymbol{\beta}))'$. The goodness of fit can be tested with the help of quadratic forms of the type

$$(\boldsymbol{M} - \boldsymbol{e}(\hat{\boldsymbol{\beta}}))'\boldsymbol{V}(\boldsymbol{M} - \boldsymbol{e}(\hat{\boldsymbol{\beta}}))$$

where $\boldsymbol{V}$ is a suitable $k \times k$ matrix.

For testing the *hypothesis* that $\boldsymbol{\beta} = \boldsymbol{\beta}_0$, we can use statistics of the form

$$\sum_{j=1}^{k} \left(M_j - e_j(\boldsymbol{\beta}_0)\right)^2 W_j$$

or

$$\sum_{i=1}^{m}\sum_{t=1}^{N} \left(X_t^i - p_t^i(\boldsymbol{\beta}_0)\right)^2 W_{it}$$

using appropriate normalizations and weights W.

The theory underlying the preceding goodness of fit and hypothesis testing statistics is contained in the next two theorems. The theorems follow readily from

Theorem 3.1 and several applications of the multivariate Martingale Central Limit Theorem as given in Andersen and Gill (1982, Appendix II). In both theorems, the crucial fact that $E_{\beta_0} p_t^i(\beta)$ lies in a nonrandom compact subset of $(0,1)$, for all i and t, and for all β in a neighborhood of β_0, is guaranteed by Assumption (A.1) or the weakened assumption mentioned following (A.1)-(A.3).

Theorem 3.2. *Let $C_1, \dots, C_k$, be a partition of $\mathcal{R}^d$. Then we have as $mN \to \infty$*
(i)

$$\sqrt{mN}\Big((\boldsymbol{M} - \boldsymbol{e}(\boldsymbol{\beta}_0))'/mN, (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)'\Big)' \xrightarrow{\mathcal{D}} \boldsymbol{N}(\boldsymbol{0}, \boldsymbol{\Sigma})$$

where $\boldsymbol{\Sigma}$ is a square matrix of dimension $d+k$,

$$\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{A} & \boldsymbol{B}' \\ \boldsymbol{B} & \boldsymbol{\Lambda}^{-1}(\boldsymbol{\beta}_0) \end{pmatrix}.$$

Here $\boldsymbol{A}$ is a diagonal $k \times k$ matrix with the jth *diagonal element given by*

$$\sigma_j^2 \equiv \int_{C_j} \frac{e^{\boldsymbol{\beta}_0' \boldsymbol{z}}}{\left(1 + e^{\boldsymbol{\beta}_0' \boldsymbol{z}}\right)^2} \nu(d\boldsymbol{z}),$$

$\boldsymbol{\Lambda}^{-1}(\boldsymbol{\beta}_0)$ is the limiting $d \times d$ inverse of the information matrix, and the jth *column of $\boldsymbol{B}$ is given by*

$$\boldsymbol{\Lambda}^{-1}(\boldsymbol{\beta}_0) \int_{C_j} \frac{e^{\boldsymbol{\beta}_0' \boldsymbol{z}}}{\left(1 + e^{\boldsymbol{\beta}_0' \boldsymbol{z}}\right)^2} \boldsymbol{z} \nu(d\boldsymbol{z}).$$

(ii)

$$\frac{(\boldsymbol{e}(\hat{\boldsymbol{\beta}}) - \boldsymbol{e}(\boldsymbol{\beta}_0))}{\sqrt{mN}} - \sqrt{mN} \boldsymbol{B}' \boldsymbol{\Lambda}(\boldsymbol{\beta}_0)(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \xrightarrow{P} 0.$$

(iii) *As $mN \to \infty$, the asymptotic distribution of the statistic*

$$\chi^2(\boldsymbol{\beta}_0) \equiv \frac{1}{mN} \sum_{j=1}^{k} (M_j - e_j(\boldsymbol{\beta}_0))^2 / \sigma_j^2$$

is χ_k^2.

