Теорія ймовірностей
та математична статистика
Вип. 1(100)/2019, с. 167–176

Teoriya Ĭmovirnosteĭ
ta Matematychna Statystyka
No. 1(100)/2019, pp. 167–176

UDC 519.233

STATISTICAL ANALYSIS OF CONDITIONALLY BINOMIAL NONLINEAR REGRESSION TIME SERIES WITH DISCRETE REGRESSORS

YU. S. KHARIN, V. A. VOLOSHKO

Abstract. The model of conditionally binomial nonlinear regression time series with discrete regressors is considered. A new frequencies-based estimator (FBE) of explicit form is constructed for this model. FBE is shown to be consistent, asymptotically normal, asymptotically effective, and to have less restrictive uniqueness assumptions w. r. t. the classical MLE. A fast recursive algorithm is constructed for FBE re-computation under model extension. Asymptotically optimal Wald test and forecasting statistic based on FBE are developed. Computer experiments on simulated data are performed for FBE.

Key words and phrases. Discrete regression time series, discrete regressors, generalized linear model, frequencies-based estimator, asymptotic efficiency, forecasting.

2010 *Mathematics Subject Classification.* 62-07, 62J12, 62F12; 62F03, 62F05, 62M20.

1. Introduction

Statistical analysis (including estimation of model parameters, hypotheses testing and forecasting) of regression time series is a classical and intensively used in practice branch of mathematical statistics [1–3]. A wide family of regression models called generalized linear models (GLM) was proposed by Nelder and Wedderburn [4] and developed by McCullagh and Nelder [5], Kedem and Fokianos [1]. A basic approach to statistical parameter estimation for GLM is a classical maximum likelihood estimation (MLE). Being asymptotically effective (attaining the Cramer–Rao bound), on practice MLE often faces well known computational troubles like numerical maximization of the likelihood (because of absence of its explicit form for the estimators), non-uniqueness (many local maxima of the likelihood) for "significantly non-canonical" link functions, high computational complexity.

Nowadays, discrete-valued time series are used in many applied fields: genetics [6], information protection [7–9], meteorology [1], psychology [6], finance [6]. In this paper we consider a GLM-based regression model of binomial count time series with discrete regressors. Using our approach proposed firstly for the binary time series [10], we construct a new consistent, asymptotically normal and asymptotically effective frequencies-based estimator (FBE) free of usual MLE drawbacks mentioned above. In particular, FBE has an explicit form, provides fast iterative model extension, and performs well for any kinds of smooth enough link functions. Distinctive feature of FBE is that it essentially uses discreteness of regressors. Using FBE we construct Wald test for testing hypotheses on the true values of model parameters and forecasting statistics.

The paper includes Introduction (Section 1), five main sections, and Conclusion (Section 7). In Section 2, mathematical model is defined. Section 3 is devoted to FBE construction and its basic properties. Section 4 contains results on statistical hypotheses testing and statistical forecasting based on FBE. In Section 5, computational features of FBE are analyzed and compared to the ones of classical MLE. Section 6 provides numerical results on simulated data.

2. Mathematical model

We consider the following regression model determined by Binomial conditional probability distribution:

$$\mathscr{L}\{y|x\} \sim \mathrm{Bi}(N, \theta_x),\ \mathsf{P}\{y = u|x\} = \binom{N}{u}\theta_x^u(1-\theta_x)^{N-u},\ y, u \in \mathbf{Y}, \tag{1}$$

where

$$\theta_x = F(\textstyle\sum_{i=1}^m \mathbf{a}_i\psi_i(x)),\ x \in \mathbf{X}. \tag{2}$$

Here $x \in \mathbf{X}$ is an observable nonrandom discrete regressor taking values from a finite set $\mathbf{X}$ with cardinality $|\mathbf{X}|$, $2 \le |\mathbf{X}| < \infty$; $y \in \mathbf{Y}$ is an observable dependent random variable taking values from the finite set $\mathbf{Y} := \{0, 1, \dots, N\}$ with cardinality $|\mathbf{Y}| = N+1$, $N \in \mathbb{N}$; the function $\theta_x : \mathbf{X} \to [0,1]$ given by (2) determines parameters of Binomial distribution (1); $F(\cdot)$ is some known cumulative distribution function (c.d.f.); $\psi_i(\cdot) : \mathbf{X} \to \mathbb{R}$, $i = 1, \dots, m$, are m known linearly independent base functions ($m \le |\mathbf{X}|$); $\mathbf{a} = (\mathbf{a}_i)_{i=1}^m \in \in \mathbb{R}^m$ is a vector of $m \in \mathbb{N}$ unknown model parameters to be estimated.

Note that the generalized regression model with K discrete regressors $x^{(1)} \in \mathbf{X}^{(1)}, \dots,$ $x^{(K)} \in \mathbf{X}^{(K)}$, where $\mathbf{X}^{(1)}, \dots, \mathbf{X}^{(K)}$ are any finite sets, can be reduced to the model (1), (2) by definition of the composed regressor $x = (x^{(1)}, \dots, x^{(K)}) \in \mathbf{X} = \mathbf{X}^{(1)} \times \dots \times \mathbf{X}^{(K)}$.

Remark 1. We will also consider a "wrapping" regression model determined by the wrapping model's parameter $\Theta = (\theta_x)_{x\in\mathbf{X}}$ according to (1). In this sense equation (2) determines the embedded model with parameter $\mathbf{a} \in \mathbb{R}^m$.

Introduce some regularity assumptions on the c.d.f. $F(\cdot)$:

A.1: $0 < F(z) < 1$ for all $z \in \mathbb{R}$;
A.2: $F(\cdot)$ and its inverse $F^{-1}(\cdot)$ are twice continuously differentiable;
A.3: derivative $F'(z)$ never equals 0 or ∞.

