Prediction and Classification of Non-stationary Categorical Time Series

by K. Fokianos, B. Kedem

T.R. 96-49

ISR
INSTITUTE FOR SYSTEMS RESEARCH

Sponsored by
the National Science Foundation
Engineering Research Center Program,
the University of Maryland,
Harvard University,
and Industry

Prediction and Classification of Non-stationary Categorical Time Series

Konstantinos Fokianos
Benjamin Kedem
Department of Mathematics
and
Institute for Systems Research
University of Maryland,
College Park, MD 20742, USA

February 1996

Abstract

Partial Likelihood analysis of a general regression model for the analysis of non-stationary categorical time series is presented, taking into account stochastic time dependent covariates. The model links the probabilities of each category to a covariate process through a vector of time invariant parameters. Under mild regularity conditions, we establish good asymptotic properties of the estimator by appealing to martingale theory. Certain diagnostic tools are presented for checking the adequacy of the fit.

Keywords: Non-stationarity, classification, prediction, asymptotic theory, partial likelihood, goodness of fit.

1 Introduction

Categorical time series arise in numerous applications many of which are reported in the recent books by [8], [10], and [15][Ch. 9]. Examples of categorical time series include signals quantized at several levels, clipped binary time series, and any multi-response longitudinal data observed on an ordinal or nominal scale. And just as with "ordinary" time series the problem of forecasting or prediction in categorical series is of importance, except that usually it concerns the estimation of a future transition probability given past data and auxiliary information. In this regard, the prediction problem is essentially synonymous with the problem of classification of a future value in one of several categories given the past. In some cases, the complete dependence structure is known thus making statistical inference relatively easy. For example, when the series can be regarded as a homogeneous Markov chain, the inference problem can be attacked using the methods of [4], but when the complete dependence structure is unknown, the problem becomes quite challenging.

Recent advances in categorical time series owe greatly to the introduction of generalized linear models and *link* functions as described in [21]. Accordingly, the one step transition probability is conveniently parametrized via the link, and this goes along well with conditional inference, allowing for some form of non-stationarity. Conditional inference where a Markov assumption is made can be found in [17],[16], [29] [22], [5], [9], [13], [20], [19] to name only a few.

This paper is a generalization of [28] who only dealt with logistic regression for binary time series. We perform conditional inference using *partial likelihood*, a concept introduced by [7], and extended and ramified by [31], [27]. Partial likelihood simplifies conditional inference–for example, it obviates the Markov assumption–and is particularly useful for time series where the dependence is unknown, let alone the knowledge of joint distributions. Furthermore, as noted by [23], partial likelihood inference allows missing values. Indeed, all that is needed is a nested sequence of histories.

Following [31] and [27], we first give the definition of partial likelihood, and then setup the model. We next discuss the large sample theory. This is followed by some diagnostic tools.

2 The Mathematical Setup

2.1 Partial Likelihood

Assume that an individual observes a stochastic process, say (x_t, y_t), $t = 1, \ldots, N$. In principle, we can write down the joint distribution of all the observations up to time N, by employing the law of total probability; that is ([31])

$$f(x_1, y_1, x_2, y_2, \ldots, x_N, y_N) = [\prod_{t=1}^{N} f(y_t \mid d_t)][\prod_{t=1}^{N} f(x_t \mid c_t)] \tag{1}$$

where $d_t = (y_1, x_1, \ldots, y_{t-1}, x_{t-1})$ and $c_t = (y_1, x_1, \ldots, y_{t-1}, x_{t-1}, y_t)$.

[7] defined the second product on the right hand side of (1) as the *Partial Likelihood.* It is helpful to note that the σ-field generated by c_{t-1} is contained in the one generated by c_t. This is a key feature which motivates our definition(see [27], and [28]).

Definition 2.1 *Let $\mathcal{F}_t$, $t = 0, 1, \ldots$ be an increasing sequence of σ-fields, $\mathcal{F}_0 \subset \mathcal{F}_1 \subset \mathcal{F}_2 \ldots$, and let $X_1, X_2, \ldots$ be a sequence of random variables on some common probability space such that X_t is $\mathcal{F}_t$ measurable. Denote the density of X_t given $\mathcal{F}_{t-1}$ by $f_t(x_t; \beta)$, where $\beta \in R^p$ is a parameter. The partial likelihood (PL) function relative to β, $\mathcal{F}_t$, and the data $X_1, X_2, \ldots, X_N$, is given by the product*

$$PL(\beta; X_1, \ldots, X_N) = \prod_{t=1}^{N} f_t(x_t; \beta) \tag{2}$$

This definition generalizes both likelihood and conditional likelihood. Unlike (full) likelihood, partial likelihood does not require complete knowledge of the joint distribution of the covariates. Unlike conditional likelihood, complete covariate information need not be known throughout the period of observation. Partial likelihood takes into account only what is known to the observer up to the time of actual observation.

The vector β that maximizes (2) is called the maximum partial likelihood estimator (MPLE). Its asymptotic distribution has been studied by several authors (see [31] ; [28]). In the context of survival analysis and counting processes see [2], [3], for example. The key point is that the gradient of the logarithm of (2) is a martingale with respect to the nested sequence of histories $\mathcal{F}_t$.

2.2 A General Model

We introduce now some notation and terminology which will be found useful in the sequel. Suppose that we actually observe a non-stationary categorical time series, say $\{y_t,\ t = 0, 1, \ldots, N\}$. Let m denote the number of possible categories and assume that the $t'th$ observation is given by a vector $y_t = (y_{t1}, \ldots, y_{tq})'$ of length $q = m - 1$, where

$$y_{tj} = \begin{cases} 1 & \text{if the } j \overset{th}{=} \text{ category is observed at time } t \\ 0 & \text{otherwise} \end{cases}$$

Let $p_t = (p_{t1}, \ldots p_{tq})'$ denote the corresponding vector of conditional probabilities given $\mathcal{F}_{t-1}$. In other words $p_{tj} = P(y_{tj} = 1 \parallel \mathcal{F}_{t-1})$ for $j = 1, \ldots, q$. The σ-algebra $\mathcal{F}_{t-1}$ represents the whole available information to the observer up to and including time t. For the $m'th$ category, put

$$y_{tm} = 1 - \sum_{j=1}^{q} y_{tj} \tag{3}$$

and

$$p_{tm} = 1 - \sum_{j=1}^{q} p_{tj} \tag{4}$$

Assume that $\mathbf{Z}_{t-1}$ is a $p \times q$ matrix that represents a covariate process. In other words each response y_{tj} corresponds to a vector of *random time dependent covariates*, say $z_{(t-1)j}$, which is the $j \overset{th}{=}$ column of $\mathbf{Z_{t-1}}$. The covariate matrix usually consists of any lagged values of the response process and (or) any exogenous variables that evolve in time simultaneously with the response variable. Moreover lagged values of the exogenous variables are allowed as well as any interactions between the response and the covariates.

