

СТАТИСТИЧЕСКОЕ ОЦЕНИВАНИЕ ПАРАМЕТРОВ МНОЖЕСТВЕННОЙ ЛИНЕЙНОЙ РЕГРЕССИИ ПРИ НАЛИЧИИ КЛАССИФИКАЦИИ НАБЛЮДЕНИЙ

Е. С. Агеева

ВВЕДЕНИЕ

В статистике часто встречается регрессионная модель. Ею описываются многие процессы в технике, экономике, медицине и т.д. В данной работе рассмотрена множественная регрессионная модель в случае, когда сами зависимые данные не наблюдаются, а наблюдаются только множества (классы), в которые попадают эти данные.

Подобные модели с классифицированными данными появились давно [4]. В литературе рассматривается случай так называемых “округлённых данных” (rounded data). Округление данных может быть вызвано точностью измерительного прибора или накопительного устройства. Такие проблемы возникают в различных моделях: во временных рядах авторегрессии и скользящего среднего [1], регрессионных моделях [2] и т.д. Во многих статьях рассматривается влияние округления на оценку математического ожидания и дисперсии для случайных величин, распределённых по нормальному закону [1], [3], [5]. Модель, рассмотренная в работе, является обобщением rounded data в регрессии.

1. МАТЕМАТИЧЕСКАЯ МОДЕЛЬ

Рассмотрим модель нелинейной множественной регрессии:

$$Y_t = F(X_t; \theta^0) + \xi_t, \quad t = 1, \dots, n, \quad (1)$$

где n объём выборки; $\theta^0 = (\theta_1^0, \dots, \theta_m^0)^T \in \Theta \subseteq R^m$ – неизвестный вектор параметров; $X_t = (X_t^1, \dots, X_t^N)^T \in X \subseteq R^N$ – наблюдаемый вектор регрессоров; $Y_t \in R^1$ – ненаблюдаемая зависимая переменная; $\xi_t \in R^1$ – случайная величина ошибок с нормальной плотностью распределения вероятностей с математическим ожиданием 0 и дисперсией $0 < \sigma^2 < \infty$; $F(\cdot) : X \times \Theta \rightarrow R^1$ – функция регрессии.

Будем предполагать, что план эксперимента $\{X_t\}_{t=1}^n$ задаётся вручную, то есть является неслучайным. Считаем, что $\{\xi_t\}_{t=1}^n$ – независимые в совокупности.

Определена последовательность K непересекающихся борелевских множеств ($K \geq 2$):

$$A_1, \dots, A_K \in B(R^1), \cup_{k=1}^K A_k = R^1, A_i \cap A_j = \emptyset, i \neq j.$$

Эта система борелевских множеств задаёт классификацию Y_t :

$$Y_t \text{ относится к классу } \Omega_{\nu_t}, \text{ если } Y_t \in A_{\nu_t}, \nu_t \in \{1, \dots, K\}. \quad (2)$$

Предположим, что множества $A_1, \dots, A_K \in B(R^1)$ являются интервалами и имеют следующий вид:

$$A_k = (a_{k-1}, a_k], k=1, \dots, K, a_0 = -\infty, a_K = +\infty. \quad (3)$$

Вместо точных наблюдений $Y_1, \dots, Y_n$ наблюдаются лишь соответствующие номера классов $\nu_1, \dots, \nu_n \in \{1, \dots, K\}$. Задача заключается в том, чтобы по классифицированным наблюдениям $\nu_1, \dots, \nu_n$ и значениям регрессоров $X_1, \dots, X_n$ построить оценки для неизвестного вектора параметров θ^0 и дисперсии ошибок σ^2 .