Further, as in [1, 5], we call $G := F^{-1} : (0,1) \to \mathbb{R}$ a link function.
Introduce some notations:

- $\mathbf{F} : (z_i) \mapsto (F(z_i))$ is the elementwise transform of some vector $z = (z_i)$ by the c.d.f. $F(\cdot)$; transform $\mathbf{G}$ is the same for the link function G;
- $\Psi(x) = (\psi_1(x), \dots, \psi_m(x))'$ is a vector-column of base functions at point $x \in \mathbf{X}$;
- $\Psi \in \mathbb{R}^{m\times|\mathbf{X}|}$ is $(m \times |\mathbf{X}|)$-matrix of columns $\Psi(x)$, $x \in \mathbf{X}$; informally, Ψ is also treated as the set of base functions $\{\psi_i\}_{i=1}^m$ (rows of matrix Ψ);
- the parameter of the model (1), (2) and the parameter of its wrapping model (1) are related as follows:

$$\Theta(\mathbf{a}) = (F(\mathbf{a}'\Psi(x)))_{x\in\mathbf{X}} = \mathbf{F}(\mathbf{a}'\Psi); \tag{3}$$

- $\mathbb{R}^{\mathbf{X}}$ is a $|\mathbf{X}|$-dimensional space of functions $f : \mathbf{X} \to \mathbb{R}$;
- $\Pi_\Psi \subset \mathbb{R}^{\mathbf{X}}$ is a linear span of the basis $\{\psi_i\}$:

$$\Pi_\Psi = \mathrm{span}(\psi_1, \dots, \psi_m); \tag{4}$$

- for two sets of functions $\phi^j = \{\phi_i^j \in \mathbb{R}^{\mathbf{X}}\}_{i=1}^{m_j}$, $j = 1, 2$, and for some matrix $h \in \mathbb{R}^{\mathbf{X}\times\mathbf{X}}$ denote Gram $(m_1 \times m_2)$-matrix of dot products:

$$\langle \phi^1, \phi^2 \rangle_h := ((\phi_i^1)' h \phi_k^2) \in \mathbb{R}^{m_1\times m_2},\ i = 1, \dots, m_1,\ k = 1, \dots, m_2; \tag{5}$$

  for $\phi = \phi^1 = \phi^2$ briefly $\langle \phi \rangle_h := \langle \phi, \phi \rangle_h$, for one-function set $\langle \phi \rangle_h =: \|\phi\|_h^2 \in \mathbb{R}$ (squared L_2-norm), for the identity matrix $h = \mathbf{Id}$: $\langle\cdot\rangle_{\mathbf{Id}} =: \langle\cdot\rangle$;
- $\mathbb{1}\{A\}$ is the indicator of event A;
- design of $T \in \mathbb{N}$ experiments (DOE) in $\mathbf{X}$:

$$\pi^T = (\pi_x^T)_{x\in\mathbf{X}},\ \pi_x^T = \tfrac{1}{T}\textstyle\sum_{t=1}^T \mathbb{1}\{x_t = x\}, \tag{6}$$

where π_x^T means a portion of experiments with the fixed value $x \in \mathbf{X}$ of the regressor.

We observe two discrete-time processes of the length $T \in \mathbb{N}$: the regressor nonrandom process $x_t \in \mathbf{X}$, and the response random process $y_t \in \mathbf{Y}$. All the pairs (y_t, x_t), $t = 1, \dots, T$, are mutually independent and follow model (1), (2). Further let us agree to denote $\theta_t := \theta_{x_t}$ for brevity.

3. Statistical estimation of parameters

Let us consider the problem of statistical estimation of parameter $\mathbf{a} \in \mathbb{R}^m$ by T pairs of observed data $Y_1^T = (y_1, \dots, y_T) \in \mathbf{Y}^T$, $X_1^T = (x_1, \dots, x_T) \in \mathbf{X}^T$. Maximum likelihood estimators (MLE) for the model parameter $\mathbf{a} = (\mathbf{a}_i)_{i=1}^m$ and for parameter $\Theta = (\theta_x)_{x \in \mathbf{X}}$ of the wrapping model are defined as follows:

$$\hat{\Theta} := \arg\max_{\Theta} L(\Theta), \tag{7}$$

$$\hat{\mathbf{a}}^* := \arg\max_{\mathbf{a}} L(\Theta(\mathbf{a})), \tag{8}$$

where $\Theta(\mathbf{a})$ is defined by (3), $L(\Theta)$ is the loglikelihood function for the model (1):

$$L(\Theta) = \textstyle\sum_{t=1}^T \lambda(y_t, \theta_t),\ \lambda(y, \theta) := \ln \binom{N}{y} + y \ln \theta + (N - y) \ln(1 - \theta). \tag{9}$$

Lemma 1. *MLE* (7) *is the vector* $\hat{\Theta} = (\hat{\theta}_x)_{x \in \mathbf{X}}$ *determined by expressions:*

$$\hat{\theta}_x = \arg\max_{0 \le \theta \le 1} L_x(\theta) = \tfrac{1}{N} \cdot \tfrac{1}{|\mathrm{T}_x|} \textstyle\sum_{t \in \mathrm{T}_x} y_t,\ x \in \mathbf{X}, \tag{10}$$

$$L_x(\theta) := \textstyle\sum_{t \in \mathrm{T}_x} \lambda(y_t, \theta),\ \mathrm{T}_x := \{t = 1, \dots, T : x_t = x\}. \tag{11}$$

Proof. Loglikelihood function (9) satisfies the separable presentation:

$$L(\Theta) = \textstyle\sum_{x \in \mathbf{X}} L_x(\theta_x),$$

and can be minimized separately by each θ_x, $x \in \mathbf{X}$. The MLE $\hat{\theta}_x$ for the parameter θ_x of a Binomial distribution $\mathrm{Bi}(N, \theta_x)$ built by the subsample $\{y_t : t \in \mathrm{T}_x\}$ has the form (10). □