The aim of this paper is to develop an asymptotic theory for a flexible and parsimonious class of models that *link* the probability of the $j \overset{th}{=}$ category with the covariate process in a certain way. This leads to an attractive parametrization, which extends ideas from the generalized linear models (GLIM) and the autoregressive moving average models (ARMA) ([6]).

Define (see [9], [13])

$$p_t = h(\mathbf{Z}'_{t-1}\beta) \tag{5}$$

Here β denotes a p dimensional vector of time invariant unknown parameters which belongs to an open set $B \subseteq R^p$. The function h is called the *link function.* We assume that the link function maps a subset $H \subseteq R^q$ bijectively onto $\{(w_1, \ldots, w_q)' : \; w_j > 0, \; j = 1, \ldots, q, \; \sum_{j=1}^{q} w_j < 1\}$. Note that the multinomial logits and the cumulative odds models fall in this category ([1]).

In our context, since each component of y_t takes the values 0 or 1, we have the multinomial probability

$$f(y_t; \beta \mid \mathcal{F}_{t-1}) = \prod_{j=1}^{m} p_{tj}(\beta)^{y_{tj}} \tag{6}$$

Consequently, the corresponding Partial Likelihood is:

$$PL(\beta) \;\; = \;\; \prod_{t=1}^{N} f(y_t; \beta \mid \mathcal{F}_{t-1})$$

$$= \prod_{t=1}^{N}\prod_{j=1}^{m} p_{tj}(\beta)^{y_{tj}} \tag{7}$$

It follows that the partial log-likelihood is given by

$$pl_N(\beta) = \sum_{t=1}^{N}\sum_{j=1}^{m} y_{tj}\log p_{tj}(\beta) \tag{8}$$

The partial score is given by the vector

$$S_N(\beta) = (\frac{\partial pl_N(\beta)}{\partial\beta_1},\ldots,\frac{\partial pl_N(\beta)}{\partial\beta_p})' \tag{9}$$

It follows that

$$S_N(\beta) = \sum_{t=1}^{N} \mathbf{Z}_{t-1}\mathbf{D}_{t-1}(\beta)(y_t - p_t(\beta)) \tag{10}$$

where $\mathbf{D}_{t-1}(\beta) = [\partial d(\mathbf{Z}'_{t-1}\beta)/\partial\gamma_{t-1}]$ with $\gamma_{t-1} = \mathbf{Z}'_{t-1}\beta$. The function d is defined as the composition of the functions h and l with l standing for the logits function. This function is defined by

$$l(p_t) = (\log(p_{t1}/p_{tm}),\ldots,\log(p_{tq}/p_{tm}))$$

The conditional information matrix is given by

$$\begin{aligned}\mathbf{G}_N(\beta) &= \sum_{t=1}^{N} Cov[\mathbf{Z}_{t-1}\mathbf{D}_{t-1}(\beta)(y_t - p_t(\beta)) \mid \mathcal{F}_{t-1}] \\ &= \sum_{t=1}^{N} \mathbf{Z}_{t-1}\mathbf{D}_{t-1}(\beta)\mathbf{\Sigma}_t\mathbf{D}'_{t-1}(\beta)\mathbf{Z}'_{t-1} \end{aligned}\tag{11}$$

with $\mathbf{\Sigma}_t(\beta)$ is the conditional covariance matrix of y_t with generic element

$$\sigma_t^{(ij)}(\beta) = \begin{cases} -p_{ti}(\beta)p_{tj}(\beta) & \text{if } i \neq j \\ p_{ti}(\beta)(1-p_{ti}(\beta)) & \text{if } i = j \end{cases}$$

for $i, j = 1\ldots, q$. The unconditional information matrix is:

$$\mathbf{F}_N(\beta) = E[\mathbf{G}_N(\beta)] \tag{12}$$

Finally, the second derivative of the partial log likelihood multiplied by -1, is

$$\mathbf{H}_N(\beta) = -\frac{\partial^2 pl_N(\beta)}{\partial\beta\partial\beta'} = \mathbf{I}_N(\beta) - \mathbf{R}_N(\beta) \tag{13}$$

where

$$\mathbf{R}_N(\beta) = \sum_{t=1}^{N}\sum_{r=1}^{q} \mathbf{Z}_{t-1}\mathbf{W}_{(t-1)r}(\beta)\mathbf{Z}'_{t-1}(y_{tr} - p_{tr}(\beta))$$

with $\mathbf{W}_{(t-1)r}(\beta) = [\partial^2 d_r(\mathbf{Z}'_{t-1}\beta)/\partial\gamma_{t-1}\partial\gamma'_{t-1}]$. Notice, that the expectation and variance have been taken above with respect to the true parameter. The maximum Partial Likelihood Estimator is the consistent solution of the $S_N(\beta) = 0$. We point out that existence and uniqueness of the estimator is not guaranteed for finite samples. However, we obtain concavity of the log-likelihood for many important applications. In the setting of independent observation these questions have been studied by several authors. Among them are [11], [30], [24], [26], [14].

3 Large Sample Theory

We prove now existence, consistency and asymptotic normality of the MPLE under regularity conditions. We will consistently suppress any notation which depends on the true parameter.

Assumption (A)

A.1 The parameter β, belongs to an open set $B \subseteq R^p$.

A.2 The covariate matrix $\mathbf{Z}_{t-1}$ almost surely lie in a nonrandom compact subset Γ of $R^{p\times q}$ such that $P[\sum_{t=1}^{N}\mathbf{Z}_{t-1}\mathbf{Z}'_{t-1} > \mathbf{0}] = 1$. Furthermore we assume that $\mathbf{Z}'_{t-1}\beta$ lies almost surely in the domain H of h for all $\mathbf{Z}_{t-1} \in \Gamma$ and $\beta \in B$.

A.3 The probability measure P which governs $\{y_t, \mathbf{Z}_{t-1}\}$, $t = 1, \ldots, N$ obeys (5) with $\beta = \beta_0$.