In verifying Theorem 3.2 it is helpful to note that by Taylor series expansion of $\boldsymbol{e}(\boldsymbol{\beta})$ to one term about $\boldsymbol{\beta}_0$, one obtains for sufficiently large mN (replacing $\boldsymbol{\beta}$ by $\hat{\boldsymbol{\beta}}$),

$$\frac{1}{\sqrt{mN}}(\boldsymbol{e}(\hat{\boldsymbol{\beta}}) - \boldsymbol{e}(\boldsymbol{\beta}_0)) \approx \sqrt{mN} \boldsymbol{B}' \boldsymbol{\Lambda}(\boldsymbol{\beta}_0)(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)$$

and therefore, by adding and subtracting $e(\beta_0)$,

$$\begin{aligned}\frac{1}{\sqrt{mN}}(M - e(\hat{\beta})) &= \frac{1}{\sqrt{mN}}\left(M - e(\beta_0) + e(\beta_0) - e(\hat{\beta})\right) \\ &\approx \frac{1}{\sqrt{mN}}(M - e(\beta_0)) - \sqrt{mN}B'\Lambda(\beta_0)(\hat{\beta} - \beta_0). \quad (9)\end{aligned}$$

It follows that the asymptotic covariance matrix of $(M - e(\hat{\beta}))/\sqrt{mN}$ is given by $A - B'\Lambda(\beta_0)B - B'\Lambda(\beta_0)B + B'\Lambda(\beta_0)B = A - B'\Lambda(\beta_0)B$.

Remark 3.1. Another useful test statistic is the quadratic form

$$\frac{1}{mN}(M - e(\hat{\beta}))'(A - B'\Lambda(\beta_0)B)^{-1}(M - e(\hat{\beta}))$$

where the inverse is a symmetric generalized inverse. The asymptotic distribution of this quadratic form is χ^2_n with $n = \text{rank}(A - B'\Lambda(\beta_0)B) \leq k - 1$.

For the next result, define $v(i,t) \equiv p^i_t(\beta_0)(1 - p^i_t(\beta_0))$ for all i, t. Then with any $a \geq 0$,

$$Y_T \equiv \sum_{i=1}^{m}\sum_{t=1}^{T}\left\{\frac{(X^i_t - p^i_t(\beta_0))^2}{(v(i,t))^a} - (v(i,t))^{1-a}\right\}$$

is a zero mean martingale with cumulative conditional variance

$$\sum_{i=1}^{m}\sum_{t=1}^{T}(v(i,t))^{1-2a}(1 - 4v(i,t)).$$

We therefore have

Theorem 3.3. *Under assumptions* (A.1)-(A.3),

$$W_a(\beta_0) \equiv \frac{\sum_{i=1}^{m}\sum_{t=1}^{N}\left\{\frac{(X^i_t - p^i_t(\beta_0))^2}{(v(i,t))^a} - (v(i,t))^{1-a}\right\}}{\left\{\sum_{i=1}^{m}\sum_{t=1}^{N}(v(i,t))^{1-2a}(1 - 4v(i,t))\right\}^{\frac{1}{2}}} \xrightarrow{D} \mathcal{N}(0,1). \quad (10)$$

The result of Theorem 3.3 is equally valid when β_0 in $p^i_t(\beta_0)$ is replaced by $\hat{\beta}$. However, the limit in (10) may not be very reliable when values $p^i_t(\hat{\beta})$ can be close to 0 or 1, unless $a \in [0, 1/2]$.

We end this section with well known results concerning the partial likelihood ratio statistic. First notice that by expanding $\log \text{PL}(\beta_0)$ to two terms about $\hat{\beta}$,

$$\log \text{PL}(\beta_0) \approx \log \text{PL}(\hat{\beta}) + S_N(\hat{\beta})(\beta_0 - \hat{\beta}) - \frac{1}{2}(\hat{\beta} - \beta_0)'I(\beta_0)(\hat{\beta} - \beta_0)$$

and that $S_N(\hat{\beta}) = \nabla \log \mathrm{PL}(\hat{\beta}) = \mathbf{0}$. It follows from (ii) of Theorem 3.1 that as $mN \to \infty$,

$$2\Big[\log \mathrm{PL}(\hat{\beta}) - \log \mathrm{PL}(\beta_0)] \approx mN(\hat{\beta} - \beta_0)'\Lambda(\beta_0)(\hat{\beta} - \beta_0)$$

is asymptotically χ^2_d. One can use this last statistic for testing the hypothesis $H_0 : \beta = \beta_0$. With a little more effort we can also test the hypothesis H_0 that c $(c \le d)$ of the components of β are equal to specific values, for example, that c of the components are equal to 0. To do that, maximize $\mathrm{PL}(\beta)$ with respect to the remaining unspecified $d - c$ parameters, and denote the maximum by $\mathrm{PL}(\beta^*)$. Clearly $\mathrm{PL}(\hat{\beta}) \ge \mathrm{PL}(\beta^*)$. Then under (A.1)-(A.3) and H_0 (see Slud (1993, Th. 6.5)), as $mN \to \infty$,

$$2[\log \mathrm{PL}(\hat{\beta}) - \log \mathrm{PL}(\beta^*)\Big] \xrightarrow{D} \chi^2_c. \tag{11}$$