Theorem 1. *If the DOE* (6) *converges under* $T \to \infty$ *to some nonsingular distribution* $\pi = (\pi_x)$*:* $\pi_x^T \to \pi_x > 0$*,* $x \in \mathbf{X}$*, then there exist sequences* $\hat{\Theta} = \hat{\Theta}(T) \in (0,1)^{\mathbf{X}}$ *and* $\hat{\mathbf{a}}^* = \hat{\mathbf{a}}^*(T) \in \mathbb{R}^m$*, such that:*

1) *with probability going to* 1*, as* $T \to \infty$*,* $\hat{\Theta}$ *and* $\hat{\mathbf{a}}^*$ *are the local maximums of loglikelihood functions* $L(\Theta)$ *and* $L(\Theta(\mathbf{a}))$ *(asymptotic existence);*
2) $\hat{\Theta} \xrightarrow[T\to\infty]{\mathsf{P}} \Theta$*,* $\hat{\mathbf{a}}^* \xrightarrow[T\to\infty]{\mathsf{P}} \mathbf{a}$ *(consistency);*
3) $\sqrt{T}(\hat{\Theta} - \Theta) \xrightarrow[T\to\infty]{D} \mathcal{N}_{\mathbf{X}}(0, I^{-1}(\Theta))$*,* $\sqrt{T}(\hat{\mathbf{a}}^* - \mathbf{a}) \xrightarrow[T\to\infty]{D} \mathcal{N}_m(0, I^{-1}(\mathbf{a}))$ *(asymptotic normality).*

Here $I(\Theta) \in \mathbb{R}^{|\mathbf{X}| \times |\mathbf{X}|}$ *and* $I(\mathbf{a}) \in \mathbb{R}^{m \times m}$ *are Fisher information matrices for the parameters* Θ *and* $\mathbf{a}$ *respectively (using notations* (5)*):*

$$I(\Theta) = \mathrm{diag}(\pi_x I_{\mathrm{Bi}}(\theta_x))_{x \in \mathbf{X}}, \tag{12}$$

$$I(\mathbf{a}) = \langle \Psi \rangle_J,\ J = \mathrm{diag}\big(\pi_x I_{\mathrm{Bi}}(\theta_x)(G'(\theta_x))^{-2}\big)_{x \in \mathbf{X}} \in \mathbb{R}^{|\mathbf{X}| \times |\mathbf{X}|}, \tag{13}$$

$I_{\mathrm{Bi}}(\theta) = N/\theta(1-\theta)$ *is the Fisher information w. r. t. parameter* θ *of Binomial distribution* $\mathrm{Bi}(N, \theta)$.

Proof. The results follow from [11] under regularity conditions **A.1**–**A.3**. □

Remark 2. Diagonal matrix J defined by (13) is nonsingular and finite. It follows from nonsingularity of the limit DOE π, condition **A.3** and from nonsingularity of Fisher information $I_{\mathrm{Bi}}(\theta) > 0$ for $\theta \in (0, 1)$.

Let us now build a new plug-in statistical estimator for the parameter $\mathbf{a}$ using the MLE (7) for the wrapping model: $\hat{\mathbf{a}} = \hat{\mathbf{a}}(\hat{\Theta})$.

Lemma 2. *For the wrapping model* (1) *the following relations hold:*

$$\mathbf{a}'\Psi(x) = G(\theta_x),\ x \in \mathbf{X}. \tag{14}$$

Proof. Acting by the link function $G(\cdot)$ on the both parts of (2) gives (14). □

Using Lemma 2 and Theorem 1 let us follow [10] and build a system of $|\mathbf{X}|$ linear in $\mathbf{a} = (\mathbf{a}_i)$ equations based on relations (14) and consistent estimators (7):

$$\mathbf{a}'\Psi(x) = G(\hat{\theta}_x),\ x \in \mathbf{X}. \tag{15}$$

On practice, equations (15) are only available for some subset of the observed values $x \in \mathbf{X}$. This question is discussed in Section 5. Assuming that $m \le |\mathbf{X}|$, for the system (15) we use the minimal squares method:

$$\begin{aligned} 0 \le W(\mathbf{a}) &:= \textstyle\sum_{x,x' \in \mathbf{X}} (\mathbf{a}'\Psi(x) - G(\hat{\theta}_x))(\mathbf{a}'\Psi(x') - G(\hat{\theta}_{x'}))H_{xx'} = \\ &= \|\mathbf{a}'\Psi - \mathbf{G}(\hat{\Theta})\|_H^2 \to \min_{\mathbf{a} \in \mathbb{R}^m}, \end{aligned} \tag{16}$$

where $H = (H_{uv}) = H' \ge 0$ is a symmetric nonnegative definite weight $(|\mathbf{X}| \times |\mathbf{X}|)$-matrix such that bilinear form W is strictly positive definite on $\mathbb{R}^m$.

Theorem 2. *If* $\langle \Psi \rangle_H > 0$*, then the minimization problem* (16) *has the unique solution:*

$$\hat{\mathbf{a}} := \arg\min_{\mathbf{a} \in \mathbb{R}^m} W(\mathbf{a}) = \langle \Psi \rangle_H^{-1} \Big\langle \Psi, \mathbf{G}(\hat{\Theta}) \Big\rangle_H, \tag{17}$$

$$W(\hat{\mathbf{a}}) = \min_{\mathbf{a} \in \mathbb{R}^m} W(\mathbf{a}) = \mathbf{G}(\hat{\Theta})' \Big(H - H\Psi' \langle \Psi \rangle_H^{-1} \Psi H \Big) \mathbf{G}(\hat{\Theta}). \tag{18}$$

Proof. The proof is similar to the proof of Theorem 1 from [10] and follows from the properties of positive definite quadratic forms [12]. □

The statistic (17) is based on frequencies (10), therefore, similarly to [10] we call statistic (17) the FB-estimator (FBE) of parameter $\mathbf{a}$.