A.4 The link function h is twice continuously differentiable, $\det[\partial h(\gamma)/\partial\gamma] \neq 0$

A.5 There is a probability measure μ on $R^{p\times q}$ such that $\int_{R^{p\times q}} \mathbf{Z}\mathbf{Z}'\mu(d\mathbf{Z})$ is positive definite, such that under (5) with $\beta = \beta_0$, for Borel sets $A \subset R^{p\times q}$ we have

$$\frac{1}{N}\sum_{t=1}^{N} I_{[\mathbf{Z}_{t-1}\in A]} \xrightarrow{p} \mu(A) \ \ , as \ \ N \to \infty.$$

Assumption **A.1** and **A.4** guarantee that the second derivative of the partial log-likelihood is a continuous function of β. The Condition $\det[\partial h(\gamma)/\partial\gamma] \neq 0$, implies in particular that $\mathbf{D}_{t-1}$ is not singular, so from **A.2** the conditional information matrix is positive definite with probability 1. To see this note that for any vector $\lambda \in R^p$

$$\begin{aligned} \lambda' \mathbf{G}_N \lambda &= \lambda' (\sum_{t=1}^{N} \mathbf{Z}_{t-1} \mathbf{D}_{t-1} \mathbf{\Sigma}_t \mathbf{D}'_{t-1} \mathbf{Z}'_{t-1}) \lambda \\ &\geq \min_t \lambda_{\min} (\mathbf{D}_{t-1} \mathbf{\Sigma}_t \mathbf{D}'_{t-1}) (\lambda' \sum_{t=1}^{N} \mathbf{Z}_{t-1} \mathbf{Z}'_{t-1} \lambda') \\ &> 0 \end{aligned}$$

with $\lambda_{\min}$ denoting the minimum eigenvalue. The claim is true. Since the variance-covariance matrix is positive definite and the matrix of derivatives, $\mathbf{D}_{t-1}$, is not singular we have that the minimum eigenvalue is positive almost everywhere. It follows that the unconditional information matrix is positive definite as well. The last part of assumption **A.2** assures that we have a well defined model. The compactness assumption will be useful in deriving bounds for the asymptotics. Assumption **A.5** simply states that if g is any continuous and bounded function on Γ taking values on $R^{p \times q}$ then we have that

$$\frac{\sum_{t=1}^{N} g(\mathbf{Z}_{t-1})}{N} \xrightarrow{p} \int_{R^{p \times q}} g(\mathbf{Z}) \mu(d\mathbf{Z})$$

Thus the conditional information matrix, $\mathbf{G}_N(\beta)$ has a non-random limit

$$\frac{\mathbf{G}_N(\beta)}{N} \xrightarrow{p} \int_{R^{p \times q}} \mathbf{Z} \mathbf{D}(\beta) \mathbf{\Sigma}(\beta) \mathbf{D}'(\beta) \mathbf{Z}' \mu(d\mathbf{Z}) = \mathbf{G}(\beta) \tag{14}$$

where $\mathbf{D}(\beta) = [\partial h(\mathbf{Z}'\beta) / \partial(\mathbf{Z}'\beta)]$ and $\mathbf{\Sigma}$ has generic element

$$\sigma^{(ij)}(\beta) = \begin{cases} -h_i(\mathbf{Z}'\beta) h_j(\mathbf{Z}'\beta) & \text{if } i \neq j \\ h_i(\mathbf{Z}'\beta)(1 - h_i(\mathbf{Z}'\beta)) & \text{if } i = j \end{cases}$$

for $i, j = 1 \ldots, q$. Note that integration with respect to a matrix, means that we integrate with respect to each element of the matrix. From **(A.4)** $\mathbf{G}(\beta)$ is a positive definite matrix at the true value and therfore its inverse exist. It is important to emphasize that our approach is quite general and does not call for any Markov assumption (compare with [9]; [13]).

At this point , we want to mention that we will use the right Cholesky square root of a positive definite matrix in the sequel. More precisely, if $\mathbf{B}$ is a positive definite matrix then the right Cholesky square root, denoted by $\mathbf{B}^{\frac{1}{2}}$, is defined as the unique upper triangular matrix with positive elements such that $\mathbf{B} = (\mathbf{B}^{\frac{1}{2}})'(\mathbf{B}^{\frac{1}{2}})$. We denote by $\mathbf{B}^{\frac{t}{2}} = (\mathbf{B}^{\frac{1}{2}})'$. Our proof of consistency and asymptotic normality is based on the classical approach of Cramer, namely we first exhibit a solution of the score equations and then prove that is consistent and asymptotically normally distributed.

At this end, we will need some helpful Lemmas. The following lemma shows that the partial score process is a zero mean square integrable martingale which satisfies the conditions for an application of a martingale central limit theorem.

Lemma 3.1 *Consider the model (5) and assume that assumption* **(A)** *holds. Then the partial score process* $\{S_t, \mathcal{F}_t\}$ *is a zero mean square integrable martingale such that :*

$$\mathbf{F}_N^{-1/2} S_N \xrightarrow{D} \mathcal{N}$$

as $N \to \infty$ *,with* $\mathcal{N}$ *denoting a standard normal random vector.*

Proof: The fact that the partial score process is zero mean square integrable martingale follows from (10) and assumption **A.2**. To show it actually converges in distribution, we consider $\phi_N = \lambda' S_N$, with $\lambda \in R^p$, having in mind the Cramer-Wald device. Then ϕ_N is a univariate zero-mean martingale. Its conditional and unconditional covariance matrices are $\lambda' \mathbf{G}_N \lambda$ and $\lambda' \mathbf{F}_N \lambda$ respectively. Thus

$$\frac{\lambda' \mathbf{G}_N \lambda}{\lambda' \mathbf{F}_N \lambda} = \frac{\lambda' \mathbf{G}_N \lambda / N}{\lambda' \mathbf{F}_N \lambda / N} \xrightarrow{p} \frac{\lambda' \mathbf{G} \lambda}{\lambda' \mathbf{G} \lambda} = 1$$

upon invoking **A.2** and **A.5**. Furthermore, by letting $I_{Nt}(\epsilon)$ to be the indicator of the set $\{|\lambda' a_t|^2 \geq (\lambda' \mathbf{F}_N \lambda)^{\frac{1}{2}} \epsilon\}$ with $a_t = ps_t - ps_{t-1}$, we get

$$\begin{aligned} \frac{1}{\lambda' \mathbf{F}_N \lambda} \sum_{t=1}^{N} E[|\lambda' a_t|^2 I_{Nt}(\epsilon) \parallel \mathcal{F}_{t-1}] &\leq \frac{1}{(\lambda' \mathbf{F}_N \lambda)^{\frac{3}{2}} \epsilon} \sum_{t=1}^{N} E[|\lambda' a_t|^3 \parallel \mathcal{F}_{t-1}] \\ &\leq \frac{N M_1}{(\lambda' \mathbf{F}_N \lambda)^{\frac{3}{2}} \epsilon} \end{aligned}$$

where M_1 is a bound. Such a bound exists from **A.2**. Therefore Lindeberg's condition holds since the right hand side of the above tends to zero. The conclusion of the Lemma follows, by appealing to the Central Limit Theorem for martingales ([12, Corollary 3.1]). □

The next lemma, a consequence of the Lindeberg's condition, parallels the well-known result from linear models ([18])

Lemma 3.2 *Under* **(A)** *we have that*

$$\lambda_{\min}(\mathbf{F}_N) \to \infty$$

as $N \to \infty$, *where* $\lambda_{\min}$ *is the minimum eigenvalue of the unconditional information matrix.*

Proof: Recall that if $\mathbf{A}$ and $\mathbf{B}$ are positive definite matrices,

$$|\lambda_{\min}(\mathbf{A}) - \lambda_{\min}(\mathbf{A})| \leq c\|\mathbf{A} - \mathbf{B}\| \tag{15}$$

where the positive constant depends only on the norm of the matrix. Then, by the proof of Lemma 3.1, we get

$$|\lambda_{\min}(\frac{\mathbf{F}_N}{N}) - \lambda_{\min}(\mathbf{G})| \rightarrow \mathbf{0}$$

It follows that $\lambda_{\min}(\mathbf{F}_N) = \bigcirc(N)$ and the claim is true. □

We prove now a continuity condition. Namely, we would like to have the matrix of second derivatives as close as possible to the information matrix. This is a technical lemma and the proof is along the lines of [13].

