О СОСТОЯТЕЛЬНОСТИ ОМП ПАРАМЕТРОВ РЕГРЕССИИ ПРИ НАЛИЧИИ КЛАССИФИКАЦИИ НАБЛЮДЕНИЙ

Е. С. Агеева

ВВЕДЕНИЕ

В математической статистике и ее приложениях широко используется регрессионная модель. Хорошо исследованы случаи, когда зависимые переменные наблюдаются с выбросами или с пропусками, а также случаи цензурированных наблюдений; при этом построены робастные статистические выводы [4, 7, 8, 9]. В статье рассматривается ситуация, когда для нелинейной множественной регрессионной модели вместо истинных значений зависимой переменной наблюдаются номера классов (интервалов), в которые попадают эти значения. Эта модель является обобщением известной модели с округленными данными [1, 2, 5].

ПОСТАНОВКА ЗАДАЧИ

Пусть на вероятностном пространстве $(\Omega, \mathcal{F}, \mathbf{P})$ определена модель нелинейной множественной регрессии с N регрессорами:

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

где n объём выборки; $\theta^0\in\Theta\subseteq\mathbb{R}^m$ – неизвестный истинный вектор параметров функции регрессии; $X_t\in\mathbf{X}\subseteq\mathbb{R}^m$ – наблюдаемый неслучайный вектор регрессоров; $Y_t\in\mathbb{R}^1$ – зависимая переменная; $\xi_t\in\mathbb{R}^1$ – ненаблюдаемая случайная величина ошибок с нормальным распределением вероятностей $\mathcal{N}(0,(\sigma^0)^2)$. Предполагается, что $\{\xi_t\}_{t=1}^n$ независимы в совокупности.

Пусть задана последовательность K непересекающихся интервалов:

$$A_k=(a_{k-1},a_k],\ k\in\mathbf{K}=\{1,\ldots,K\}, \qquad (2)$$

где $-\infty=a_0<a_1<\ldots<a_{K-1}<a_K<\infty$ – заданный упорядоченный набор границ интервалов. Эти интервалы задают классификацию зависимой переменной Y_t: Y_t относится к классу ν_t, если $Y_t\in A_{\nu_t}$, $\nu_t\in\mathbf{K}=\{1,\ldots,K\}$.

Задача заключается в том, чтобы по классифицированным наблюдениям $\nu_1,\ldots,\nu_n$ и значениям регрессоров $X_1,\ldots,X_n$ построить статистические оценки для неизвестного вектора параметров $\alpha^0=(\theta^0,(\sigma^0)^2)'\in\mathbb{R}^{m+1}$.

Используя модельные предположения (1), (2), определим функцию

$$P_X(k;\alpha)=\Phi\left(\frac{a_k-F(X;\theta)}{\sigma}\right)-\Phi\left(\frac{a_{k-1}-F(X;\theta)}{\sigma}\right),\ k\in\mathbf{K},\ \alpha=(\theta,\sigma^2)'.$$

Тогда логарифмическая функция правдоподобия допускает представление:

$$l(\alpha)=\sum_{t=1}^{n}\ln P_{X_t}(\nu_t;\alpha)=\sum_{t=1}^{n}\ln\left(\Phi\left(\frac{a_{\nu_t}-F(X_t;\theta)}{\sigma}\right)-\Phi\left(\frac{a_{\nu_t-1}-F(X_t;\theta)}{\sigma}\right)\right).$$

Максимизируя функцию $l(\alpha)$ по α, найдем оценки максимального правдоподобия (ОМП):

$$\hat{\alpha}^n=(\hat{\theta}^n,(\hat{\sigma}^n)^2)':\ l(\hat{\alpha}^n)=\max_{\alpha} l(\alpha). \qquad (3)$$

СОСТОЯТЕЛЬНОСТЬ ОМП ПАРАМЕТРОВ МОДЕЛИ

Рассмотренная модель является моделью с независимыми, но неодинаково распределенными наблюдениями, поэтому при выполнении ряда предположений к ней применима теорема о состоятельности по вероятности, доказанная Ходли в [3]. Определим вспомогательные функции:

$$P_X(k;\alpha,\rho)=\sup_{|\alpha'-\alpha|\le\rho} P_X(k;\alpha'),\ \rho>0;\ \psi_X(k;r)=\sup_{|\alpha'|\ge r} P_X(k;\alpha'),\ k\in\mathbf{K},\ r>0.$$

$$R_X(\nu;\alpha^0,\alpha)=\ln\frac{P_X(\nu;\alpha)}{P_X(\nu;\alpha^0)},\quad R_X(\nu;\alpha^0,\alpha,\rho)=\ln\frac{P_X(\nu;\alpha,\rho)}{P_X(\nu;\alpha^0)},\quad V_X(\nu;\alpha^0,r)=\ln\frac{\psi_X(\nu;r)}{P_X(\nu;\alpha^0)}.$$

$$\zeta^{(B)}=\begin{cases}\zeta, & \text{если } \zeta\ge -B;\\ -B, & \text{если } \zeta<-B,\end{cases}\quad B\ge 0.$$

Теорема 1. Пусть выполнены следующие условия:

C1. $\alpha\in\Omega$, где Ω – замкнутое подмножество $\mathbb{R}^{m+1}$.