Theorem 3. *Under* $\langle \Psi \rangle_H > 0$ *and conditions* **A.1**–**A.3** *as* $T \to \infty$ *the FBE* (17) *is consistent and asymptotically normal statistical estimator:*

$$\begin{gathered} \hat{\mathbf{a}} \xrightarrow[T \to \infty]{\mathsf{P}} \mathbf{a},\ \sqrt{T}(\hat{\mathbf{a}} - \mathbf{a}) \xrightarrow[T \to \infty]{D} \mathcal{N}_m(0, \Sigma_H), \\ \Sigma_H = \langle \Psi \rangle_H^{-1} \langle \Psi \rangle_{HJ^{-1}H} \langle \Psi \rangle_H^{-1}, \end{gathered} \tag{19}$$

where the matrix J *is defined by* (13).

Proof. The proof is similar to the proof of Theorem 2 and Theorem 3 in [10] and is based on Theorem 1 of this paper and on theorem on smooth functional transformations of asymptotically normal random sequences [13]. □

Corollary 1. *For the optimal weight matrix* $H = J$ *defined by* (13) *the FBE is asymptotically efficient: it attains the Cramer–Rao bound* $\Sigma_H = I^{-1}(\mathbf{a})$ *as* $T \to \infty$.

Proof. Substitution of J into (19) instead of H gives $\Sigma_J = \langle \Psi \rangle_J^{-1}$, which equals $I^{-1}(\mathbf{a})$ according to (13). □

4. Statistical hypotheses testing and statistical forecasting

Due to asymptotic normality of the built FBE (17) proved in Theorem 3, this constructed estimator can be used for statistical hypotheses testing: $H_0 : \mathbf{a} = \mathbf{a}^0$, where $\mathbf{a}^0$ is some fixed hypothetical value, against the alternative $H_1 = \overline{H}_0$ based on the Wald criterion [14]:

$$\text{accepted } \begin{cases} H_0, & \text{if } (\hat{\mathbf{a}} - \mathbf{a}^0)' \Sigma_H^{-1} (\hat{\mathbf{a}} - \mathbf{a}^0) \le F^{-1}_{\chi^2_m}(1-\varepsilon), \\ H_1, & \text{otherwise,} \end{cases} \tag{20}$$

where $F^{-1}_{\chi^2_m}(\cdot)$ means the chi-square quantile function with m degrees of freedom, $0 < \varepsilon < 1$ is some significance level.

Corollary 2. *The Wald criterion* (20) *has the asymptotic significance level* ε *as* $T \to \infty$.

Consider now the problem of statistical forecasting of the future state $y_{T+\tau} \in \mathbf{Y}$ in $\tau \ge 1$ steps ahead based on the data $Y_1^T, (X_1^T, x_{T+\tau})$.

Corollary 3. *If* $\mathbf{a}$ *is known, the minimal mean square error of forecasting is given by the forecasting statistic*

$$\hat{y}_{T+\tau} = [NF(\textstyle\sum_{i=1}^m \mathbf{a}_i \psi_i(x_{T+\tau}))],$$

where $[z]$ *means the closest integer to the number* z.

Proof. Let $y = y_{T+\tau}$, $x = x_{T+\tau}$. As $\hat{y} = \hat{y}(x)$ is nonrandom, the mean square error $\mathsf{E}\{(\hat{y} - y)^2\} = \mathsf{D}\{y\} + |\hat{y} - \mathsf{E}\{y\}|^2 = N\theta_x(1-\theta_x) + |\hat{y} - N\theta_x|^2$, which is minimal at the closest to $N\theta_x = NF(\sum_{i=1}^m \mathbf{a}_i \psi_i(x))$ integer $\hat{y}$. □

Using Theorem 3 we propose the asymptotically optimal forecasting statistic:

$$\tilde{y}_{T+\tau} = [NF(\textstyle\sum_{i=1}^m \hat{\mathbf{a}}_i \psi_i(x_{T+\tau}))], \tag{21}$$

where $\hat{\mathbf{a}} = (\hat{\mathbf{a}}_i)$ is determined by explicit form (17).

5. Computational advantages of FBE

FBE (17) has some essential computational advantages w. r. t. MLE (8). First of all, FBE has an explicit form (17), while MLE is computed by iterative numerical procedure with known related problems. As we will see, the other advantages are: less restrictive sufficient uniqueness conditions; ability of iterative extension of the model.

Consider the following subsets of the regressor discrete space $\mathbf{X}$:

$$\mathbf{M} = \mathbf{M}(X_1^T) := \textstyle\bigcup_{t=1}^T \{x_t\},\ \mathbf{M}_+ = \mathbf{M}_+(X_1^T) := \bigcap_{y \in \{0,N\}} \bigcup_{t=1}^T \{x_t | y_t \ne y\}. \tag{22}$$

