СТАТИСТИЧЕСКОЕ ОЦЕНИВАНИЕ ПАРАМЕТРОВ МНОЖЕСТВЕННОЙ РЕГРЕССИИ ПРИ НАЛИЧИИ КЛАССИФИКАЦИИ НАБЛЮДЕНИЙ

Е. С. Агеева, Ю. С. Харин

НИИ прикладных проблем
математики и информатики
Минск, Беларусь
E-mail: helenaageeva@yahoo.ca

В данной работе рассмотрена модель множественной регрессии, в которой зависимые данные наблюдаются не полностью: вместо точных значений известны только номера классов, в которые они попадают. Для статистического оценивания параметров модели построена оценка максимального правдоподобия и исследованы ее состоятельность и асимптотическая нормальность.

Ключевые слова: регрессия, классификация наблюдений, оценки максимального правдоподобия.

ВВЕДЕНИЕ

В статистике часто встречается регрессионная модель. Ею описываются многие процессы в технике, экономике, медицине и т. д. Хорошо изучены случаи, когда зависимые данные наблюдаются с выбросами [12] или не наблюдаются вовсе [10], а также случай, когда зависимые данные являются цензурированными [3]. 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 [5]. В литературе рассматривается случай так называемых «округленных 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, распределенных по нормальному закону [1, 4, 6]. Встречаются «смешанные» модели, где разным компонентам соответствуют разные уровни округления [7]. Модель, рассмотренная в работе, является обобщением rounded data в регрессии.

МАТЕМАТИЧЕСКАЯ МОДЕЛЬ

Рассмотрим модель нелинейной множественной регрессии:

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

где n объем выборки; $\theta^0 = (\theta_1^0, ..., \theta_m^0)^T \in \Theta \subseteq R^m$ – неизвестный вектор параметров; $X_t = (X_t^1, ..., X_t^N)^T \in X \subseteq R^N$ – 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 $0 < \sigma^2 < \infty$; $F(\cdot): X \times \Theta \to R^1$ – функция регрессии.

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

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

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

Эта система борелевских множеств задает классификацию Y_t:

$$Y_t \text{ относится к классу } \Omega_{\nu_t}, \text{ если } Y_t \in A_{\nu_t},\ \nu_t \in \{1, ..., K\}. \tag{2}$$

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

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

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

ОЦЕНКИ МАКСИМАЛЬНОГО ПРАВДОПОДОБИЯ

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

$$P_{X_t,\theta,\sigma^2}\{Y_t \in A_k\} = P_{X_t,\theta,\sigma^2}\{\nu_t \in k\} = P_{X_t}(\nu_t;\, \theta, \sigma^2) = \tfrac{1}{\sqrt{2\pi}\sigma}\int_{A_k} e^{-\frac{(z-F(X_t,\theta))^2}{2\sigma^2}} dz,\ k=1, \ldots, K.$$

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

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

Максимизируя функцию $l(\theta, \sigma^2)$ по θ и σ^2, найдем оценки максимального правдоподобия [11]:

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

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

$$l(\theta, \sigma^2) = \sum\nolimits_{t=1}^n \ln\left(\Phi\left(\tfrac{a_{\nu_t} - F(X_t;\theta)}{\sigma}\right) - \Phi\left(\tfrac{a_{\nu_t - 1} - F(X_t;\theta)}{\sigma}\right)\right).$$

Теорема 1. Пусть Є – замкнутое подмножество R^m; существует такое $\bar{\sigma}^2$, что $\bar{\sigma}^2 \le \sigma^2$; $2 < K < +\infty$. 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 $(\theta, \sigma^2) \in \Theta \times [\bar{\sigma}^2, \infty)$, $\theta \ne \theta^0$, $\sigma^2 \ne \sigma^{02}$, начиная с некоторого объема выборки $n > n_1$ для $[pn]+1$ наблюдений из $\{X_t\}_{t=1}^n$ верно

$$E_{\theta^0, \sigma^{02}}\{\ln P_{X_t}(\nu_t; \theta, \sigma^2)\} - E_{\theta^0, \sigma^{02}}\{\ln P_{X_t}(\nu_t; \theta^0, \sigma^{02})\} \le -d;$$

для любой последовательности $\{\theta^i : \theta^i \in \Theta, i \in N\}$ такой, что $|\theta^i| \xrightarrow[i \to \infty]{} \infty$, выполнено $|F(X, \theta^i)| \xrightarrow[i \to \infty]{} \infty$, $X \in X \subseteq R^N$; 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 также пусть в точке (θ^0, σ^{0^2}) информационная матрица Фишера невырожденная и

$$\lim_{n \to \infty} |\tfrac{1}{n} I_n(\theta^0, \sigma^{0^2})| \ne 0.$$

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

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

КОМПЬЮТЕРНОЕ МОДЕЛИРОВАНИЕ

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

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

Оценки максимального правдоподобия находятся градиентным методом [9]. 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 рисунке 1 представлен график зависимости $\bar{V}$ от n.