Lemma 3.3 *Under* $(\mathbf{A})$ *the following continuity condition holds*

$$\sup_{\tilde{\beta} \in O_N(\delta)} \|\mathbf{F}_N^{-\frac{1}{2}}(\mathbf{H}_N(\tilde{\beta}) - \mathbf{G}_N)\mathbf{F}_N^{-\frac{t}{2}}\| \xrightarrow{p} 0$$

with $O_N(\delta) = \{\tilde{\beta} : \|\mathbf{F}_N^{\frac{t}{2}}(\tilde{\beta} - \beta)\| \leq \delta\}$, *holds for any* $\delta > 0$.

Proof: Let $\lambda \in R^p$, with $\lambda \neq 0$ and assume without loss of generality that $\|\lambda\| = 1$. We will show the equivalent condition, for any $\delta > 0$

$$\sup_{\tilde{\beta} \in O_N(\delta)} \lambda' \mathbf{F}_N^{-\frac{1}{2}}(\mathbf{H}_N(\tilde{\beta}) - \mathbf{G}_N)\mathbf{F}_N^{-\frac{t}{2}}\lambda \xrightarrow{p} 0 \tag{16}$$

using once more Cramer-Wold Device. By decomposing $\mathbf{H}_N(\tilde{\beta}) = \mathbf{G}_N(\tilde{\beta}) - \mathbf{R}_N(\tilde{\beta})$ we need really to show

$$g_N = \sup_{\tilde{\beta} \in O_N(\delta)} \lambda' \mathbf{F}_N^{-\frac{1}{2}}(\mathbf{G}_N(\tilde{\beta}) - \mathbf{G}_N)\mathbf{F}_N^{-\frac{t}{2}}\lambda \xrightarrow{p} 0 \tag{17}$$

and

$$\sup_{\tilde{\beta} \in O_N(\delta)} \lambda' \mathbf{F}_N^{-\frac{1}{2}}\mathbf{R}_N(\tilde{\beta})\mathbf{F}_N^{-\frac{t}{2}}\lambda \xrightarrow{p} 0 \tag{18}$$

hold simultaneously. Define the vectors $w_{(t-1)N} = \lambda' \mathbf{F}_N^{-\frac{1}{2}}\mathbf{Z}_{t-1}$, for $1 \leq t \leq N$, and $w_N = \sum_{t=1}^N w'_{(t-1)N} w_{(t-1)N}$. Then we have that

$$g_N = \sup_{\tilde{\beta} \in O_N(\delta)} \sum_{t=1}^{N} w'_{(t-1)N}(\mathbf{L}_{t-1}(\tilde{\beta}) - \mathbf{L}_{t-1})w_{(t-1)N}$$

where $\mathbf{L}_{t-1}(\beta) = \mathbf{D}_{t-1}\mathbf{\Sigma}_t\mathbf{D}'_{t-1}$ for $t = 1, \ldots, N$. It follows that

$$g_N \leq w_N \sup_{\tilde{\beta}\in O_N(\delta),t} \|\mathbf{L}_{t-1}(\tilde{\beta}) - \mathbf{L}_{t-1}\|$$

Using **A.2** , $\sup_t \|\mathbf{L}_{t-1}(\tilde{\beta}) - \mathbf{L}_{t-1}\|$ can be estimated from above by a continuous function of $\tilde{\beta}$ with a zero at $\tilde{\beta} = \beta$. Notice that $\{O_N(\delta)\}$ shrinks to β. Hence

$$\sup_{\tilde{\beta}\in O_N(\delta),t} \|\mathbf{L}_{t-1}(\tilde{\beta}) - \mathbf{L}_{t-1}\| \to 0$$

By applying Markov's inequality we have that

$$\begin{aligned} P[|g_N| \geq \epsilon] &\leq \frac{E[g_N]}{\epsilon} \\ &\leq E[w_N] \sup_{\tilde{\beta}\in O_N(\delta),t} \|\mathbf{L}_{t-1}(\tilde{\beta}) - \mathbf{L}_{t-1}\| \\ &= \lambda'[\frac{\mathbf{F}_N}{N}]^{-\frac{1}{2}} \frac{\sum_{t=1}^N E[\mathbf{Z}_{t-1}\mathbf{Z}'_{t-1}]}{N} [\frac{\mathbf{F}_N}{N}]^{-\frac{t}{2}} \lambda \\ &\quad \sup_{\tilde{\beta}\in O_N(\delta),t} \|\mathbf{L}_{t-1}(\tilde{\beta}) - \mathbf{L}_{t-1}\| \to 0 \end{aligned}$$

since the other terms converge to a limit by the continuity of the square root and the assumption **A.5**. By further decomposition we obtain

$$\sup_{\tilde{\beta}\in O_N(\delta)} \sum_{t=1}^N w'_{(t-1)N}(\mathbf{W}_{(t-1)j}(\tilde{\beta}) - \mathbf{W}_{(t-1)j})w_{(t-1)N}(y_{tj} - p_{tj}) \xrightarrow{p} 0 \tag{19}$$

$$\sup_{\tilde{\beta}\in O_N(\delta)} \sum_{t=1}^N w'_{(t-1)N}\mathbf{W}_{(t-1)j}(\tilde{\beta})w_{(t-1)N}(p_{tj} - p_{tj}(\tilde{\beta})) \xrightarrow{p} 0 \tag{20}$$

$$\sum_{t=1}^N w'_{(t-1)N}\mathbf{W}_{(t-1)j}w_{(t-1)N}(y_{tj} - p_{tj}) \xrightarrow{p} 0 \tag{21}$$

for any j, $1 \leq j \leq q$, jointly are sufficient for (18). The proofs of (19), (20), are the same as that of (17). To prove (21), consider the increments of (21), that is