4. Asymptotic Relative Efficiency

How efficient are maximum partial likelihood estimates compared with the usual maximum likelihood (ML) estimates? While no general answer to this question is possible, we can give an answer in the case of an $AR(p)$ process as discussed in Section 2.1. In this special case, a fully specified model for $\{X_t, \boldsymbol{Z}_t\}$ is readily available, and a comparison is possible via the information matrices corresponding to partial and full likelihoods.

Referring to Section 2.1, we make the following simplifying assumptions. First, set $r = 0$, $\lambda = 1$, and consider the stationary $AR(p)$ process, $Y_t = \beta_1 + \beta_2 Y_{t-1} + \cdots + \beta_{p+1}Y_{t-p} + \epsilon_t$. Here, $\beta = (\beta_1, \beta_2, \ldots, \beta_{p+1})'$, $X^1_t = X_t = I_{[Y_t \ge 0]}$, and $\boldsymbol{Z}^1_{t-1} = \boldsymbol{Z}_{t-1} = (1, Y_{t-1}, Y_{t-2}, \ldots, Y_{t-p})'$, and

$$\epsilon_t = Y_t - \beta' \boldsymbol{Z}_{t-1} \tag{12}$$

are i.i.d. logistic random variables with density, $f(x) = e^x/(1+e^x)^2$. The variable ϵ_t is independent of $\mathcal{F}_{t-1} = \sigma(\boldsymbol{Z}_s, s < t)$. Under the assumption of stationarity, let $\boldsymbol{Z}$ be distributed as $\boldsymbol{Z}_{t-1}$. Also let ϵ be distributed as ϵ_t, independently of $\boldsymbol{Z}$.

The PL information matrix (7) corresponding to (5) and (6), is given by

$$\Lambda^{\mathrm{PL}}(\beta_0) \equiv \Lambda(\beta_0) = E_{\beta_0}\left[\frac{e^{\beta_0' \boldsymbol{Z}}}{(1 + e^{\beta_0' \boldsymbol{Z}})^2} \boldsymbol{Z}\boldsymbol{Z}'\right] = E_{\beta_0}[f(\beta_0' \boldsymbol{Z})\boldsymbol{Z}\boldsymbol{Z}']. \tag{13}$$

The likelihood $L(\beta)$ based on $Y_{p+1}, \ldots, Y_N$, is given by

$$L(\beta) = \prod_{t=p+1}^{N} f(\epsilon_t) \tag{14}$$

where ϵ_t is given in (12). The information matrix is obtained in exactly the same manner as in the PL case, as the limit of the information about $\boldsymbol{\beta}$ per observation. This information matrix, equal to the inverse of the asymptotic covariance matrix for the ML estimator of $\boldsymbol{\beta}$ when the true parameter value is $\boldsymbol{\beta}_0$, is given by

$$\Lambda^{\mathrm{L}}(\boldsymbol{\beta}_0) \equiv 2E_{\boldsymbol{\beta}_0}\left[\frac{e^{\epsilon}}{(1+e^{\epsilon})^2}\boldsymbol{Z}\boldsymbol{Z}'\right] = 2E_{\boldsymbol{\beta}_0}[f(\epsilon)\boldsymbol{Z}\boldsymbol{Z}'] = \frac{1}{3}E_{\boldsymbol{\beta}_0}[\boldsymbol{Z}\boldsymbol{Z}'] \qquad (15)$$

upon noting that $\int_{-\infty}^{\infty} f^2(x)dx = 1/6$.

