УДК 519.2

И.А. Бодягин, Ю.С. Харин

Прогнозирование авторегрессионных временных рядов при наличии цензурирования

Рассматривается задача статистического прогнозирования авторегрессионных временных рядов при наличии интервального цензурирования. Построена оптимальная прогнозирующая статистика, для нее вычислен условный среднеквадратический риск. Для авторегрессии первого порядка проведено сравнение оптимальной прогнозирующей статистики с прогнозирующими статистиками, часто используемыми на практике. Представлены численные результаты.

Ключевые слова: авторегрессия, прогнозирование, цензурирование, риск.

Задача статистического прогнозирования возникает во многих приложениях: в медицине, экономике, метеорологии, технике, астрономии [1]. Для описания временных рядов с зависимыми наблюдениями и прогнозирования будущих значений широко применяется модель авторегрессии [1]. На практике значения временного ряда часто наблюдаются с искажениями различных типов: выбросы, пропуски, гетероскедастичность [2], цензурирование [3] и др.; обзор типов искажений и их математические описания представлены в [2]. Цензурирование временного ряда заключается в том, что часть наблюдений ряда известна точно, а об остальных наблюдениях известно лишь, что они принадлежат некоторым числовым интервалам. Такая ситуация может возникать из-за наличия у приборов конечных пределов измерения, высокой стоимости проведения точных измерений, разладки оборудования и других причин.

Цензурированные выборки независимых наблюдений подробно изучены в математической теории надежности [4]. Однако статистический анализ цензурированных временных рядов остается малоизученным и актуальным направлением исследований.

Математическая модель

Пусть временной ряд x_t описывается моделью AP(p) авторегрессии порядка $p \in \mathrm{N}$ [1]:

$$x_t = \sum_{i=1}^{p} \theta_i x_{t-i} + u_t, t \in Z, \quad (1)$$

где $\{\theta_i\}_{i=1}^{p}$ — коэффициенты авторегрессии такие, что все корни порождающего характеристического многочлена $z^p - \sum_{j=1}^{p} \theta_j z^{p-j}$ лежат внутри единичного круга; $\{u_t\}$ — независимые в совокупности одинаково распределенные случайные величины, имеющие нормальный закон распределения вероятностей: $L\{u_t\} = \mathrm{N}(0, \sigma^2)$.

Пусть вместо значений временного ряда наблюдаются случайные события:

$$A_t^* = \{x_t \in A_t\}, t \in \{1, \ldots, T\}, \quad (2)$$

где $\{A_t\}$ — заданные борелевские множества, $T > p$ — длительность наблюдения. При наличии интервального цензурирования возможны два случая: 1) A_t состоит из одного элемента ($A_t = \{x_t\}$), тогда значение x_t известно точно; 2) A_t является числовым интервалом ($A_t = [a_t, b_t)$, $a_t < b_t$), тогда имеет место интервальное цензурирование значения x_t, а интервал $[a_t, b_t)$ называется интервалом цензурирования. Статистическое прогнозирование будущего значения $x_{T+1} \in \mathrm{R}$ заключается в вычислении оценки $\hat{x}_{T+1} \in \mathrm{R}$ на основе имеющейся информации о наступлении событий $A_1^*, \ldots, A_T^*$:

$$\hat{x}_{T+1} = f\left(A_T^*, A_{T-1}^*, \ldots, A_1^*\right). \quad (3)$$

Погрешность прогнозирования будем характеризовать условным риском прогнозирования

$$r_T(f) = \mathrm{E}\left\{\left(x_{T+1} - \hat{x}_{T+1}\right)^2 \middle| A_T^*, \ldots, A_1^*\right\} \geq 0, \quad (4)$$

т.е. среднеквадратической ошибкой прогнозирования.