$$u_{(t-1)N} = w'_{(t-1)N}\mathbf{W}_{(t-1)j}w_{(t-1)N}(y_{tj} - p_{tj})$$

Then we see, that

$$E[u_{(t-1)N} \parallel \mathcal{F}_{t-1}] = 0$$

and

$$\begin{aligned} Var[u_{(t-1)N} \parallel \mathcal{F}_{t-1}] &= w'_{(t-1)N}\mathbf{W}_{(t-1)j}w_{(t-1)N} \\ & \quad Var[y_{tj} - p_{tj} \parallel \mathcal{F}_{t-1}]w'_{(t-1)N}\mathbf{W}'_{(t-1)j}w_{(t-1)N} \\ &\leq K(w'_{(t-1)N}w_{(t-1)N})^2 \end{aligned}$$

where K is a bound on $\|\mathbf{W}_{(t-1)j}\|^2 Var[y_{tj} - p_{tj} \parallel \mathcal{F}_{t-1}]$. Actually, the above two relations make clear that $\{u_{(t-1)N},\ t = 1, \ldots, N\}$ are the orthogonal increments of a square integrable zero mean martingale. It follows that

$$E(\sum_{t=1}^{N} u_{(t-1)N}) = 0$$

and

$$\begin{aligned} Var[\sum_{t=1}^{N} u_{(t-1)N}] &\leq K\sum_{t=1}^{N} E[w'_{(t-1)N}w_{(t-1)N}] \\ &\leq K \sup_t E[w'_{(t-1)N}w_{(t-1)N}]E[w_N] \end{aligned}$$

However

$$\begin{aligned} \sup_t E(w'_{(t-1)N}w_{(t-1)N}) &= \sup_t \lambda'\mathbf{F}_N^{-\frac{1}{2}}E(\mathbf{Z}_{t-1}\mathbf{Z}'_{t-1})\mathbf{F}_N^{-\frac{t}{2}}\lambda \\ &\leq \lambda'\mathbf{F}_N^{-1}\lambda \sup_{\mathbf{Z}_{t-1}\in\Gamma} \|E(\mathbf{Z}_{t-1})\|^2 \\ &\leq \frac{\sup_{\mathbf{Z}_{t-1}\in\Gamma}\|E(\mathbf{Z}_{t-1})\|^2}{\lambda_{\min}(\mathbf{F}_N)} \to 0 \end{aligned}$$

Since $E[w_N]$ is bounded, from its convergence, relation (21) holds and therefore the continuity condition was established. □

We prove now the main result of this section.

Theorem 3.1 *Under* (**A**), *the probability that a locally unique maximum partial likelihood estimator exists converges to one. Moreover there exists a sequence of maximum partial likelihood estimators $\hat{\beta}_N$ which is consistent and asymptotically Normal.*

$$\sqrt{N}(\hat{\beta}_N - \beta_0) \xrightarrow{D} \mathcal{N}(0, \mathrm{G}^{-1}(\beta_0))$$

Proof: From Lemma 3.1 we get that

$$(\mathbf{F}_N^{-\frac{1}{2}}S_N,\ \ \mathbf{F}_N^{-\frac{1}{2}}\mathbf{G}_N\mathbf{F}_N^{-\frac{1}{2}}) \xrightarrow{D} (\mathcal{N},\mathbf{I})$$

where $\mathcal{N}$ is a standard normal vector . Choosing $\mathbf{G}_N^{\frac{1}{2}}$ such that $\mathbf{F}_N^{-\frac{1}{2}}\mathbf{G}_N^{\frac{1}{2}}$ to be the Cholesky square root of $\mathbf{F}_N^{-\frac{1}{2}}\mathbf{G}_N\mathbf{F}_N^{-\frac{1}{2}}$, we have from the continuity of the square root that

$$(\mathbf{F}_N^{-\frac{1}{2}}S_N,\ \ \mathbf{F}_N^{-\frac{1}{2}}\mathbf{G}_N^{\frac{1}{2}}) \xrightarrow{D} (\mathcal{N},\mathbf{I})$$

where $\mathcal{N}$ is as above.
We first prove asymptotic existence and consistency. By Taylor expansion we have that

$$pl_N(\tilde{\beta}) = pl_N(\beta_0) + (\tilde{\beta}-\beta_0)'S_N - \frac{1}{2}(\tilde{\beta}-\beta_0)'\mathbf{H}_N(\tilde{\tilde{\beta}})(\tilde{\beta}-\beta_0)$$

where $\tilde{\tilde{\beta}}$ lies between $\tilde{\beta}$ and β_0. Equivelantly

$$pl_N(\tilde{\beta}) - pl_N(\beta_0) = (\tilde{\beta}-\beta_0)'S_N - \frac{1}{2}(\tilde{\beta}-\beta_0)'\mathbf{H}_N(\tilde{\tilde{\beta}})(\tilde{\beta}-\beta_0) \tag{22}$$

Let now $\tilde{\lambda} = \mathbf{F}_N^{\frac{t}{2}}(\tilde{\beta}-\beta_0)/\delta$. Then it follows by choosing δ such that $\tilde{\lambda}'\tilde{\lambda} = 1$ that $(\tilde{\beta}-\beta_0)' = \tilde{\lambda}'\mathbf{F}_N^{-\frac{1}{2}}\delta$. Substituting into (22), we have

$$pl_N(\tilde{\beta}) - pl_N(\beta_0) = \delta\tilde{\lambda}'\mathbf{F}_N^{-\frac{1}{2}}S_N - \frac{\delta^2}{2}\tilde{\lambda}'\mathbf{F}_N^{-\frac{1}{2}}\mathbf{H}_N(\tilde{\tilde{\beta}})\mathbf{F}_N^{-\frac{t}{2}}\tilde{\lambda} \tag{23}$$

We are going to prove that for every $\eta > 0$ there exists N and δ such that

$$P[pl_N(\tilde{\beta}) - pl_N(\beta_0) < 0 \ \ \forall\tilde{\beta} \in \partial O_N(\delta)] \geq 1-\eta \tag{24}$$

This shows that, with probability tending to one, there exists a local maximum inside $O_N(\delta)$. From (23), we recognize that it is sufficient to show

$$P[\|\mathbf{F}_N^{-\frac{1}{2}}S_N\|^2 \leq \delta^2\frac{\lambda^2_{\min}(\mathbf{F}_N^{-\frac{1}{2}}\mathbf{H}_\mathbf{N}(\tilde{\tilde{\beta}})\mathbf{F}_N^{-\frac{t}{2}})}{4}] \geq 1-\eta \tag{25}$$

This is so because of the inequality

$$\tilde{\lambda}\mathbf{F}_N^{-\frac{1}{2}}S_N - \frac{\delta}{2}\tilde{\lambda}\mathbf{F}_N^{-\frac{1}{2}}\mathbf{H}_N(\tilde{\tilde{\beta}})\mathbf{F}_N^{-\frac{t}{2}}\tilde{\lambda} \leq \|\mathbf{F}_N^{-\frac{1}{2}}S_N\|^2 - \frac{\delta}{2}\lambda_{\min}(\mathbf{F}_N^{-\frac{1}{2}}\mathbf{H}_N(\tilde{\tilde{\beta}})\mathbf{F}_N^{-\frac{t}{2}})$$