Since $f(x) \le 1/4$, it follows immediately from (13), (15), that for every vector $\boldsymbol{b} \in \mathcal{R}^{p+1}$,

$$\boldsymbol{b}'\Lambda^{\mathrm{PL}}(\boldsymbol{\beta}_0)\boldsymbol{b} \le \frac{3}{4}\boldsymbol{b}'\Lambda^{\mathrm{L}}(\boldsymbol{\beta}_0)\boldsymbol{b}. \qquad (16)$$

Thus, any scalar parameter derived from $\boldsymbol{\beta}$ linearly can be estimated with *asymptotic relative efficiency* (ARE) at best 3/4 via the PL logistic regression method as compared with ML analysis for *AR*(*p*). The worst ARE is obtained when $p_t(\boldsymbol{\beta}_0)$ is often close to 1 or 0, i.e., the case where prediction is very good!

5. Other Link Functions

In the foregoing analysis, we have modeled $p_t^i(\boldsymbol{\beta})$ in (5) in terms of $F(\boldsymbol{\beta}'\boldsymbol{Z}_{t-1}^i)$, where $F(x) = 1/(1+\exp(-x))$ is a standard logistic distribution function. However, other "link" functions F can be used. One attractive link is defined by $F \equiv \Phi$, where Φ is the standard normal distribution function. In this case we obtain what is known as *probit* model,

$$\tilde{p}_t^i(\boldsymbol{\beta}) \equiv P_{\boldsymbol{\beta}}(X_t^i = 1 \mid \mathcal{F}_{t-1}) = \Phi(\boldsymbol{\beta}'\boldsymbol{Z}_{t-1}^i). \qquad (17)$$

Virtually every aspect of our analysis under (5) has an analog under (17). Thus, under (17), in the *AR*(*p*) example for $X_t = I_{[Y_t \ge 0]}$ with variance 1, the errors are now Gaussian instead of logistic, and the ARE calculations give

$$\Lambda^{\mathrm{L}}(\boldsymbol{\beta}_0) = E_{\boldsymbol{\beta}_0}[\boldsymbol{Z}\boldsymbol{Z}']$$

and

$$\Lambda^{\mathrm{PL}}(\boldsymbol{\beta}_0) = E_{\boldsymbol{\beta}_0}\left[\boldsymbol{Z}\boldsymbol{Z}'\frac{\phi^2(\boldsymbol{\beta}_0'\boldsymbol{Z})}{\Phi(\boldsymbol{\beta}_0'\boldsymbol{Z})(1-\Phi(\boldsymbol{\beta}_0'\boldsymbol{Z}))}\right]$$

where $\phi(x)$ is the standard normal density. It can be checked that the upper bound for the asymptotic relative efficiency is $2/\pi$, and again the ARE is much worse when $p_t(\boldsymbol{\beta}_0)$ values often fall near 1 and 0.

The proofs of theorems analogous to Theorems (3.1)-(3.2) are very similar to those in the logistic case, with extra complication due to the fact that the log partial likelihood will no longer be a.s. concave for finite time series.

6. Application to Rainfall-Runoff Data

We apply the logistic regression model (2) in the analysis of daily rainfall-runoff data obtained by the National Weather Service in the Bird Creek Ohio watershed, and described in Yakowitz (1987). The data were collected in intervals of 13-15 weeks during each of the years 1939-1964. In what follows, we regard daily runoff Y_t as the response variable, and rainfall R_t as the explanatory variable. Since flooding is of interest, it is natural to try to understand the relationship between level exceedances $X_t = I_{[Y_t > r]}$ and the explanatory variables. Our primary goal is to illustrate how the model (2) can be estimated and pass goodness of fit tests in a particular situation when the linear autoregressive model turns out not to be adequate.

The 26 years of data were split into a testing set (the 10 years 1939-48, consisting of 1031 rainfall-runoff pairs), and a training set (the 16 years 1949-64, consisting of 1691 rainfall-runoff pairs). Since the models we contemplate involve explanatory variables defined from $(Y_{t-1}, \ldots, Y_{t-4}, R_t, \ldots, R_{t-3})$, we lose 4 observations per year which lead to 991 rainfall-runoff pair observations in the testing set, and 1627 in the training set. The threshold values chosen are $r = 1, 3$ *cubic ft/sec*. During the training period 1949-64, there were 401 and 87 positive responses (i.e., level-upcrossings) corresponding to levels $r = 1, 3$, which respectively were chosen as moderate and high levels. The corresponding respective numbers for the testing period 1939-48 were 244, 56.