C2. $P_{X_t}(\nu_t;\alpha)$ – функция, полунепрерывная сверху по α, равномерно по t.

C3. Существуют $\rho^*=\rho^*(\theta)>0$ и $r>0$ такие, что для некоторых $\delta>0$ и $M>0$ сразу для всех t, t=1,…,n, выполнено

$$\mathbf{E}_{X_t,\alpha^0}\{R^{(0)}_{X_t}(\nu_t;\alpha^0,\alpha,\rho)\}^{1+\delta}\le M,\ 0\le\rho\le\rho^*;\ \mathbf{E}_{X_t,\alpha^0}\{V^{(0)}_{X_t}(\nu_t;\alpha^0,r)\}^{1+\delta}\le M.$$

C4. Существует B>0, для которого

$$\overline{\lim_{n\to\infty}}\left(\tfrac{1}{n}\sum_{t=1}^{n}\mathbf{E}_{X_t,\alpha^0}\{R_{X_t}(\nu_t;\alpha^0,\alpha,\rho)\}\right)^{(B)}<0,\ \alpha\ne\alpha^0;$$

$$\overline{\lim_{n\to\infty}}\left(\tfrac{1}{n}\sum_{t=1}^{n}\mathbf{E}_{X_t,\alpha^0}\{V_{X_t}(\nu_t;\alpha^0,r)\}\right)^{(B)}<0.$$

C5. $R_{X_t}(\nu_t;\alpha^0,\alpha,\rho)$, $V_{X_t}(\nu_t;\alpha^0,r)$ – *измеримые функции от* ν_t.

Тогда ОМП $\hat{\alpha}^n$ *состоятельна по вероятности.*

Следующая теорема, предложенная мной, накладывает условия сильной состоятельности ОМП.

Теорема 2. Пусть выполнено условие С1, а также справедливы условия:

У1. Существуют ρ>0 и r>0 такие, что для некоторого M>0 сразу для всех t, t=1,…,n, выполнено

$$\mathbf{D}_{X_t,\alpha^0}\{R_{X_t}(\nu_t;\alpha^0,\alpha,\rho)\}\le M;\ \mathbf{D}_{X_t,\alpha^0}\{V_{X_t}(\nu_t;\alpha^0,r)\}^{1+\delta}\le M.$$

У2. Для ρ>0 и r>0 из У1 выполнено

$$\overline{\lim_{n\to\infty}}\frac{1}{n}\sum_{t=1}^{n}\mathbf{E}_{X_t,\alpha^0}\{R_{X_t}(\nu_t;\alpha^0,\alpha,\rho)\}<0,\ \alpha\neq\alpha^0;$$

$$\overline{\lim_{n\to\infty}}\frac{1}{n}\sum_{t=1}^{n}\mathbf{E}_{X_t,\alpha^0}\{V_{X_t}(\nu_t;\alpha^0,r)\}<0.$$

Тогда ОМП $\hat{\alpha}^n$ сильно состоятельна.

Применение теорем в данном виде на практике не представляется удобным в силу сложности вида условий. Поэтому в следующей теореме предложен ряд достаточных условий, проще проверяемых на практике.

Теорема 3. Пусть выполнены следующие условия:

У1*. Число классов ограничено: 2<K<+∞; Θ – замкнутое ограниченное подмножество $\mathbb{R}^m$; известно такое $\overline{\sigma}^2>0$, что $\overline{\sigma}^2\le(\sigma^0)^2$; $\mathbf{X}\subseteq\mathbb{R}^N$ – компакт.

У2*. Функция F(X;θ) непрерывна на $\mathbf{X}\times\Theta$ и для любого фиксированного значения θ∈Θ ограничена на $\mathbf{X}\subseteq\mathbb{R}^N$.

У3*. Для любого ε>0 существует δ=δ(ε)>0 такое, что для любых α∈Ω^ε, где $\Omega^\varepsilon=\{\alpha\in\Theta\times[\overline{\sigma}^2,\infty):|\alpha-\alpha^0|\ge\varepsilon\}$, выполнено

$$\underline{\lim_{n\to\infty}}\frac{1}{n}\sum_{t=1}^{n}\mathbf{I}_{\{|F(X_t;\theta^0)-F(X_t;\theta)|\ge\delta\}}=b,\ 0<b=b(\theta,\theta^0,\delta,F(\cdot))<1.$$

У4*. Для любого R>0 существует r>0 такое, что выполнено

$$\underline{\lim_{n\to\infty}}\frac{1}{n}\sum_{t=1}^{n}\mathbf{I}_{\left\{\inf\limits_{|\theta|\ge r}|F(X_t;\theta)|\ge R\right\}}=q,\ 0<q<q(R,F(\cdot))<1.$$

Тогда ОМП $\hat{\alpha}^n$ сильно состоятельна.