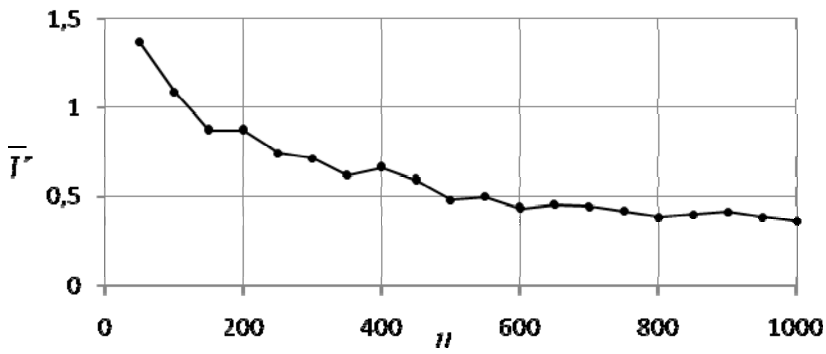

Рис. 1. График зависимости $\overline{V}$ от *n*

Ряд численных экспериментов был также проведен для случая нелинейной регрессии, в частности функции Кобба – Дугласса [8]:

$$Y_t = \theta_1^0 (X_t^1)^{\theta_2^0} (X_t^2)^{\theta_3^0} + \xi_t, \ t = 1, ..., n.$$

Оценки максимального правдоподобия находятся градиентным методом [9]. По методу Монте-Карло для каждого объема выборки n проводим Q = 10 экспериментов и вычисляем статистики:

$$\overline{V} = \tfrac{1}{Q} \sum_{q=1}^{Q} \sqrt{(\hat{\theta}_1^q - \theta_1^0)^2 + (\hat{\theta}_2^q - \theta_2^0)^2 + (\hat{\theta}_3^q - \theta_3^0)^2}.$$

Компьютерное моделирование будем проводить при $\theta_1^0 = 1$, $\theta_2^0 = 3$, $\theta_3^0 = 4$, σ² = 1, *K* = 4, $a_1 = 10$, $a_2 = 40$, $a_3 = 60$. $\{X_t\}_{t=1}^n$ – равномерная сетка на [0,2]×[0,2]. На рисунке 2 представлен график зависимости $\overline{V}$ от *n*.

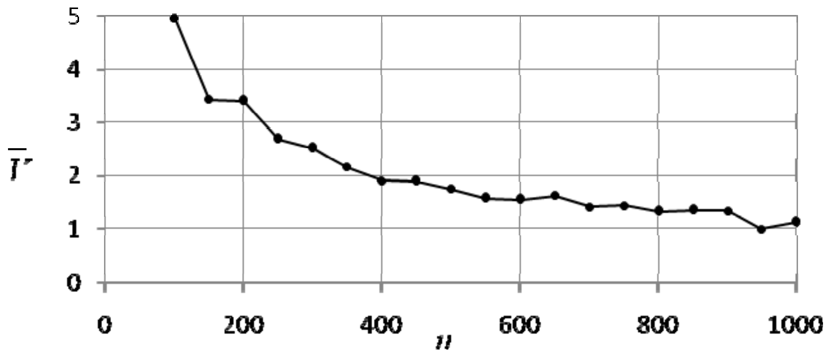

Рис. 2. График зависимости $\overline{V}$ от *n*

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. Найдены условия сильной состоятельности ОМП $(\hat{\theta}, \hat{\sigma}^2)$ и асимптотической нормальности ОМП $\hat{\theta}$. Результаты компьютерного моделирования иллюстрируют теоретические выкладки.

ЛИТЕРАТУРА

1. *Bai, Z.* Statistical Analysis for Rounded Data / Z. Bai, S. Zheng, B. Zhang, Z. Hu // J. Statist. Plann. Inferense. 2009. 139, № 8, 2526–2542.

2. *Dempster, A. P.* Rounding error in regression: the appropriateness of Sheppard corrections / A. P. Dempster, D. B. Rubin // J. Roy. Statist. Soc. Ser. B. 1983. 45, 51–59.

3. *Nelson, W.* Linear estimation of a regression relationship from censored data (part I) / W. Nelson, G. J. Hahn // Technometrics. 1972. Vol. 14. 247–269.

4. *Sen Roy, S.* Estimation of regression parameters in the presence of outliers in response / S. Sen Roy, S. Guriab // Statistics. 2009. 43, № 6. P 531–539.

5. *Sheppard, W. F.* On the calculation of the most probable values of frequency constants for data arranged according to equidistant divisions of a scale / W. F. Sheppard // Proc. London Math. Soc. 1898. 29, 353–380.

6. *Vardeman, S. B.* Likelihood-based statistical estimation from quantization data / S. B. Vardeman, C. S. Lee // IEEE Trans on Instru. Measure. 2005. 54, 409–414.

7. *Wright, D. E.* A mixture model for rounded data / D. E. Wright, I. Bray // The Statistician. 2003. 52, Part I, 3–13.

8. *Бородич, С. А.* Эконометрика / С. А. Бородич. Минск: Новое знание, 2001. 408 с.

9. *Калитин Н. Н.* Численные методы / Н. Н. Калитин. М.: Наука, 1978. 512 с.

10. *Литтл, Р. Дж. А.* Статистический анализ данных с пропусками / Р. Дж. А. Литтл, Д. Б. Рубин. М.: Финансы и статистика, 1990. 336 с.

11. *Харин, Ю. С.* Математическая и прикладная статистика / Ю. С. Харин, Е. Е. Жук. Минск: БГУ, 2005. 279 с.

12. *Хьюбер, П.* Робастность в статистике / П. Хьюбер. М.: Мир, 1984. 304 с.