Residuals plots and partial likelihood ratio statistics (11) obtained from a preliminary fitting of the model (2) for the 1939-48 data, suggest the importance of the covariates

$$Z_t' = \Big(1, R_t, Y_{t-1}, R_t Y_{t-1}, R_{t-1}, R_t R_{t-1}, R_{t-2}, R_{t-1} R_{t-2}, Y_{t-2}, Y_{t-1} Y_{t-2}\Big).$$

Other covariates that we examined did not appear to play a significant role. In particular, our attempts to discover covariates to account for year to year differences in runoff, did not produce new covariates worth including in the logistic model (2).

Our first results are given in Table 1. The table gives the estimated values, from the 1949-64 training data set, of the components $\hat{\beta}_i$ of the MPLE $\hat{\boldsymbol{\beta}}$, and the standardized values $\hat{\beta}_i / \sqrt{\mathrm{Var}(\hat{\beta}_i)}$, corresponding to $X_t = I_{[Y_t > r]}$, $r = 1, 3$. In each case, the 10-dimensional covariate vector is $\boldsymbol{Z}_t$ defined above. Attempts to add the residuals of linear regression as additional covariates resulted in only a slight change in the maximum log-partial likelihood and were therefore neglected.

To begin to assess model adequacy for the logistic model (2) fitted to the 1949-64 rainfall-runoff data, we calculated the goodness of fit statistics $\chi^2(\hat{\boldsymbol{\beta}})$, $W_a(\hat{\boldsymbol{\beta}})$ described in Theorems 3.2, 3.3, replacing $\boldsymbol{\beta}_0$ with the MPLE $\hat{\boldsymbol{\beta}}$ obtained from the 1949-64 data. Thus, the goodness of fit statistics $\chi^2(\hat{\beta})$ and $W_a(\hat{\beta})$ were evaluated from both the training data set (1949-64) and from the testing data set (1939-48), using the *same* $\hat{\beta}$ obtained from the 1949-64 training data. Also σ_j^2 was estimated in each case from the data to which the χ^2 test was being applied, using the estimator

$$\hat{\sigma}_j^2 \equiv \frac{1}{N}\sum_{t=1}^{N} I_{[Z_{t-1}\in C_j]} p_t(\hat{\beta})(1-p_t(\hat{\beta})).$$

For the statistic $\chi^2(\hat{\beta})$, the partition cells $C_j \subset \mathcal{R}^{10}$ are defined as the intersections of all sets satisfying the conditions that R_t is in one of the three intervals $[0, 0.004]$, $(0.004, 0.008]$, or $(0.008, \infty)$, $R_{t-1}+R_t$ is in one of the intervals $[0, 0.01]$, $(0.01, 0.02]$, or $(0.02, \infty)$, and Y_{t-1} is in $[0, 0.5]$, $(0.5, 1]$, or $(1, \infty)$. This gives $k = 3^3 = 27$ partition cells. However, because β_0 is replaced by $\hat{\beta}$, $\chi^2(\hat{\beta})$ is (asymptotically) stochastically smaller than χ^2_{27} on the training data (due to having estimated β_0 from the same data), and stochastically larger than χ^2_{27} on the testing data (since $\hat{\beta}$ is approximately independent of $(\boldsymbol{M} - e(\beta_0))$ for the testing data). An indication of this is provided by invoking (9). Thus, for the training data,

$$\begin{aligned} E[\chi^2(\hat{\beta})] &\approx k - 2\mathrm{tr}\boldsymbol{B}'\Lambda(\beta_0)\boldsymbol{B}\boldsymbol{A}^{-1} + \mathrm{tr}\boldsymbol{B}'\Lambda(\beta_0)\boldsymbol{B}\boldsymbol{A}^{-1} \\ &= k - \sum_{j=1}^{k}(\boldsymbol{B}'\Lambda(\beta_0)\boldsymbol{B})_{jj}/\sigma_j^2 \end{aligned}$$

while for the testing data the middle term $-2\mathrm{tr}\boldsymbol{B}'\Lambda(\beta_0)\boldsymbol{B}\boldsymbol{A}^{-1}$ is absent due to approximate independence of the training and testing data, and hence

$$\begin{aligned} E[\chi^2(\hat{\beta})] &\approx k + \mathrm{tr}\boldsymbol{B}'\Lambda(\beta_0)\boldsymbol{B}\boldsymbol{A}^{-1} \\ &= k + \sum_{j=1}^{k}(\boldsymbol{B}'\Lambda(\beta_0)\boldsymbol{B})_{jj}/\sigma_j^2. \end{aligned}$$

The estimated value of $\sum_{j=1}^{k}(\boldsymbol{B}'\Lambda(\beta_0)\boldsymbol{B})_{jj}/\sigma_j^2$ was 5.3 for $r = 1$, and 3.6 for $r = 3$. The difference is attributed to the very different response probabilities $p_t(\hat{\beta})$ corresponding to different levels r.