Рассмотрим задачу построения оптимальной прогнозирующей статистики (ОПС) $f_0(\cdot)$, минимизирующей условный риск (4), в случае известных параметров модели (1), (2):

$$r_T(f_0) = \inf_{f(\cdot)} r_T(f). \tag{5}$$

Основные результаты

Теорема 1. Если временной ряд x_t наблюдается при наличии цензурирования общего вида (2), то среди всех прогнозирующих статистик вида (3) оптимальная по критерию минимума риска (5) прогнозирующая статистика определяется условным математическим ожиданием:

$$\hat{x}_{T+1} = f_0\left(A_T^*,\ldots,A_1^*\right) = \mathrm{E}\left\{x_{T+1}\middle|A_T^*,\ldots,A_1^*\right\},\ r_T(f_0) = \mathrm{D}\left\{x_{T+1}\middle|A_T^*,\ldots,A_1^*\right\}. \tag{6}$$

Доказательство. Преобразуем условный риск (4):

$$r_T(f) = \mathrm{E}\left\{\left(x_{T+1}-\hat{x}_{T+1}\right)^2\middle|A_T^*,\ldots,A_1^*\right\} = \mathrm{E}\left\{\left(x_{T+1}-f\left(A_T^*,\ldots,A_1^*\right)\right)^2\middle|A_T^*,\ldots,A_1^*\right\} =$$

$$= \mathrm{E}\left\{\left(\left(x_{T+1}-\mathrm{E}\left\{x_{T+1}\middle|A_T^*,\ldots,A_1^*\right\}\right)+\left(\mathrm{E}\left\{x_{T+1}\middle|A_T^*,\ldots,A_1^*\right\}-f\left(A_T^*,\ldots,A_1^*\right)\right)\right)^2\middle|A_T^*,\ldots,A_1^*\right\}.$$

Заметим, что второе слагаемое зависит только от $A_T^*,\ldots,A_1^*$:

$$r_T(f) = \mathrm{E}\left\{\left(x_{T+1}-\mathrm{E}\left\{x_{T+1}\middle|A_T^*,\ldots,A_1^*\right\}\right)^2\middle|A_T^*,\ldots,A_1^*\right\}+\left(\mathrm{E}\left\{x_{T+1}\middle|A_T^*,\ldots,A_1^*\right\}-f\left(A_T^*,\ldots,A_1^*\right)\right)^2 +$$

$$+2\,\mathrm{E}\left\{\left(x_{T+1}-\mathrm{E}\left\{x_{T+1}\middle|A_T^*,\ldots,A_1^*\right\}\right)\middle|A_T^*,\ldots,A_1^*\right\}\left(\mathrm{E}\left\{x_{T+1}\middle|A_T^*,\ldots,A_1^*\right\}-f\left(A_T^*,\ldots,A_1^*\right)\right) =$$

$$= \mathrm{D}\left\{x_{T+1}\middle|A_T^*,\ldots,A_1^*\right\}+\left(\mathrm{E}\left\{x_{T+1}\middle|A_T^*,\ldots,A_1^*\right\}-f\left(A_T^*,\ldots,A_1^*\right)\right)^2 \to \min_{f(\cdot)},$$

Из этого представления следует, что (6) есть решение задачи (5). ■

Теорема 1 является обобщением известного результата [1] в ситуации, когда цензурирование отсутствует.

Рассмотрим случай, когда цензурированы только последние q, $(0\le q\le T)$, значений временного ряда, а остальные $T-q$ наблюдений известны точно. Обозначим

$$\mu(t,m) = \theta_1 x_{t-1}+\ldots+\theta_m x_{t-m} = \sum_{i=1}^{m}\theta_i x_{t-i},\ t,m\in\mathrm{N}.$$

Теорема 2. Пусть в рамках модели (1), (2) наблюдаются значения $x_1,\ldots,x_{T-q}$ и случайные события $A_{T-q+1}^* = \left\{x_{T-q+1}\in\left[a_{T-q+1},b_{T-q+1}\right)\right\}$, …, $A_T^* = \left\{x_T\in\left[a_T,b_T\right)\right\}$. Тогда ОПС имеет вид:

$$\hat{x}_{T+1} = \frac{\int_{a_T}^{b_T}\ldots\int_{a_{T-q+1}}^{b_{T-q+1}}\mu(T+1,p)\,p(x_T,\ldots,x_{T-q+1}|x_{T-q},\ldots,x_1)\,dx_{T-q+1}\ldots dx_T}{\int_{a_T}^{b_T}\ldots\int_{a_{T-q+1}}^{b_{T-q+1}}p(x_T,\ldots,x_{T-q+1}|x_{T-q},\ldots,x_1)\,dx_{T-q+1}\ldots dx_T}. \tag{7}$$