Consequently we have that

$$P[\|\mathbf{F}_N^{-\frac{1}{2}} S_N\|^2 \leq \delta^2 \frac{\lambda^2_{\min}(\mathbf{F}_N^{-\frac{1}{2}} \mathbf{H}_N(\tilde{\tilde{\beta}})\mathbf{F}_N^{-\frac{t}{2}})}{4}] \geq 1 - \frac{E[\mathbf{F}_N^{-\frac{1}{2}} S_N\|^2]}{\delta^2 \frac{\lambda^2_{\min}(\mathbf{F}_N^{-\frac{1}{2}} \mathbf{H}_N(\tilde{\tilde{\beta}})\mathbf{F}_N^{-\frac{t}{2}})}{4}}$$

Since $E[\|\mathbf{F}_N^{-\frac{1}{2}} S_N\|^2] = p$ and the denominator is bounded, the above expression can become arbitrarily small. The last claim follows from 15 and lemma 3.3. Asymptotic existence therefore was established. More specifically, we have that there exists a sequence $\{\hat{\beta}_N\}$ of PMLE's such that for any $\eta > 0$, there is a δ, N_1 with

$$P[\hat{\beta}_N \in O_N(\delta)] \geq 1 - \eta \ \forall N \geq N_1 \tag{26}$$

From Lemmas 3.1 and 3.3 we obtain that $\mathbf{H}_N(\beta)$ is positive definite throughout $O_N(\delta)$ with probability converging to 1. Therefore the MPLE $\hat{\beta}_N$ is also locally unique. Consistency was established as well, upon noting

$$\begin{aligned} 1 - \eta &\leq P[\|\mathbf{F}_N^{\frac{t}{2}}(\hat{\beta}_N - \beta_0)\| \leq \delta] \\ &\leq P[\|\hat{\beta}_N - \beta_0\| \leq \frac{\delta}{\lambda_{\min}(\mathbf{F}_N)}]. \end{aligned}$$

We prove now asymptotic normality. By Taylor expansion around $\hat{\beta}_N$, and using the mean value theorem for multivariate function we obtain

$$S_N = \tilde{\mathbf{H}}_N(\hat{\beta}_N - \beta_0) \tag{27}$$

where $\tilde{\mathbf{H}}_N = \int_0^1 \mathbf{H}_N(\beta_0 + s(\hat{\beta}_N - \beta_0)ds$ and the integration is taken elementwise. We need to show that

$$\mathbf{F}_N^{-\frac{1}{2}} \tilde{\mathbf{H}}_N \mathbf{F}_N^{-\frac{t}{2}} \xrightarrow{p} \mathbf{I} \tag{28}$$

But

$$\begin{aligned} \mathbf{F}_N^{-\frac{1}{2}} \tilde{\mathbf{H}}_N \mathbf{F}_N^{-\frac{t}{2}} &= \mathbf{F}_N^{-\frac{1}{2}} (\tilde{\mathbf{H}}_N - \mathbf{G}_N) \mathbf{F}_N^{-\frac{t}{2}} + \mathbf{F}_N^{-\frac{1}{2}} \mathbf{G}_N \mathbf{F}_N^{-\frac{t}{2}} \\ &\xrightarrow{p} \mathbf{0} + \mathbf{I} = \mathbf{I} \end{aligned}$$

This is so because for $N \to \infty$, the MPLE is consistent so that $\tilde{\mathbf{H}}_N - \mathbf{G}_N$ behaves the same as $\mathbf{H}_N - \mathbf{G}_N$. Invoking Lemmas 3.1 and 3.3 we have that (28) holds. Therefore, from (27)

$$\mathbf{F}_N^{-\frac{1}{2}} S_N = (\mathbf{F}_N^{-\frac{1}{2}} \tilde{\mathbf{H}}_N \mathbf{F}_N^{-\frac{t}{2}})(\mathbf{F}_N^{\frac{t}{2}}(\hat{\beta}_N - \beta_0))$$

Thus

$$\mathbf{F}_N^{\frac{t}{2}}(\hat{\beta}_N - \beta_0) \to \mathcal{N}$$

But

$$\mathbf{G}_N^{\frac{t}{2}}(\hat{\beta}_N - \beta_0) = \mathbf{G}_N^{\frac{t}{2}}\mathbf{F}_N^{-\frac{t}{2}}\mathbf{F}_N^{\frac{t}{2}}(\hat{\beta}_N - \beta_0) \to \mathcal{N}$$

since $\mathbf{G}_N^{\frac{t}{2}}\mathbf{F}_N^{-\frac{t}{2}} \xrightarrow{p} \mathbf{I}$. From the continuity of the square root

$$\frac{\mathbf{G}_N^{\frac{t}{2}}}{\sqrt{N}} \xrightarrow{p} \mathbf{G}^{\frac{t}{2}}$$

An application of Slutsky's theorem yields to the conclusion of the theorem □.

Corollary 3.1 Under **(A)** we have

$$\sqrt{N}(\hat{\beta}_N - \beta_0) - \frac{1}{N}\mathbf{G}^{-1}S_N \xrightarrow{p} 0$$

Proof: Using again Slutsky's theorem and the continuity of the square root we obtain that

$$\frac{1}{N}\mathbf{G}^{-1}S_N \xrightarrow{D} \mathcal{N}(0, \mathbf{G}^{-1})$$

The claim follows from the previous theorem and Slutsky's theorem once again □.

Now, assume that each component of the link function is log-concave, that is $\log h_j$ is concave for every $j = 1, \dots, m$ with $h_m = 1 - \sum_{j=1}^{q} h_j$. It follows that the logarithm is concave and if the parameter space B is R^p we obtain the following:

Corollary 3.2 Suppose **(A)** holds. Assume further that $\log h_j$ is concave for $j = 1, \dots, m$. Then the probability that a unique maximum partial likelihood estimator exists converges to one. Any such sequence is consistent and asymptotically normal as in Theorem 3.1.