The statistics $W_a(\hat{\beta})$, $a = 0, 1$, are asymptotically normal; however, some care is needed in interpreting the results due to division by $[p_t(\hat{\beta})(1-p_t(\hat{\beta}))]^a$. This is so because many of the logistic probabilities $p_t(\hat{\beta})$ are either quite close to 0 or 1, a tendency already seen in the previous example. For this reason $W_0(\hat{\beta})$ is more reliable than $W_1(\hat{\beta})$.

The values of $\chi^2(\hat{\boldsymbol{\beta}})$, $W_0(\hat{\boldsymbol{\beta}})$, $W_1(\hat{\boldsymbol{\beta}})$ are given in Table 2, leading to some interesting conclusions. The table indicates that the logistic model (2) with $X_t = I_{[Y_t > r]}$ for $r = 1, 3$ are adequate for both the training and testing data because the three statistics admit relatively small values, except for $W_1(\hat{\boldsymbol{\beta}})$ for the test data with $r = 3$ (the choice of r is discussed in Chiu and Kedem (1990), Kedem and Pavlopoulos (1991)). However, as remarked above, the fact that the predicted response probabilities $p_t(\hat{\boldsymbol{\beta}})$ are so often very close to 0 or 1 makes the statistic $W_1(\hat{\boldsymbol{\beta}})$ less reliable and more difficult to interpret than the sum of the *unnormalized* squared residuals $W_0(\hat{\boldsymbol{\beta}})$. At any rate, since comparatively few of the response values $X_t = I_{[Y_t > 3]}$ are 1, the large sample theory on which we rely for the validity of our analysis may be less reliable than in the case of $r = 1$. This is the chronic problem associated with predictive models for exceedances of high levels – that there is relatively little data where high levels are exceeded.

In conclusion we would like to note that, as in Yakowitz (1987) and following our reasoning in Section 2.1, we have gone through numerous linear regression as well as linear autoregression fits, assuming logistic errors, using different thresholds r and various covariates. The result was that the prediction of $X_t = I_{[Y_t > r]}$ was much better with our logistic model than with linear autoregression. Thus, a further interesting conclusion from fitting the logistic model (2) to the rainfall-runoff data, is that while the logistic model appears to fit reasonably well with both levels $r = 1, 3$, it does not appear to be compatible with a single linear model for rainfall-runoff. That is, the coefficients $\hat{\boldsymbol{\beta}}$ are markedly different for the different levels r, violating (4).

Table 1. Logistic regression parameter estimates $\hat{\beta}_i$ obtained from the 1949-64 rainfall-runoff data.

		$r = 1$		$r = 3$	
i	Covariate	$\hat{\beta}_i$	$\frac{\hat{\beta}_i}{\sqrt{\mathrm{Var}(\hat{\beta}_i)}}$	$\hat{\beta}_i$	$\frac{\hat{\beta}_i}{\sqrt{\mathrm{Var}(\hat{\beta}_i)}}$
1	Intercept	-6.31	-13.6	-6.25	-13.9
2	R_t	161.20	6.9	99.70	6.6
3	Y_{t-1}	4.38	9.6	1.05	5.8
4	$R_t Y_{t-1}$	58.50	1.4	64.40	3.9
5	R_{t-1}	70.90	3.0	-50.50	-1.9
6	$R_t R_{t-1}$	-2509.00	-1.5	3392.00	2.3
7	R_{t-2}	-63.00	-4.3	-23.50	-1.2
8	$R_{t-1} R_{t-2}$	-2124.00	-1.2	-2498.00	-2.2
9	Y_{t-2}	0.24	0.6	0.29	2.4
10	$Y_{t-1} Y_{t-2}$	0.31	1.4	-0.04	-2.7

Table 2. Goodness of fit statistics on training and testing data for the logistic model (2) with coefficients displayed in Table 1.