Доказательство. Оценку (6) в силу (1) можно представить следующим образом:

$$\hat{x}_{T+1} = \mathrm{E}\left\{x_{T+1}\middle|A_T^*,\ldots,A_{T-q+1}^*,x_{T-q},\ldots,x_1\right\} = \mathrm{E}\left\{\sum_{i=1}^{p}\theta_i x_{T+1-i}+u_{T+1}\middle|A_T^*,\ldots,A_{T-q+1}^*,x_{T-q},\ldots,x_1\right\} =$$

$$= \mathrm{E}\left\{\mu(T+1,p)\middle|A_T^*,\ldots,A_{T-q+1}^*,x_{T-q},\ldots,x_1\right\},$$

так как случайная величина u_{T+1} не зависит от $A_T^*,\ldots,A_{T-q+1}^*$, $x_{T-q},\ldots,x_1$ и $\mathrm{E}\{u_{T+1}\}=0$. Вычислив полученное математическое ожидание, получим требуемое равенство (7). ■

Введём обозначение:

$$\Psi(x,y,m,s,u,v)=\left(u\,\varphi((x-m)/s)-v\,\varphi((y-m)/s)\right)\left(\Phi((y-m)/s)-\Phi((x-m)/s)\right)^{-1},\ x,y,m,s,u,v\in\mathrm{R},$$

где $\varphi(x)=\left(1/\sqrt{2\pi}\right)\exp(-x^2/2)$, $\Phi(x)=\int_{-\infty}^{x}\varphi(t)dt$ — соответственно плотность и функция распределения вероятностей стандартного нормального закона N(0,1).

Следствие. Если в рамках модели (1), (2) наблюдаются значения $x_1,\ldots,x_{T-1}$ и случайное событие $A_T^*=\{x_T\in[a_T,b_T)\}$ (q=1), то ОПС (4) имеет вид:

$$\hat{x}_{T+1}=\theta_1\mu(T,p)+\textstyle\sum_{i=2}^{p}\theta_i x_{T-i+1}+\theta_1\sigma\Psi(a_T,b_T,\mu(T,p),\sigma,1,1).$$

Доказательство. Воспользовавшись известным соотношением для модели АР(p) [1]:

$$p(x_T|x_{T-1},\ldots,x_1)=\varphi((x_T-\mu(T,p))/\sigma)/\sigma$$

и теоремой 2 для q=1, получим требуемое соотношение. ■

Если среди последних q значений временного ряда имеются не только цензурированные наблюдения, но и k ($1\le k\le q$) известных наблюдений $x_{l_1},\ldots,x_{l_k}$ (T-q+$1\le l_1<\ldots<l_k\le T$), то ОПС может быть получена из (7) предельным переходом $b_{l_1}\to a_{l_1},\ldots,b_{l_k}\to a_{l_k}$.

Рассмотрим частный случай модели (1) —авторегрессию первого порядка (p=1):

$$x_t=\theta x_{t-1}+u_t,\ t\in Z, \tag{8}$$

и q=1, причем предполагается, что параметры модели θ и σ известны. Для этого случая исследуем зависимость условного риска прогнозирования от длины интервала цензурирования и проведем сравнение ОПС $f_0(\cdot)$ с прогнозирующими статистиками, часто используемыми на практике [3]. В этом случае последнее значение x_T временного ряда цензурировано интервалом $[a_T,b_T)$, а предпоследнее значение x_{T-1} известно точно. Поскольку в данном случае результат зависит только от одного интервала цензурирования $A_T=[a_T,b_T)$, то для упрощения обозначений вместо a_T и b_T будем писать a и b. Используя теорему 2, может быть доказана следующая теорема.