The set $\mathbf{M}$ contains all the values of regressors $x_t \in \mathbf{X}$ met in the observed regressor process X_1^T. Also $\mathbf{M}$ is representable as a support of DOE (6) and the set of $x \in \mathbf{X}$ for which $\mathrm{T}_x \ne \varnothing$ in (11). The set $\mathbf{M}_+ \subset \mathbf{M}$, in its turn, contains those values $x \in \mathbf{X}$ for which there exist $t, \tau \in \{1, \dots, T\}$: $x_t = x_\tau = x$, $y_t < N$, $y_\tau > 0$. The estimates (10) are only defined for $x \in \mathbf{M}$. The equations (15), according to (10), are only defined for $x \in \mathbf{M}_+$. Indeed, $G(\hat{\theta}_x)$ is finite iff $\hat{\theta}_x \notin \{0, 1\}$ which is equivalent to $x \in \mathbf{M}_+$. Therefore, similarly to the case of binary autoregressive model [10], we consider the sparse FBE that use some subset of equations (15) instead of their full set of the size $|\mathbf{X}|$. Namely, we use the following restrictions on the weight matrix H:

$$H \in \mathcal{H}_M,\ \mathcal{H}_M := \{H = H' \ge 0 : H_{x,x'} \equiv 0, \forall (x, x') \notin M^2\},\ M \subset \mathbf{X}. \tag{23}$$

If the weight matrix H in (17) satisfies (23), then we call corresponding FBE (17) an M-sparse FBE. As it was mentioned above, the equations (15) are only possible for

$x \in \mathbf{M}_+$, whence only $\mathbf{M}_+$-sparse FBEs can be built by the observations (Y_1^T, X_1^T). For each $M \subset \mathbf{X}$ we denote by $[M]$ the corresponding projector matrix:

$$[M] := (\mathbb{1}\{x = x' \in M\})_{x,x' \in \mathbf{X}} \in \mathcal{H}_M. \tag{24}$$

The sufficient uniqueness conditions for MLE (8) and FBE (17) are as follows:

$\mathbf{U}^*$: $\ln F(z)$, $\ln(1 - F(z))$ are strictly concave on $z \in \mathbb{R}$, and $\langle \Psi \rangle_{[\mathbf{M}_+]}$ is positive definite;

$\mathbf{U}$: $\langle \Psi \rangle_H$ is positive definite.

Theorem 4. *Under condition* $\mathbf{U}^*$ *(*$\mathbf{U}$*) the MLE* (8) *(the FBE* (17)*) is unique.*

Proof. The uniqueness of the FBE (17) under condition $\mathbf{U}$ follows from Theorem 2. Conditions $\mathbf{U}^*$ are given in [15, 16] as a sufficient ones for MLE uniqueness for binary and binomial GLM models. □

Remark 3. C.d.f. $F(\cdot)$ satisfying the condition $\mathbf{U}^*$ is called weakly log-concave [17]. In particular, logistic and normal c.d.f.'s satisfy it, while Cauchy c.d.f. and the others heavy-tailed ones do not. The conditions $\mathbf{U}$ and $\mathbf{U}^*$ related to the basis Ψ are close to each other and are equivalent for the weight matrices $H \in \mathcal{H}_{\mathbf{M}_+}$ of a general form, when H is strictly positive definite restricted to $\mathbf{M}_+$.

According to the formula (17), computational complexity of M-sparse FBE is linear in the cardinality $|M|$, i. e. in the number of equations (15) used. So the small subsets $M \subset \mathbf{X}$ lead to reduced computational complexity of M-sparse FBEs (17), (23). On the other hand, M-sparse FBEs use incomplete data, which leads to a decrease in efficiency (accuracy) of estimation. Therefore some balance is needed between efficiency and computational complexity. Further we find the best attainable efficiency of the M-sparse FBE subject to a fixed subset M.

Let us use a measure of efficiency of FBE (17) which we call risk:

$$\mathbf{R}(\hat{\mathbf{a}}) := \operatorname{tr}(\Sigma_H \cdot I(\mathbf{a})) \geq m, \tag{25}$$

where $I(\mathbf{a})$ and Σ_H are respectively the Fisher information matrix (13) and the asymptotic covariance matrix (19). The minimum risk value $\mathbf{R} = m$ corresponds to the Cramer–Rao bound that is reached at $H = J$ due to Corollary 1. Introduce some notation:

- for two m-dimensional subspaces Π_1, Π_2 of a Hilbert space denote

$$\cos^{-2}(\Pi_1, \Pi_2) := \left\| \left\langle \varphi^1, \varphi^2 \right\rangle^{-1} \right\|_{\mathfrak{F}}^2 = \textstyle\sum_{i=1}^m 1/\cos^2 \alpha_i, \tag{26}$$

  where $\|D\|_{\mathfrak{F}}^2 = \sum_{i,j} D_{ij}^2$ is a squared Frobenius norm of a matrix, $\varphi^i = \{\varphi_j^i\}_{j=1}^m \subset \subset \Pi_i$, $i = 1, 2$, are two arbitrary orthonormal bases, $\{\alpha_i\}_{i=1}^m$ are the Jordan's principal angles [18] between Π_1 and Π_2;
- under the notation (4) and (13), denote the following m-dimensional subspaces of $\mathbb{R}^{\mathbf{X}}$:

$$\Pi_1 = J^{1/2}\Pi_\Psi \subset \mathbb{R}^{\mathbf{X}},\ \Pi_2 = J^{-1/2}H\Pi_\Psi \subset \mathbb{R}^{\mathbf{X}}. \tag{27}$$

Theorem 5. *Risk* (25) *of the FBE* (17) *has the form*

$$\mathbf{R}(\hat{\mathbf{a}}) = \cos^{-2}(\Pi_1, \Pi_2). \tag{28}$$

Proof. The proof of this theorem is similar to the proof of Theorem 4 in [10] for the space $\mathbb{R}^{\mathbf{X}}$ instead of $\mathbb{R}^{\{0,1\}^s}$. □