	Training data (1627 obs.)		Test data (991 obs.)	
Statistic	$r = 1$	$r = 3$	$r = 1$	$r = 3$
$\chi^2(\hat{\beta})$	42.40	18.90	53.10	18.70
$W_0(\hat{\beta})$	-0.93	0.10	-0.73	1.64
$W_1(\hat{\beta})$	0.28	-0.26	0.01	16.50

Acknowledgement

We would like to express our gratitude to Professor Sid Yakowitz for providing us with the rainfall-runoff data and for his 1987 reference. We also would like to thank the referees for their useful and constructive remarks. This work was supported by grants AFOSR-89-0049, ONR-89-J-1051, and NSF CDR-88-03012.

References

Andersen, P. K. and Gill, R. D. (1982). Cox's regression model for counting processes: A large sample study. *Ann. Statist.* **10**, 1100–1120.

Arjas, E. and Haara, P. (1984). A marked point process approach to censored failure data with complicated covariates. *Scand. J. Statist.* **11**, 193–209.

Arjas, E. and Haara, P. (1987). A logistic regression model for hazard: Asymptotic results. *Scand. J. Statist.* **14**, 1–18.

Basawa, I. and Scott, D. (1983). *Asymptotic Optimal Inference for Non-Ergodic Models. Lecture Notes in Statist.* **17**, Springer-Verlag, New York.

Berkson, J. (1944). Application of the logistic function to bio-assay. *J. Amer. Statist. Assoc.* **39**, 357–365.

Chiu, L. S. and Kedem, B. (1990). Estimating the exceedance probability of rain rate by logistic regression. *J. Geophys. Res.* **95**, 2217–2227.

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

Cox, D. R. (1972). Regression models and life tables. *J. Roy. Statist. Soc. Ser.B* **34**, 187–202.

Cox, D. R. (1975). Partial likelihood. *Biometrika* **62**, 69–76.

Fahrmeir, L. (1992). State space modeling and conditional mode estimation for categorical time series. In IMA volume *New Directions in Time Series Analysis* (Edited by D. Brillinger et al.), 87–109, Springer-Verlag, New York.

Fahrmeir, L. and Kaufmann, H. (1987). Regression models for non-stationary categorical time series. *J. Time Series Anal.* **8**, 147–160.

Gail, M. H., Santner, T. J. and Brown, C. C. (1980). An analysis of comparative carcinogenesis experiments based on multiple times to tumor. *Biometrics* **36**, 255–266.

Kaufmann, H. (1987). Regression models for nonstationary categorical time series: Asymptotic estimation theory. *Ann. Statist.* **15**, 79–98.

Kedem, B. (1980). *Binary Time Series.* Dekker, New York.

Kedem, B. and Pavlopoulos, H. (1991). On the threshold method for rainfall estimation: Choosing the optimal threshold level. *J. Amer. Statist. Assoc.* **86**, 626–633.

Keenan, D. M. (1982). A time series analysis of binary data. *J. Amer. Statist. Assoc.* **77**, 816–821.

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.

Muenz, L. R. and Rubinstein, L. V. (1985). Markov models for covariate dependence of binary sequences. *Biometrics* **41**, 91–101.

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

Schoenfeld, D. (1980). Chi-squared goodness-of-fit tests for the proportional hazards regression model. *Biometrika* **67**, 145–153.

Slud, E. (1992). Partial likelihood for continuous-time stochastic processes. *Scand. J. Statist.* **19**, 97–109.

Slud, E. (1993). *Martingale Methods in Statistics*. Forthcoming book.

Slud, E. and Kedem, B. (1988). Partial likelihood analysis of time series models with application to rainfall-runoff data. Report MD88-11-ES/BK, Department of Mathematics, University of Maryland, College Park.

Wong, W. H. (1986). Theory of partial likelihood. *Ann. Statist.* **14**, 88–123.

Yakowitz, S. (1987). Nearest-neighbour methods for time series analysis. *J. Time Series Anal.* **8**, 235–247.

Mathematics Department, University of Maryland, College Park, MD 20742, U.S.A.
Mathematics Department and Institute for Systems Research, University of Maryland, College Park, MD 20742, U.S.A.

(Received November 1991; accepted July 1993)