РЕЗУЛЬТАТЫ КОМПЬЮТЕРНОГО МОДЕЛИРОВАНИЯ

Для компьютерного моделирования возьмем модель:

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

где $\theta^0=(1,3,4)'$; $(\sigma^0)^2=9$; $K=4$; $a_1=10$; $a_2=40$; $a_3=60$. $\{X_t\}_{t=1}^n$ представляют собой узлы равномерной сетки на $[0,2]\times[0,2]$. Для нахождения ОМП использовался метод градиентного спуска решения экстремальной задачи (3) [6]. По методу Монте-Карло для каждого значения n проводилось $Q=100$ экспериментов и вычислялась статистика:

$$V_{\alpha^0}^n = \frac{1}{Q}\sum_{q=1}^{Q}\left((\hat{\theta}_1^{\,n,q}-\theta_1^0)^2+(\hat{\theta}_2^{\,n,q}-\theta_2^0)^2+(\hat{\theta}_3^{\,n,q}-\theta_3^0)^2+(\hat{\sigma}^{n,q\,2}-\sigma^{0\,2})^2\right).$$

На рисунке 1 представлен график зависимости статистики $V_{\alpha^0}^n$ от объема выборки n, иллюстрирующий состоятельность оценки $\hat{\alpha}^n$.

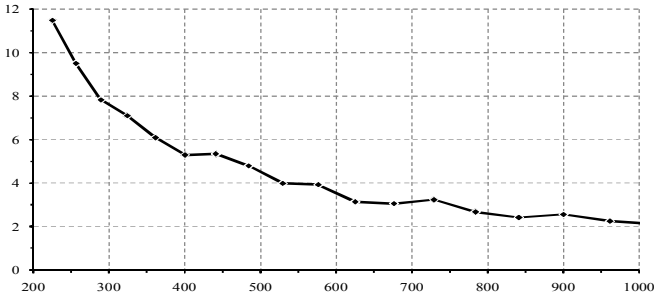

Рис. 1 График зависимости $V_{\alpha^0}^n$ от n

ЗАКЛЮЧЕНИЕ

В статье исследована модель множественной нелинейной регрессии (1), в которой зависимые данные наблюдаются косвенно: вместо их точных значений известны только номера классов, в которые они попадают. Найдены условия состоятельности по вероятности и сильной состоятельности ОМП параметров функции регрессии.

Литература

1. *Bai Z., Zhen S., Zhang B. H, Z.* Statistical Analysis for Rounded Data. J. Statist. Plann. Inferense, 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. *Hoadley B.* Asymptotic properties of the maximum likelihood estimators for the independent not identically distributed case. Ann. Math. Statist., 1971, Vol. 42, no. 4, 1977–1991.
4. *Nelson W., Hahn G. J.* Linear estimation of a regression relationship from censored data (part I). Technometrics, 1972, Vol. 14, 247–269.
5. *Sen Roy S., Guriab S.* Estimation of regression parameters in the presence of outliers in response. Statistics, 2009, 43, no. 6, 531–539.
6. *Калитин Н. Н.* Численные методы. М.: Наука, 1978.

7. *Литтл Р. Дж. А, Рубин Д. Б.* Статистический анализ данных с пропусками // М.: Финансы и статистика. 1990.
8. *Харин Ю. С.* Оптимальность и робастность в статистическом прогнозировании // Мн.: БГУ. 2008.
9. *Хьюбер Дж. П.* Робастность в статистике. М.: Мир. 1984.

СИСТЕМА ОБЪЕКТНО-РЕЛЯЦИОННОГО ОТОБРАЖЕНИЯ

В. К. Агекян

ВВЕДЕНИЕ

Большинство современных языков программирования используют объектно-ориентированный подход. В программах, использующих доступ к базам данных (БД), у программистов чаще всего возникает необходимость создавать классы, объекты которых соответствуют записям в соответствующей классу таблице БД. При этом все изменения данных, введенных конечным пользователем в прикладной программе, записываются сначала в эти объекты, а затем эти изменения синхронизируются с БД [1]. То же самое удобно делать при чтении информации из БД: сначала инициализировать объекты, наполняя их содержимым из таблиц БД, а затем манипулировать свойствами этих объектов. Набор действий, необходимый для синхронизации объектов в памяти и записей инвариантен относительно самих классов, поэтому целесообразно создать набор стандартных процедур, позволяющих записывать данные из таблиц БД в объекты программы и синхронизировать изменения объектов с БД. Эту задачу решают системы объектно-реляционного отображения.

Задачей данной работы является реализация собственной системы объектно-реляционного отображения на языке программирования Java, предоставляющий следующий функционал для работы с БД:

- возможность отображения Java классов на соответствующие сущности БД;
- возможность выполнения основных операций по добавлению и извлечению данных из БД;
- реализация объектно-ориентированного метода генерации динамических запросов;
- поддержка механизма транзакций.

Областью применения разработанной системы является разработка приложений, написанных на языке программирования Java, выполняющие работу с реляционными БД.