Theorem 6. *Let* $|\langle \Psi_M \rangle| \neq 0$ *for* $M \subset \mathbf{X}$. *Then the minimal risk* (25) *among the* M*-sparse FBEs* (17) *is attained at* $H_M = J_M$ *and has the value*

$$\mathbf{R}_M^*(\mathbf{a}) := \min_{H \in \mathcal{H}_M} \mathbf{R}(\hat{\mathbf{a}}) = \operatorname{tr}\Big(\langle \Psi_M \rangle_{J_M}^{-1} \langle \Psi \rangle_J \Big). \tag{29}$$

Here J is defined by (13)*, J_M is an $(M \times M)$-submatrix of J, Ψ_M is a submatrix of base functions $\{\psi_i\}$ restricted to M.*

Proof. The proof of this theorem is similar to the proof of Theorem 5 in [10] and is derived from (28) by the following scheme. From $H \in \mathcal{H}_M$ it follows that $\Pi_2 \subset \Pi_M$, where $\Pi_M \subset \mathbb{R}^{\mathbf{X}}$ is a subspace of functions $\xi : \mathbf{X} \to \mathbb{R}$ having support $\operatorname{supp}\xi \subset M$. The subspaces Π_1, Π_2 and $\Pi_* = [M]\Pi_1$ (see (24)) are m-dimensional, which follows from $\langle\Psi\rangle_H > 0$ and $|J| \neq 0$ (see Remark 2). The value (29) equals $\cos^{-2}(\Pi_1, \Pi_*)$, hence we get from (28) and Lemma 5 in [10] that

$$\mathbf{R}^*_M(\mathbf{a}) = \cos^{-2}(\Pi_1, \Pi_*) \leq \cos^{-2}(\Pi_1, \Pi_2) = \mathbf{R}(\hat{\mathbf{a}}).$$

This lower bound of the risk $\mathbf{R}(\hat{\mathbf{a}})$ is attained at $H_M = J_M$ where $\Pi_2 = \Pi_*$. □

One more advantage of the FBE (17) w. r. t. the MLE (8) is the fast iterative model extension algorithm. By model extension, we mean adding a new base function $\psi_{m+1}(\cdot)$ to the base $\Psi = \{\psi_i\}_{i=1}^m$.

The fast "Gram–Schmidt-based" algorithm of re-estimation of a parameter $\mathbf{a}$ is as follows. We have to orthonormalize the base $\{\psi_i\}$ w. r. t. the dot product $\langle\cdot,\cdot\rangle_H$. The weight matrix H is assumed to be fixed during the iterative process. On the m-th iteration the algorithm stores $\mathcal{O}(m^2)$ auxiliary coefficients: the inverse Gram matrix $\mathfrak{G}^m = \langle\Psi\rangle_H^{-1} \in \mathbb{R}^{m\times m}$; Cholesky decomposition $\mathfrak{G}^m = R'R$, where $R = (\mathfrak{R}_{i,j})_{i,j=1}^m$ is a lower triangular matrix ($\mathfrak{R}_{i,j} \equiv 0$ at $j > i$). The i-th function of the resulting orthonormal base is $\tilde{\psi}_i = \sum_{j=1}^i \mathfrak{R}_{i,j}\psi_j$;

$$\epsilon_i(\xi) := \langle\tilde{\psi}_i, \xi\rangle_H = \textstyle\sum_{j=1}^i \mathfrak{R}_{i,j}\langle\psi_j, \xi\rangle_H,\ E_i(\xi) := \sum_{j=1}^i \epsilon_j^2(\xi),$$

are respectively the i-th coefficient of $\xi : \mathbf{X} \to \mathbb{R}$ in the base $\{\tilde{\psi}_i\}$, and a squared norm of the orthogonal projection of ξ onto $\operatorname{span}(\psi_j)_{j=1}^i$ w. r. t. $\langle\cdot,\cdot\rangle_H$. Under the notation (5) the FBE $\hat{\mathbf{a}} = \hat{\mathbf{a}}^m \in \mathbb{R}^m$ is computed by the recurrence

$$\hat{\mathbf{a}}_i^m = \hat{\mathbf{a}}_i^{m-1} + \epsilon_m(\mathbf{G}(\hat{\Theta}))\cdot\mathfrak{R}_{m,i},\ i = 1,\ldots,m, \tag{30}$$

where $\hat{\mathbf{a}}_i^m \equiv 0$ at $i > m$. The approximation error (18) is computed as follows:

$$W(\hat{\mathbf{a}}^m) = \|\mathbf{G}(\hat{\Theta})\|_H^2 - E_m(\mathbf{G}(\hat{\Theta})). \tag{31}$$

The recurrences for the inverse Gram matrix and Cholesky decomposition are:

$$\mathfrak{R}_{m,m} = \big(\|\psi_m\|_H^2 - E_{m-1}(\psi_m)\big)^{-1/2}, \tag{32}$$

$$\mathfrak{R}_{m,i} = -\mathfrak{R}_{m,m}\textstyle\sum_{j=1}^{m-1}\langle\psi_m,\psi_j\rangle_H\mathfrak{G}_{i,j}^{m-1},\ i = 1,\ldots,m-1, \tag{33}$$

$$\mathfrak{G}_{i,j}^m = \mathfrak{G}_{i,j}^{m-1} + \mathfrak{R}_{m,i}\mathfrak{R}_{m,j},\ i,j = 1,\ldots,m. \tag{34}$$

Here $\mathfrak{G}_{i,j}^{m-1} \equiv 0$ at $\max\{i,j\} = m$. Initial values are $\mathbf{a}^0 = 0$, $\mathfrak{G}^0 = 0$, $\mathfrak{R}_{1,1} = 1/\|\psi_1\|_H$.