2. ОЦЕНКИ МАКСИМАЛЬНОГО ПРАВДОПОДОБИЯ

Наблюдаются дискретные случайные величины $\{\nu_t\}_{t=1}^n$, связанные с Y_t стохастической зависимостью, порождаемой (1)-(3):

$$P_{X_t, \theta, \sigma^2} \{Y_t \in A_k\} = P_{X_t, \theta, \sigma^2} \{\nu_t \in k\} = P_{X_t}(\nu_t; \theta, \sigma^2) =$$

$$\frac{1}{\sqrt{2\pi}\sigma} \int_{A_k} e^{-\frac{(z - F(X_t, \theta))^2}{2\sigma^2}} dz, k=1, \dots, K.$$

В силу независимости $\{\nu_t\}_{t=1}^n$ логарифмическая функция правдоподобия имеет вид:

$$l(\theta, \sigma^2) = \sum_{t=1}^n \ln P_{X_t}(\nu_t; \theta, \sigma^2).$$

Максимизируя функцию $l(\theta, \sigma^2)$ по θ и σ^2 , найдём оценки максимального правдоподобия [6]:

$$\hat{\theta}, \hat{\sigma}^2: l(\hat{\theta}, \hat{\sigma}^2) = \max_{\theta, \sigma^2} l(\theta, \sigma^2).$$

Лемма 1. Если множества $A_1, \dots, A_K \in B(R^1)$ имеют вид (3), то логарифмическую функцию правдоподобия можно записать в виде:

$$l(\theta, \sigma^2) = \sum_{t=1}^n \ln \left(\Phi \left(\frac{a_{v_t} - F(X_t; \theta)}{\sigma} \right) - \Phi \left(\frac{a_{v_t-1} - F(X_t; \theta)}{\sigma} \right) \right)$$

Теорема 1. Пусть Θ – замкнутое подмножество R^m ; существует такое $\bar{\sigma}^2$, что $\bar{\sigma}^2 \leq \sigma^2$; $2 < K < +\infty$. И пусть существуют такие $0 < d$, $0 < p < 1$, что план эксперимента $\{X_t : X_t \in X \subseteq R^N\}_{t=1}^n$ обладает следующим свойством: для любого $(\theta, \sigma^2) \in \Theta \times [\bar{\sigma}^2, \infty)$, $\theta \neq \theta^0$, $\sigma^2 \neq \sigma^{0^2}$, начиная с некоторого объёма выборки $n > n_1$ для $[pn] + 1$ наблюдений из $\{X_t\}_{t=1}^n$ верно

$$E_{\theta^0, \sigma^{0^2}} \{ \ln P_{X_t}(v_t; \theta, \sigma^2) \} - E_{\theta^0, \sigma^{0^2}} \{ \ln P_{X_t}(v_t; \theta^0, \sigma^{0^2}) \} \leq -d;$$

для любой последовательности $\{\theta^i : \theta^i \in \Theta, i \in N\}$ такой, что $|\theta^i| \xrightarrow{i \rightarrow \infty} \infty$, выполнено $|F(X, \theta^i)| \xrightarrow{i \rightarrow \infty} \infty$, $X \in X \subseteq R^N$; для любого фиксированного значения $\theta \in \Theta$ функция $F(X, \theta)$ ограничена на $X \subseteq R^N$. Тогда ОМП $(\hat{\theta}, \hat{\sigma}^2)$ является сильно состоятельной, т.е.

$$(\hat{\theta}, \hat{\sigma}^2) \xrightarrow{P=1} (\theta, \sigma^2).$$

Информационная матрица Фишера в точке (θ, σ^2) для модели (1)-(3) будет иметь вид:

$$I_n(\theta, \sigma^2) = \sum_{t=1}^n B_{X_t}^{\theta, \sigma^2} \nabla_{\theta} F(X_t, \theta) (\nabla_{\theta} F(X_t, \theta))^T,$$

где $B_X^{\theta, \sigma^2} = \sum_{k=1}^K \frac{\frac{1}{\sigma^2} (\varphi(\frac{a_k - F(X, \theta)}{\sigma}) - \varphi(\frac{a_{k-1} - F(X, \theta)}{\sigma}))^2}{P_X(k; \theta, \sigma^2)}.$

Теорема 2. Пусть выполнены условия теоремы 1. А так же пусть в точке (θ^0, σ^{0^2}) информационная матрица Фишера невырожденная и

$$\lim_{n \rightarrow \infty} |\frac{1}{n} I_n(\theta^0, \sigma^{0^2})| \neq 0.$$