4 Goodness of fit Statistics

A question which arises naturally after every procedure involving regression is that of goodness of fit. Our approach is to classify the responses $\mathbf{y}_t$ according to mutually exclusive events in terms of the covariates $\mathbf{Z}_{t-1}$ (see [25]; [28]). Suppose that $A_1, \ldots, A_k$ constitute a partition of $R^{p \times q}$. For $l = 1, \ldots, k$ define

$$M_l = \sum_{t=1}^{N} I_{[\mathbf{Z}_{t-1} \in A_l]} y_t$$

and

$$E_l(\beta) = \sum_{t=1}^{N} I_{[\mathbf{Z}_{t-1} \in A_l]} p_t(\beta)$$

where I is the indicator of the set $\{\mathbf{Z}_{t-1} \in A_l\}$, for $l = 1, \ldots, k$. Let $M_N = (M_1', \ldots, M_k')'$, $E_N(\beta) = (E_1'(\beta), \ldots, E_k'(\beta))'$. If we let $I_{t-1} = (I_{[\mathbf{Z}_{t-1} \in A_1]}, \ldots, I_{[\mathbf{Z}_{t-1} \in A_k]})'$ we can see that

$$d_N(\beta) = M_N - E_N(\beta) = \sum_{t=1}^{N} I_{t-1} \otimes (y_t - p_t(\beta))$$

with $\otimes$ denotes Kronecker product. It follows that $d_N(\beta)$ is a zero mean square integrable martingale that satisfies all the conditions needed for an application of the Central Limit Theorem under our previous assumptions. Thus

$$\frac{d_N}{\sqrt{N}} \overset{p}{\to} \mathcal{N}(0, \mathbf{C})$$

where $\mathbf{C} = \oplus_{l=1}^{k} \mathbf{C}_l$, the direct sum of k matrices [1], and $\mathbf{C}_l$ is a $q \times q$ symmetric matrix given by

$$\mathbf{C}_l(\beta_0) = \begin{bmatrix} \int_{A_l} p_1(\beta_0)(1 - p_1(\beta_0))\mu(d\mathbf{Z}) & \cdots & -\int_{A_l} p_1(\beta_0)p_q(\beta_0)\mu(d\mathbf{Z}) \\ \vdots & \ddots & \vdots \\ -\int_{A_l} p_1(\beta_0)p_q(\beta_0)\mu(d\mathbf{Z}) & \cdots & \int_{A_l} p_q(\beta_0)(1 - p_q(\beta_0))\mu(d\mathbf{Z}) \end{bmatrix}$$

From the above result we have the following proposition:

Proposition 4.1 *As $N \to \infty$, the asymptotic distribution of the statistic*

$$\chi^2(\beta_0) = \frac{1}{N} \sum_{l=1}^{k} d_l'(\beta_0) \mathbf{C}_l^{-1}(\beta_0) d_l(\beta_0) \tag{29}$$

is chi-square with kq degrees of freedom.

We are going to demonstrate now another theorem which gives rise to another goodness of fit statistic.

[1] $\mathbf{A} \oplus \mathbf{B}$ creates a partitioned diagonal matrix, having $\mathbf{A}$, $\mathbf{B}$ on the main diagonal.

Theorem 4.1 Suppose that **(A)** hold. Let $A_1, \ldots, A_k$ be a partition of $R^{p\times q}$. Then we have as $N \to \infty$

1. $$\sqrt{N}(\frac{d'_N}{N}, (\hat{\beta}_N - \beta_0)) \overset{D}{\to} \mathcal{N}(0, \mathbf{\Gamma})$$

   where $\mathbf{\Gamma}$ is a square matrix of dimension $p + kq$

   $$\mathbf{\Gamma} = \begin{bmatrix} \mathbf{C} & \mathbf{B}' \\ \mathbf{B} & \mathbf{G}^{-1} \end{bmatrix}$$

   Here $\mathbf{C}$ is as in Proposition (29), $\mathbf{G}$ is the limiting $p \times p$ information matrix, and the $l \ \underline{\underline{th}}$ column of $\mathbf{B}$ is given by the matrix

   $$\mathbf{G}^{-1} \int_{A_l} \mathbf{ZD\Sigma}\mu(d\mathbf{Z})$$

2. We also have, as $N \to \infty$ that

   $$\frac{E_N(\hat{\beta}_N) - E_N(\beta_0)}{\sqrt{N}} - \sqrt{N}\mathbf{B}'\mathbf{G}(\hat{\beta}_N - \beta_0) \overset{P}{\to} 0$$

Proof: For proving (1) we only need to observe from Corollary 3.1 that for some integer N greater than N_0 we have that

$$\frac{1}{\sqrt{N}}(d'_N, (\hat{\beta} - \beta_0)) \overset{P}{\approx} \frac{1}{\sqrt{N}}(d'_N, \mathbf{G}^{-1}S_N) \tag{30}$$

Now we know that d_N and S_N are martingales which obey the conditions for an application of a Central Limit Theorem for martingales. It follows that jointly (using again the Cramer-Wold device) the vector on the right hand side of the above equation converges to normal as $N \to \infty$. We only need to compute the asymptotic covariance matrix of its component. We have

$$\frac{1}{N}\mathbf{G}^{-1}S_N \sum_{t=1}^{N} I_{[\mathbf{Z}_{t-1}\in A_l]}(y_t - p_t) = \frac{1}{N}\mathbf{G}^{-1} \sum_{s=1}^{N} \mathbf{Z}_{s-1}\mathbf{D}_{s-1}(y_s - p_s) \sum_{t=1}^{N} I_{[\mathbf{Z}_{t-1}\in A_l]}(y_t - p_t)$$

But for $s < t$

$$E[\mathbf{Z}_{s-1}\mathbf{D}_{s-1}(y_s - p_s)I_{[\mathbf{Z}_{t-1}\in A_l]}(y_t - p_t)] = E[\mathbf{Z}_{s-1}\mathbf{D}_{s-1}(y_s - p_s)I_{[\mathbf{Z}_{t-1}\in A_l]}E[(y_t - p_t) \parallel \mathcal{F}_{t-1}]] = 0$$

Therefore, we have from assumption **A.5** that

$$\begin{aligned} E[\frac{1}{N}\mathbf{G}^{-1}S_N\sum_{t=1}^{N} I_{[\mathbf{Z}_{t-1}\in A_l]}(y_t-p_t)] &= E[\frac{1}{N}\mathbf{G}^{-1}\sum_{s=1}^{N}\mathbf{Z}_{s-1}\mathbf{D}_{s-1}(y_s-p_s)\sum_{t=1}^{N} I_{[\mathbf{Z}_{t-1}\in A_l]}(y_t-p_t)] \\ &= E[\sum_{t=1}^{N} I_{[\mathbf{Z}_{t-1}\in A_l]}\mathbf{Z}_{t-1}\mathbf{D}_{t-1}\mathbf{\Sigma}_t] \xrightarrow{p} \mathbf{G}^{-1}\int_{A_l}\mathbf{Z}\mathbf{D}\mathbf{\Sigma}\mu d(\mathbf{Z}) \end{aligned}$$

