О СТАТИСТИЧЕСКОМ ОЦЕНИВАНИИ ПАРАМЕТРОВ РЕГРЕССИИ ПРИ НАЛИЧИИ СЛУЧАЙНОГО ЦЕНЗУРИРОВАНИЯ

Е. С. Агеева

МОДЕЛЬ

Рассмотрим модель множественной линейной регрессии, заданную уравнением [1]:

$$Y_t = \sum_{i=1}^{M} \theta_i X_t^i + \theta_0 + \xi_t, t = 1,...,n, \quad (1)$$

где $\{\xi_t\}$ – независимые в совокупности нормальные одинаково распределённые случайные величины с математическим ожиданием 0 и дисперсией $0 < \sigma^2 < \infty$; $\{\theta_0, \theta_1, ..., \theta_M\}^T$ называются коэффициентами регрессии, $\{X_t^i\}$, $i = 1,...,M$ – регрессорами, $\{\xi_t\}$ – случайными величинами ошибок. Будем предполагать σ^2 известным. Нами будут наблюдаться значения регрессоров $\{X_t^i\}$, $i = 1,...,M$, $t = 1,...,n$ и события $Y_1 \in [a_1, b_1], ..., Y_n \in [a_n, b_n]$ вместо точных значений $Y_1, ..., Y_n$; здесь $[a_t, b_t]$, $t = 1,...,n$ – интервал цензурирования, $a_t \le b_t$.

Требуется построить оценки максимального правдоподобия для вектора регрессионных коэффициентов $\theta = (\theta_0, \theta_1, ..., \theta_M)^T$ и получить асимптотическое выражение этой оценки в асимптотике $\Delta_t = b_t - a_t \to 0$, используя разложение в ряд по Δ_t. Для найденных оценок требуется получить выражения для смещения и матрицы вариаций.

ОЦЕНКИ И ИХ СВОЙСТВА

Пусть матрица $X = \begin{pmatrix} 1 & X_1^1 & ... & X_1^M \\ 1 & X_2^1 & ... & X_2^M \\ ... & ... & ... & ... \\ 1 & X_n^1 & ... & X_n^M \end{pmatrix} = (X_{t,j})_{\substack{t=1,...,n \\ j=1,...,M+1}}$ – это $n \times (M+1)$- матрица эксперимента, $\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{t^2}{2}} dt$ – функция Лапласа.

Теорема 1. В случае множественной линейной регрессии (1) с интервальным цензурированием оценка максимального правдоподобия $\hat{\theta} = (\hat{\theta}_0, \hat{\theta}_1, ..., \hat{\theta}_M)^T$ является решением системы $M+1$ нелинейных уравнений:

$$\sum_{t=1}^{n} X_{t,j} \frac{\partial P_t}{\partial \theta_0} \Big/ P_t = 0, \;\; j = 1, ..., M+1 \qquad (2)$$

при условии, что $(M+1)\times(M+1)$-матрица $A(\hat{\theta})$ отрицательно определена: $A(\hat{\theta}) \prec 0$. Здесь

$$P_t = \Phi((b_t - \theta_0 - \theta_1 X_t^1 - ... - \theta_M X_t^M)/\sigma) - \Phi((a_t - \theta_0 - \theta_1 X_t^1 - ... - \; - \theta_M X_t^M)/\sigma),$$

$$A(\hat{\theta}) = \left(\sum_{t=1}^{n} X_{t,i} X_{t,j} \left(\frac{1}{\sigma} \frac{\partial^2 P_t}{\partial {\theta_0}^2} P_t - \left(\frac{\partial P_t}{\partial \theta_0} \right)^2 \right) \Big/ P_t^2 \right)_{i,j=1}^{M+1}.$$

Оценка максимального правдоподобия находится из условия [1]:

$$l(\theta) = l(\theta_0, \theta_1, ..., \theta_M) = \sum_{t=1}^{n} \ln P(Y_t \in [a_t, b_t]) = \sum_{t=1}^{n} \ln P_t \xrightarrow[\theta_0, \theta_1, ..., \theta_M]{} \max. \qquad (3)$$

Раскладывая функцию $l(\theta)$ по степеням $\Delta_t = b_t - a_t$, получим 3 асимптотических выражения для оценки максимального правдоподобия, имеющих вид:

$$\hat{\theta} = C^{-1} F. \qquad (4)$$

Здесь

$$C_{ij} = \frac{1}{n} \sum_{t=1}^{n} B_t X_{t,i} X_{t,j}, \; F_i = \frac{1}{n} \sum_{t=1}^{n} B_t X_{t,i} \frac{a_t + b_t}{2}, \; i, j = 1, ..., M+1, \qquad (5)$$

где в случае $B_t = 1, \;\; t = 1, ..., n$ порядок аппроксимации равен 1, а в случае

$B_t = 1 - \frac{{\Delta_t}^2}{12\sigma^2}$ или $B_t = \beta_t = \frac{\Delta_t e^{-\frac{\Delta_t}{8\sigma^2}}}{12\sqrt{2\pi}\sigma^2 (\Phi(\frac{\Delta_t}{2\sigma}) - \Phi(-\frac{\Delta_t}{2\sigma}))}$, $t = 1, ..., n$ порядок аппроксимации равен 3.