Тогда ОМП $\hat{\theta}$ асимптотически нормально распределена:

$$L \left\{ (I_n(\theta^0, \sigma^{0^2})^{-\frac{1}{2}})^T (\hat{\theta} - \theta^0) \right\} \xrightarrow{n \rightarrow \infty} N_m(0_m, I_m).$$

3. КОМПЬЮТЕРНОЕ МОДЕЛИРОВАНИЕ

Численные эксперименты будем проводить для простой линейной регрессии ($m=2$, $N=1$):

$$Y_t = \theta_1^0 + \theta_2^0 X_t + \xi_t, \quad t = 1, \dots, n.$$

Оценки максимального правдоподобия находятся градиентным методом [7]. По методу Монте-Карло для каждого объема выборки n проводим $Q=100$ экспериментов и вычисляем статистики:

$$\bar{V} = \frac{1}{Q} \sum_{q=1}^Q \sqrt{(\hat{\theta}_1^q - \theta_1^0)^2 + (\hat{\theta}_2^q - \theta_2^0)^2}.$$

Компьютерное моделирование будем проводить при $\theta_1^0 = 2$, $\theta_2^0 = 4$,

$\sigma^2=1$, $K=4$, $a_1=15$, $a_2=20$, $a_3=25$. $\{X_t\}_{t=1}^n$ – равномерная сетка на

$[0,10]$. На рис. 1 представлен график зависимости $\bar{V}$ от n .

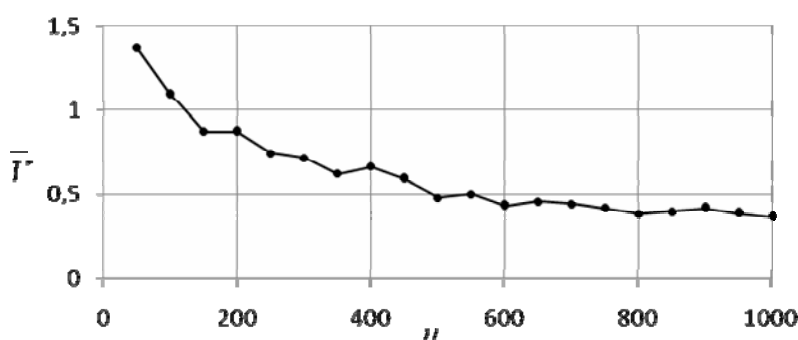


Рис. 1. График зависимости $\bar{V}$ от n

ЗАКЛЮЧЕНИЕ

В данной работе рассмотрена регрессионная модель, в которой зависимые данные наблюдаются не полностью: вместо точных значений известны только номера классов, в которые они попадают. Для нахождения параметров модели предлагаются оценки максимального правдоподобия. Найденны условия сильной состоятельности ОМП $(\hat{\theta}, \hat{\sigma}^2)$ и асимптотической нормальности ОМП $\hat{\theta}$.

Литература

1. Bai Z., Zheng, S., Zhan, B., Hu, Z. Statistical Analysis for Rounded Data // J. Statist. Plann. Inference. 2009. 139, no. 8, 2526–2542.

2. *Dempster A.P., Rubin, D.B.* Rounding error in regression: the appropriateness of Sheppard corrections // J. Roy. Statist. Soc. Ser. B. 1983. 45, 51–59.
3. *Sen Roy, S., Guriab S.* Estimation of regression parameters in the presence of outliers in response // Statistics. 2009. 43, no. 6, 531–539.
4. *Sheppard W. F.* On the calculation of the most probable values of frequency constants for data arranged according to equidistant divisions of a scale. // Proc. London Math. Soc. 1898. 29, 353–380.
5. *Vardeman S. B., Lee C. S.* Likelihood-based statistical estimation from quantization data. IEEE Trans on Instru. Measure. 2005. 54, 409–414.
6. *Харин Ю.С.* Математическая и прикладная статистика / Ю.С. Харин, Е.Е. Жук. Мн.: БГУ, 2005.
7. *Калитин Н.Н.* Численные методы. М.: Наука, 1978.