Using the described algorithm (30)–(34) we can add new base functions ψ_{m+1} adaptively, i. e. minimizing the error (31) among the candidates for ψ_{m+1}. We propose the following empirical choice for subset $M \subset \mathbf{X}$ and for weight matrix $H \in \mathcal{H}_M$:

$$M = M_\alpha := \min\{M \subset \mathbf{M}_+ : \pi^T_M \geq \alpha\},\ H = \hat{J} = \operatorname{diag}\left(\pi_x^T\frac{I_{\mathrm{Bi}}(\hat{\theta}_x)}{(G'(\hat{\theta}_x))^2}\right)_{x\in M}, \tag{35}$$

where the DOE π^T is defined by (6), $0 \leq \alpha \leq \alpha_* = \pi^T_{\mathbf{M}_+} \leq 1$. The choice (35) for H is based on Theorem 6. The choice (35) for M gathers the most "data-supported" values $x \in \mathbf{X}$ by some threshold α of data portion used.

6. Results of computer experiments

For the experiments we use time series simulated for the model (1), (2) with the following parameters: $N = 3$, $\mathbf{Y} = \{0, 1, 2, 3\}$; $\mathbf{X} = \{0, 1, \ldots, 31\}$; $F(z) = \frac{1}{2} + \pi^{-1} \arctan(z)$ is the Cauchy c.d.f; $m = 5$, $\{\psi_i(x)\}_{i=1}^5$ are the first five polynomials orthonormal on $\mathbf{X}$ w. r. t. the uniform probability measure, $\deg \psi_i = i - 1$, $i = 1, \ldots, 5$; $\mathbf{a}^0 = (1, -1, 2, -2, 1)$ is true model parameter.

The plots in Figure 1 illustrate dependence of the risk (25) on time series length T (both risk and time series length are in logarithmic scale on the plots). The risk (25) was evaluated by Monte Carlo method with $K = 1000$ replications:

$$\hat{\mathbf{R}}(\hat{\mathbf{a}}) = \frac{T}{K} \sum_{k=1}^{K} \|\hat{\mathbf{a}}^{(k)} - \mathbf{a}^0\|^2_{I(\mathbf{a})},$$

where $\hat{\mathbf{a}}^{(k)}$ is the value of statistic (17) in the k-th replication. The both plots in Figure 1 have the lower level $\log_2(\mathbf{R}(\hat{\mathbf{a}})) \approx 2.32$ corresponding to the Cramer–Rao bound $\mathbf{R}(\hat{\mathbf{a}}) = 5$.

The left plot in Figure 1 is constructed for the case of uniform DOE (6): $\pi_x^T = 1/32$, $x \in \mathbf{X}$. As it is seen from this plot, the use of the estimated optimal weight matrix $H = \hat{J}$ in (17) gives significant gain (w. r. t. the case $H = \mathbf{Id}$) in the estimation accuracy and asymptotical convergence of the FBE (17) to the Cramer–Rao bound.

The right plot in Figure 1 is constructed for $H = \hat{J}$ and for the case of nonuniform DOE (6) with the portions π_x^T ranging between 1/24 and 1/48. The right plot shows that the loss of estimation accuracy caused by the use of sparse FBE (α is a π^T-measure of sparse subset of equations (15) used) is more significant for large time series length T. For "small data" the corresponding loss of estimation accuracy is less dramatic.

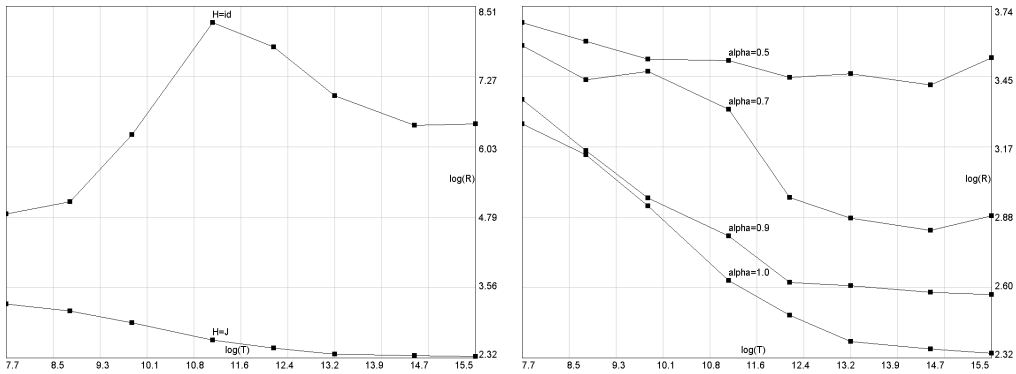

Figure 1. Dependence of the risk (25) on time series length T (both in logarithmic scale); left plot: different weight matrices H (unit matrix $H = \mathbf{Id}$ versus the estimated optimal matrix $H = \hat{J}$); right plot: different parameters α of the sparse subset of regressors (35), $H = \hat{J}$

7. Conclusion

The following new results are obtained in the paper for conditionally binomial nonlinear regression time series with discrete regressors:

1) a new frequencies-based estimator (FBE) of explicit form is constructed;
2) FBE is shown to be consistent, asymptotically normal and asymptotically effective;
3) FBE is shown to have less restrictive uniqueness assumptions w. r. t. the classical MLE;

4) a fast recursive algorithm is constructed for FBE re-computation under model extension;
5) asymptotically optimal Wald test and forecasting statistic based on FBE are developed;
6) FBE is successfully tested in computer experiments on simulated data.

The authors plan to use the developed results in robust statistical analysis [19–22].

Acknowledgement

The authors gratefully acknowledge the valuable recommendations of the Referee and of the Associated Editor during the paper preparation.