Теорема 3. Пусть для модели (8) наблюдаются значение x_{T-1} и случайное событие $A_T^*=\{x_T\in[a,b)\}$, тогда ОПС и её условный риск имеют вид:

$$\hat{x}_{T+1}=f_0\left(A_T^*,x_{T-1}\right)=\theta^2x_{T-1}+\theta\sigma\Psi(a,b,\theta x_{T-1},\sigma,1,1), \tag{9}$$

$$r_T(f_0)=(1+\theta^2)\sigma^2-\left(\theta\sigma\Psi(a,b,\theta x_{T-1},\sigma,1,1)\right)^2+\theta^2\sigma\Psi\left(a,b,\theta x_{T-1},\sigma,a-\theta\ x_{T-1},b-\theta\ x_{T-1}\right). \tag{10}$$

Следствие. В условиях теоремы 3 для условного риска прогнозирования справедливо асимптотическое разложение при $\tau=b-a\to 0$:

$$r_T(f_0)=\sigma^2+\theta^2\tau^2/12-\theta^2\tau^4\left(3a^2-6a\theta x_{T-1}+3\theta^2x_{T-1}^2+2\sigma^2\right)/720\sigma^4+o(\tau^4).$$

Доказательство. Учитывая дифференцируемость функции $\Psi(\cdot)$ в (10) по τ, воспользуемся формулой Тейлора с остаточным членом в форме Пеано и получим требуемое соотношение для условного риска. ■

Из доказанного следствия получаем, что безусловный риск ОПС имеет следующее асимптотическое разложение при $\tau=b-a\to 0$:

$$\mathrm{E}\{r_T(f_0)\}=\sigma^2+\theta^2\tau^2/12-\theta^2\tau^4\left(3a^2+3\theta^2\sigma^2/\left(1-\theta^2\right)+2\sigma^2\right)/720\sigma^4+o(\tau^4). \tag{11}$$

Одной из возможных альтернативных прогнозирующих статистик является статистика [3]:

$$\hat{x}_{T+1}=f_1(A_T^*)=\theta\,\mathrm{E}\left\{x_T\mid A_T^*\right\}=\theta\,\mathrm{E}\{x_T\mid x_T\in[a,b)\}. \tag{12}$$

Теорема 4. Если для модели (8) наблюдается случайное событие $A_T^* = \{x_T \in [a,b)\}$, то прогнозирующая статистика (12) имеет вид:

$$\hat{x}_{T+1} = f_1(A_T^*) = \left(\theta\sigma/\sqrt{1-\theta^2}\right)\Psi\left(a,b,0,\theta\sigma/\sqrt{1-\theta^2},1,1\right), \qquad (13)$$

и её условный риск прогнозирования равен:

$$r_T(f_1) = \frac{\sigma^2}{1-\theta^2} - \frac{\theta^2\sigma^2}{1-\theta^2}\left(\Psi\left(a,b,0,\frac{\sigma}{\sqrt{1-\theta^2}},1,1\right)\right)^2 + \frac{\theta^2\sigma}{\sqrt{1-\theta^2}}\Psi\left(a,b,0,\frac{\sigma}{\sqrt{1-\theta^2}},a,b\right). \qquad (14)$$

Доказательство. Прогнозирующая статистика (12) имеет вид

$$\hat{x}_{T+1} = \theta\,\mathrm{E}\{x_T \mid x_T \in [a,b)\} = \theta\left(\int_a^b x\mathrm{n}\left(x\middle|0,\sigma^2/1-\theta^2\right)dx\right)\left(\int_a^b \mathrm{n}\left(x\middle|0,\sigma^2/1-\theta^2\right)dx\right)^{-1}.$$

Воспользовавшись [5] для вычисления интегралов, получим (13). Аналогично вычисляется условный риск прогнозирования (14). ■