Пусть рассматривается множество всевозможных интервалов на числовой прямой, имеющих фиксированную длину $\Delta > 0$:

$$S_\Delta \{(a,b) : a, b \in R, b - a = \Delta\} = \{(a, a+\Delta) : a \in R\} \subset \mathrm{B}(R).$$

Пусть далее $F(x), x \in R, -$ некоторая абсолютно непрерывная функция распределения вероятностей, задающая на $\mathrm{B}(R)$ вероятностную меру $P\{\cdot\}$, абсолютно непрерывную относительно меры Лебега:

$$P\{(a,b)\} = P\{(a,a+\Delta)\} = F(a+\Delta) - F(a) \geq 0, a \in R. \quad (6)$$

Случайный интервал $(a,a+\Delta) \in S_\Delta$ полностью определяется заданием его левой границы a. Её плотность распределения вероятностей:

$$p_a(x;\Delta) = \frac{F(x+\Delta) - F(x)}{\int_{-\infty}^{\infty} (F(y+\Delta) - F(y))dy} \geq 0. \quad (7)$$

Лемма 1. Пусть $F(x), x \in R$ – произвольная симметричная функция распределения вероятностей, такая что $\lim_{x\to+\infty} xF(-x) = 0$. Тогда функция

$$p(x;\Delta) = \frac{F(x+\Delta) - F(x)}{\Delta}, x \in R$$

является плотностью распределения вероятностей левой границы a случайного интервала $(a,a+\Delta) \in S_\Delta$. В частности, если $F(x) = \Phi(x), x \in R$ – функция распределения вероятностей стандартной нормальной случайной величины, то

$$p(x;\Delta) = \frac{\Phi(x+\Delta) - \Phi(x)}{\Delta}, x \in R.$$

По лемме 1 плотность распределения вероятностей нижней границы a_t равна

$$p_{a_t}(x;\Delta) = \frac{\Phi(\frac{x+\Delta_t - \theta_0 - \theta_1 X_t^1 - ... - \theta_M X_t^M}{\sigma}) - \Phi(\frac{x - \theta_0 - \theta_1 X_t^1 - ... - \theta_M X_t^M}{\sigma})}{\Delta_t}. \quad (8)$$

Лемма 2. Если случайная величина a_t имеет плотность распределения вероятностей, заданную формулой (8), то её математическое ожидание и второй момент имеют следующий вид:

$$E\{a_t\} = \theta_0 + \theta_1 X_t^1 + ... + \theta_M X_t^M - \frac{\Delta_t}{2},$$

$$E\{a_t^2\} = \sigma^2 + (\theta_0 + \theta_1 X_t^1 + ... + \theta_M X_t^M)^2 - (\theta_0 + \theta_1 X_t^1 + ... + \theta_M X_t^M)\Delta_t + \frac{\Delta_t^2}{3}.$$

Теорема 2. Если матрица C, определённая в (5), невырожденная, то оценки (4) являются несмещёнными, а матрица вариаций равна:

$$V(\hat{\theta}) = C^{-1}BC^{-1}, \quad (9)$$

где элементы матрицы B имеют вид:

$$B_{ij} = \frac{1}{n^2}\sum_{t=1}^{n} X_{t,i}X_{t,j}B_t{}^2(\sigma^2 + \frac{\Delta_t{}^2}{12}), i, j = 1,...,M+1. \qquad (10)$$

Теорема 3. Пусть $|B_t| \le D_1$, $\sigma^2 + \frac{\Delta_t{}^2}{12} \le D_2$, $t = 1,...,n$, а ряды $\frac{1}{n}\sum_{t=1}^{n}|X_t^i X_t^j|, \frac{1}{n}\sum_{t=1}^{n}|X_t^i|, i, j = 1,...,M$ ограничены сверху константой D_3, сразу для всех n, $i, j = 1,...,M$. Тогда элементы матрицы вариаций $V(\hat{\theta})$ будут стремиться к 0 при $n \to \infty$.

Следствие 1. В условиях теоремы 3 оценки (4) являются состоятельными по вероятности.

Теорема 4. Пусть $D = (\delta_{ij}\frac{\Delta_i{}^2}{12})_{i,j=1}^{n}$, $Y = (\frac{a_1 + b_1}{2},...,\frac{a_n + b_n}{2})^T$, а матрица C, определённая в (5), невырожденная. Тогда несмещённой, состоятельной по вероятности оценкой дисперсии ошибок σ^2 будет статистика

$$\hat{\sigma}^2 = \frac{1}{n-M-1}\left((Y - X\hat{\theta})'(Y - X\hat{\theta}) - tr((I_n - X(X'X)^{-1}X')D)\right). \qquad (11)$$

ЗАКЛЮЧЕНИЕ

В работе рассмотрена модель множественной линейной регрессии при наличии случайного цензурирования. Получены три асимптотических выражения ОМП в асимптотике $\Delta_t \to 0$. Для нижней границы интервала цензурирования a_t, $t = 1,...,n$, найдена плотность распределения вероятностей $p_{a_t}(x)$. Доказана несмещённость и асимптотическая состоятельность предложенных оценок, а так же найдены матрицы вариаций для каждой из них. Для дисперсии ошибок σ^2 предложена несмещённая оценка.

Литература

1. *Харин Ю. С., Жук Е. Е.* Математическая и прикладная статистика: учеб. пособие. Мн., БГУ, 2005.
2. *Gang Li, Cun-Hui Zhang.* Linear regression with interval censored data // The Annals of Statistics. 1998. Vol. 26. N. 4. PP. 1306–1327.
3. *Gomez G.,Espinal A., Lagakos W.* Inference for a linear model with an interval-censored covariate // Statistics in medicine. 2003. Vol. 22. P. 409–425.
4. *Koul H., Susarla V., Ryzin V.J.* Regression analysis with randomly right-censored data // The Annals of Statistics. 1981. Vol. 9. N. 6. P. 1276–1288.
5. *Zhou L.* A simple censored median regression estimator // Statistica Sinica. 2006. Vol. 16. P.1043–1058.