The first part of the theorem follows. For proving the second part, we have by Taylor's expansion

$$\begin{aligned} E_N^l(\hat{\beta}_N) &\approx E_N^l(\beta_0) + [\frac{\partial E_N^l(\beta)}{\partial\beta}]_{\beta_0}(\hat{\beta}_N-\beta_0) + o_p(\|\hat{\beta}_N-\beta_0\|) \\ &= E_N^l(\beta_0) + [\sum_{t=1}^{N} I_{[\mathbf{Z}_{t-1}\in A_l]}\frac{\partial p_t}{\partial\beta}](\hat{\beta}_N-\beta_0) + o_p(\|\hat{\beta}_N-\beta_0\|) \\ &= E_N^l(\beta_0) + [\sum_{t=1}^{N}\mathbf{Z}_{t-1}I_{[\mathbf{Z}_{t-1}\in A_l]}\frac{\partial p_t}{\partial\gamma_{t-1}}](\hat{\beta}_N-\beta_0) + o_p(\|\hat{\beta}_N-\beta_0\|) \\ &= E_N^l(\beta_0) + [\sum_{t=1}^{N}\mathbf{Z}_{t-1}I_{[\mathbf{Z}_{t-1}\in A_l]}\frac{\partial p_t}{\partial l}\frac{\partial l}{\partial p_t}\frac{\partial p_t}{\partial\gamma_{t-1}}](\hat{\beta}_N-\beta_0) + o_p(\|\hat{\beta}_N-\beta_0\|) \\ &= E_N^l(\beta_0) + [\sum_{t=1}^{N} I_{[\mathbf{Z}_{t-1}\in A_l]}\mathbf{Z}_{t-1}\mathbf{D}_{t-1}\mathbf{\Sigma}_t](\hat{\beta}_N-\beta_0) + o_p(\|\hat{\beta}_N-\beta_0\|) \end{aligned}$$

where l is the logits function and $\gamma_{t-1}=\mathbf{Z}'_{t-1}\beta$. So the desired result follows □.
Remark: From the second part of the Theorem 4.1 we obtain that

$$\begin{aligned} \frac{1}{N}(M_N-E_N(\hat{\beta}_N)) &= \frac{1}{N}(M_N-E_N(\beta_0)+E_N(\beta_0)-E_N(\hat{\beta}_N)) \\ &\approx \frac{1}{N}(M_N-E_N(\beta_0)-\sqrt{N}\mathbf{B}'\mathbf{G}(\hat{\beta}_N-\beta_0) \end{aligned}$$

It follows that the asymptotic covariance matrix of $(M_N-E_N(\hat{\beta}_N))/N$ is given by $\mathbf{C}-\mathbf{B}'\mathbf{G}\mathbf{B}$. So another useful statistic is

$$\frac{1}{N}(M_N-E_N(\hat{\beta}_N)'[\mathbf{C}-\mathbf{B}'\mathbf{G}\mathbf{B}]^{-1}(M_N-E_N(\hat{\beta}_N))$$

where the inverse is a symmetric generalized inverse. The asymptotic distribution of this statistic is again chi-square but the number of degrees of freedom is less or equal to $kq-1$.

References

[1] A. Agresti, *Categorical data analysis*, Wiley, New York, 1990.

[2] P. K. Andersen and R. D Gill, *Cox's regression models for counting process: a large sample approach*, Annals of Statistics **10** (1982), 88–123.

[3] E. Arjas and P. Haara, *A logistic regression model for hazard: Asymptotic results*, Scandinavian Journal of Statistics **14** (1987), 1-18.

[4] P. Billingsley, *Statistical inference for markov processes*, Univ. Chicago Press, Chicago, 1961.

[5] E. G. Bonney, *Logistic regression for dependent binary observations*, Biometrics **43** (1987), 951–973.

[6] G. E. P. Box and G. M. Jenkins, *Time series analysis: Forecasting and control*, 2nd ed., Holden-Day, San Francisco, 1976.

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

[8] J. P. Diggle, K-Y. Liang, and L. S. Zeger, *Analysis of longitudinal data*, Oxford University Press, New York, 1994.

[9] L. Fahrmeir and H. Kaufmann, *Regression models for nonstationary categorical time series*, Journal of Time Series Analysis **8** (1987), 147–160.

[10] L. Fahrmeir and G. Tutz, *Multivariate statistical modelling based on generalized linear models*, Springer-Verlag, New York, 1994.

[11] S. J. Haberman, *The analysis of frequency data*, University of Chicago Press, Chicago, 1974.

[12] P. Hall and C. C. Heyde, *Martingale limit theorems and its applications*, Academic Press, New York, 1980.

[13] H. Kaufmann, *Regression models for nonstationary time series: Asymptotic estimation theory*, Annals of Statistics **15** (1987), 79–98.

[14] ______, *On existence and uniqueness of maximum likelihood estimates in quantal and ordinal response models*, Metrika **13** (1989), 291–313.

[15] B. Kedem, *Time series analysis by higher order crossings*, IEEE Press, New York, 1994.

[16] D. M. Keenan, *A time series analysis of binary data*, Journal of American Statistical Association **77** (1982), 816–821.

[17] E. L. Korn and A. S. Whittemore, *Methods for analyzing panel studies of accute health effects of air pollution*, Biometrics **35** (1979), 795–802.

[18] T. Z. Lai and C. Z. Wei, *Least squares estimation in stochastic regression models with applications to identification and control of dynamic systems*, Annals of Statistics **10** (1982), 154–166.

[19] W. K. Li, *Time series models based on generalized linear models: some further results*, Biometrics **50** (1994), 506–511.

[20] K.-Y. Liang and S. L. Zeger, *A class of logistic regression models for multivariate binary time series*, Journal of American Statistical Association **84** (1989), 447–451.

[21] P. McCullagh and J. A Nelder, *Generalized linear models*, 2nd ed., Chapman and Hall, London, 1989.

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

[23] S. Murphy and B. Li, *Projected partial likelihood and its application to longidutinal data*, Biometrika **82** (1995), 399–406.

[24] W. J. Pratt, *Concavity of the log-likelihood*, Journal of American Statistical Association **76** (1981), 103–106.

[25] D Schoenfeld, *Chi-square goodness-of-fit test for the proportional hazards regression model*, Biometrika **67** (1980), 145–153.

[26] M.J Silvapulle, *On the existence of maximum likelihood estimates for the binomial response models*, Journal of Royal Statistical Society **B43** (1981), 310–313.

[27] E. Slud, *Partial likelihood for continuous time stochastic processes*, Scandinavian Journal of Statistics **19** (1992), 97–109.

[28] E. Slud and B. Kedem, *Partial likelihood analysis of logistic regression and autoregression*, Statistica Sinica **4** (1994), 89–106.

[29] R. D. Stern and R. Coe, *A model fitting analysis of daily rainfall data*, Journal of Royal Statistical Society **A147** (1984), 1–34.

[30] R. W. M. Weddeburn, *On the existence and uniqueness of the maximum likelihood estimates*, Biometrika **63** (1976), 27–32.

[31] W. H. Wong, *Theory of partial likelihood*, Annals of Statistics **14** (1986), 88–123.