Следствие. В условиях теоремы 4 для условного риска прогнозирования справедливо асимптотическое разложение при $\tau = b - a \to 0$:

$$r_T(f_1) = \sigma^2 + \theta^2\tau^2/12 - \theta^2\left(1-\theta^2\right)^2\tau^4\left(3a^2 + 2\sigma^2/(1-\theta^2)\right)/720\sigma^4 + o(\tau^4).$$

Доказательство. Проводится аналогично доказательству следствия теоремы 3. ■

Сравнивая $\mathrm{E}\{r_T(f_0)\}$ и $r_T(f_1)$ при $\tau \to 0$, замечаем, что усредненный риск ОПС (11) меньше риска прогнозирующей статистики (12).

Рассмотрим ещё одну часто используемую прогнозирующую статистику:

$$\hat{x}_{T+1} = f_2\left(A_T^*\right) = \theta(a+b)/2. \qquad (15)$$

Теорема 5. Если для модели (8) наблюдается случайное событие $A_T^* = \{x_T \in [a,b)\}$, то условный риск прогнозирования для статистики (15) равен:

$$r_T(f_2) = \sigma^2/1-\theta^2 + \theta^2(a+b)^2/4 - \left(\theta^2\sigma/\sqrt{1-\theta^2}\right)\Psi\left(a,b,0,\sigma/\sqrt{1-\theta^2},b,a\right). \qquad (16)$$

Доказательство. Проводится аналогично доказательству теоремы 4. ■

Следствие. Пусть выполнены условия теоремы 5 и $\tau = b - a \to 0$. Тогда для условного риска справедливо следующее асимптотическое разложение:

$$r_T(f_2) = \sigma^2 + \theta^2\tau^2/12 + \theta^2\left(1-\theta^2\right)^2\tau^4\left(a^2 - \sigma^2/(1-\theta^2)\right)/360\sigma^4 + o(\tau^4).$$

Доказательство. Проводится аналогично доказательству следствия теоремы 3. ■

Легко видеть, что при τ близких к нулю, риск прогнозирования статистики (15) больше, чем риск прогнозирования для статистик (9) и (12).

Численные результаты

Для сравнения прогнозирующих статистик (9), (12) и (15) проведены компьютерные эксперименты. Применялся метод Монте-Карло с числом прогонов N=10000. Для моделирования временного ряда использованы значения параметров: p=1, θ=0.8, σ=1, q=1, T=100, $\tau \in \{0,0.5,\ldots,15\}$, по наблюдению x_T строился интервал цензурирования (a,b) длины τ, где $a = x_T - \alpha\tau$ и $b = x_T + \alpha(1-\tau)$, α — случайная величина, равномерно распределённая на $[0,1]$.

На рисунке 1(а) изображены графики зависимостей экспериментальных значений риска для всех трёх прогнозирующих статистик от τ. Как видно из рисунка, ОПС (9) имеет наименьший риск, риск статистики (12) принимает большие значения, а риск статистики (15) возрастает очень быстро и уже при малых τ принимает достаточно большие значения.

На рисунках 1(б) – 1(г) изображены усредненные теоретические значения риска прогнозирования для статистик (9), (12) и (15) в зависимости от τ, вычисленные по формулам (10), (14) и (16) соответственно, и 95%-ные доверительные границы риска.