References

1. B. Kedem, K. Fokianos, *Regression Models for Time Series Analysis*, Wiley, Hoboken, 2002.
2. Yu. S. Kharin, *Robustness in Statistical Forecasting*, Springer, Cham, Heidelberg, New York, Dordrecht, London, 2013.
3. O. O. Dashkov, A. G. Kukush, *Consistency of the orthogonal regression estimator in an implicit linear model with errors in variables*, Theory Probab. Math. Statist., **97** (2018), 45–55.
4. J. Nelder, R. Wedderburn, *Generalized linear models*, J. Royal Statistical Society. Series A, **35** (1972), no. 3, 370–384.
5. P. McCullagh, J. A. Nelder, *Generalized Linear Models*, Chapman and Hall, London, 1989.
6. C. H. Weiss, *An Introduction to Discrete-Valued Time Series*, John Wiley and Sons Ltd, 2018.
7. Yu. S. Kharin, E. V. Vecherko, *Statistical estimation of parameters for binary Markov chain models with embeddings*, Discrete Mathematics and Applications, **23** (2013), no. 2, 153–169.
8. Yu. S. Kharin, E. V. Vecherko, *Detection of embeddings in binary Markov chains*, Discrete Mathematics and Applications, **26** (2016), no. 1, 13–29.
9. V. A. Voloshko, *Steganographic capacity for one-dimensional Markov cover*, Discrete Mathematics and Applications, **27** (2017), no. 4, 247–268.
10. Yu. S. Kharin, V. A. Voloshko, E. A. Medved, *Statistical estimation of parameters for binary conditionally nonlinear autoregressive time series*, Mathematical Methods of Statistics, **27** (2018), no. 2, 103–118.
11. L. Fahrmeir, H. Kaufmann, *Consistency and Asymptotic Normality of the Maximum Likelihood Estimator in Generalized Linear Models*, The Annals of Statistics, **13** (1985), no. 1, 342–368.
12. B. Noble, J. W. Daniel, *Applied Linear Algebra*, Prentice-hall, Englewood Cliffs, 1988.
13. A. N. Shiryaev, *Probability*, Springer, New York, 1995.
14. A. Wald, *Tests of statistical hypotheses concerning several parameters when the number of observations is large*, Trans. Amer. Math. Soc., **54** (1943), no. 3, 426–482.
15. S. J. Haberman, *Maximum likelihood estimates in exponential response models*, The Annals of Statistics, **5** (1977), no. 5, 815–841.
16. R. W. M. Wedderburn, *On the Existence and Uniqueness of the Maximum Likelihood Estimates for Certain Generalized Linear Models*, Biometrika, **63** (1976), no. 1, 27–32.
17. M. Bagnoli, T. Bergstrom, *Log-Concave Probability and Its Applications*, University of Michigan, 1989.
18. C. Jordan, *Essai sur la géométrie à n dimensions*, Bulletin de la Société Mathématique de France, **3** (1875), 103–174.
19. Yu. Kharin, *Robustness of clustering under outliers*, Lecture Notes in Computer Science, **1280** (1997), 501–511.
20. Yu. Kharin, E. Zhuk, *Filtering of multivariate samples containing "outliers" for clustering*, Pattern Recognition Letters, **19** (1998), 1077–1085.
21. Yu. Kharin, *Robustness of the mean square risk in forecasting of regression time series*, Communications in Statistics – Theory and Methods, **40** (2011), no. 16, 2893–2906.
22. A. Kharin, *Performance and robustness evaluation in sequential hypotheses testing*, Communications in Statistics – Theory and Methods, **45** (2016), no. 6, 1693–1709.

Research Institute for Applied Problems of Mathematics and Informatics, Belarusian State University, Minsk, Belarus
E-mail address: kharin@bsu.by

Research Institute for Applied Problems of Mathematics and Informatics, Belarusian State University, Minsk, Belarus
E-mail address: valeravoloshko@yandex.ru

Received 28.02.2019

СТАТИСТИЧНИЙ АНАЛІЗ УМОВНО БІНОМІАЛЬНИХ НЕЛІНІЙНИХ РЕГРЕСІЙНИХ ЧАСОВИХ РЯДІВ ІЗ ДИСКРЕТНИМИ РЕГРЕСОРАМИ

Ю. С. ХАРІН, В. А. ВОЛОШКО

Анотація. Розглянуто модель умовно біноміального нелінійного регресійного часового ряду з дискретними регресорами. Для цієї моделі будується нова статистична FB-оцінка параметрів на основі частот, яка має явний вигляд. Для FB-оцінки доводяться конзистентність, асимптотична нормальність, асимптотична ефективність, а також достатні умови єдиності, менш жорсткі порівняно з такими для класичної оцінки максимальної вірогідності. Будується швидкий рекурсивний алгоритм обчислення FB-оцінки при розширенні моделі. На основі FB-оцінки розробляються асимптотично оптимальні критерій Вальда і прогнозуюча статистика. Наведено результати комп'ютерних експериментів на модельних даних.

СТАТИСТИЧЕСКИЙ АНАЛИЗ УСЛОВНО БИНОМИАЛЬНЫХ НЕЛИНЕЙНЫХ РЕГРЕССИОННЫХ ВРЕМЕННЫХ РЯДОВ С ДИСКРЕТНЫМИ РЕГРЕССОРАМИ

Ю. С. ХАРИН, В. А. ВОЛОШКО

Аннотация. Рассматривается модель условно биномиального нелинейного регрессионного временного ряда с дискретными регрессорами. Для этой модели строится новая, имеющая явный вид, статистическая FB-оценка параметров на основе частот. Для FB-оценки доказываются состоятельность, асимптотическая нормальность, асимптотическая эффективность, а также достаточные условия единственности, менее жесткие в сравнении с таковыми для классической оценки максимального правдоподобия. Строится быстрый рекурсивный алгоритм вычисления FB-оценки при расширении модели. На основе FB-оценки разрабатываются асимптотически оптимальные критерий Вальда и прогнозирующая статистика. Приводятся результаты компьютерных экспериментов на модельных данных.