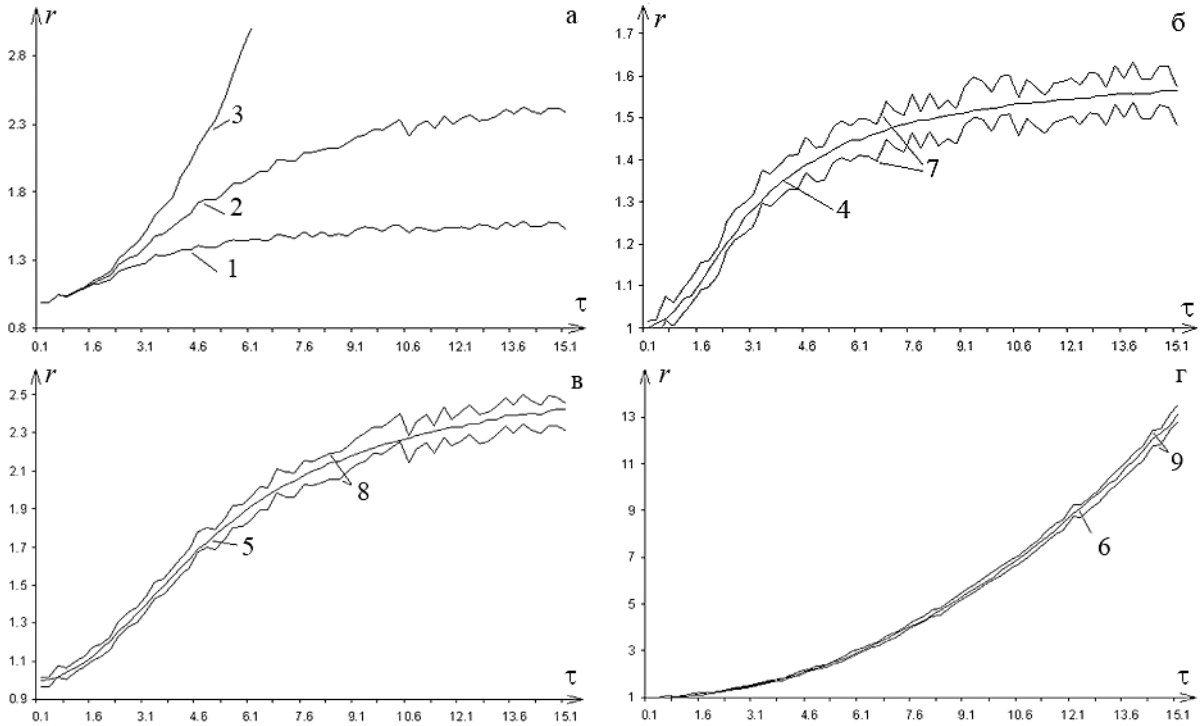

Рис.1. Результаты численных экспериментов: а) Сравнение трех прогнозирующих статистик: 1 –риск оптимальной прогнозирующей статистики (9), 2 –риск статистики (12), 3 – риск статистики (15); б) – г) сравнение теоретических и экспериментальных значений риска прогнозирующих статистик (9), (12) и (15): 4 – теоретическое значение риска статистики (9), 5 – теоретическое значение риска статистики (12), 6 – теоретическое значение риска статистики (15), 7 - 9 – 95% доверительные границы соответствующих значений риска

Таким образом, в настоящей работе найдена ОПС для авторегрессионных временных рядов при наличии цензурирования и ее риск; в случае авторегрессии первого порядка проведено сравнение ОПС с прогнозирующими статистиками, часто используемыми на практике; проведены компьютерные эксперименты, которые показали, что экспериментальные и теоретические значения риска находятся в хорошем согласии.

Литература

1. Андерсон Т. Статистический анализ временных рядов: моногр. М.:Мир, 1976.

2. Харин Ю.С. Оптимальность и робастность в статистическом прогнозировании: моногр. Мн.: БГУ, 2008.

3. Park J. W., Genton M. G., Ghosh S. K. Censored time series analysis with autoregressive moving average models // The Canadian Journal of Statistics. Vol.35, № 1. 2007. P. 151–168.

4. Gomez G., Espinal A., Lagakos W. Inference for a linear model with an interval-censored covariate // Statistics in medicine. № 22. 2003. P. 409–425.

5. Градштейн И.С., Рыжик И.М. Таблицы интегралов, сумм, рядов и произведений. М.: Физматгиз, 1963.

I.A. Badziahin, Yu.S. Kharin

Forecasting of autoregressive time series under censoring

Problems of statistical forecasting are considered for autoregressive time series observed under interval censoring. Optimal forecasting statistic is proposed, its mean-square risk is evaluated. Comparison of optimal and widely used in practice forecasting statistics is made. Numerical results are given.

Keywords: autoregression, forecasting, censoring, risk.