ПРОВЕРКА ПРОСТОЙ НУЛЕВОЙ ГИПОТЕЗЫ И СЛОЖНОЙ АЛЬТЕРНАТИВНОЙ ГИПОТЕЗЫ ДЛЯ МНОЖЕСТВЕННОЙ РЕГРЕССИИ ПРИ НАЛИЧИИ КЛАССИФИКАЦИИ НАБЛЮДЕНИЙ

Е. С. Агеева

Белорусский государственный университет
Минск, Беларусь
E-mail: helenaageeva@gmail.com

Рассмотрена модель множественной регрессии, в которой зависимые данные наблюдаются с искажениями: вместо точных значений известен только интервал, в который они попадают. Проанализирована простая нулевая и сложная альтернативная гипотезы о значении параметров модели. Для их проверки предложены два статистических теста.

Ключевые слова: регрессия, классификация, проверка гипотез.

Введение

Регрессионная модель широко применяется. Но на практике чаще встречаются различные отклонения от нее: регрессионные модели с выбросами [8, 9]; пропусками [7]; округлением [1]; цензурированием [3, 4]; группированием [2]. Для таких моделей необходимо строить эффективные статистические выводы.

Рассмотрим регрессионную модель при наличии классификации зависимой переменной: вместо точного значения мы наблюдаем лишь интервал, которому это значение принадлежит. Состоятельность оценки максимального правдоподобия (ОМП) для параметров регрессионной модели при наличии классификации наблюдений изучена в [10]. Проверим статистические гипотезы об истинном значении параметров модели.

Математическая модель

Пусть регрессионная модель имеет вид:

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

где n объем выборки; $\theta^0 = (\theta_1^0,\dots,\theta_m^0)^T \in \Theta \subseteq R^m$ – неизвестный вектор параметров; $X_t = (X,\dots,X)^T \in \mathrm{X} \subseteq R^N$ – наблюдаемый вектор регрессоров (план эксперимента $\{X_t\}_{t=1}^n$ является неслучайным); $\xi_t \in R^1$ – случайная величина ошибок с нормальной плотностью распределения вероятностей с математическим ожиданием 0 и дисперсией $0 < (\sigma^0)^2 < +\infty$ ($\{\xi_t\}_{t=1}^n$ – независимые в совокупности); $F(\cdot): \mathrm{X}\times\Theta \to R^1$ – функция регрессии. Параметром модели будем считать составной вектор $\delta^0 = (\delta_1,\dots,\delta_m,\delta_{m+1})^T = ((\theta^0)^T,(\sigma^0)^2)^T \in \Delta \subseteq R^{m+1}$.

Определена последовательность K непересекающихся интервалов ($2 \le K < +\infty$):

$$A_k = (a_{k-1}, a_k], k \in \mathrm{K} = \{1,\dots,K\}, a_0 = -\infty, a_K = +\infty. \tag{2}$$

Эта система интервалов задает классификацию Y_t: Y_t относится к классу Ω_{ν_t}, если $Y_t \in A_{\nu_t}$, $\nu_t \in \mathrm{K}$. Тогда вместо точных наблюдений $Y_1, \dots, Y_n$ наблюдаются лишь соответствующие номера классов $\nu_1, \dots, \nu_n \in \mathrm{K}$.

Задача в том, чтобы по известным интервалам $A_1, \dots, A_K$, классифицированным наблюдениям $\nu_1, \dots, \nu_n$ и значениям регрессоров $X_1, \dots, X_n$ построить статистические тесты для проверки простой нулевой гипотезы и сложной альтернативной гипотезы о значении вектора параметров δ^0.

Асимптотическая нормативность ОМП

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

$$P_{X_t}(k;\delta) = \mathbf{P}_{X_t,\delta}\{Y_t \in A_k\} = \frac{1}{\sqrt{2\pi}\sigma}\int_{A_k} e^{-\frac{(z-F(X_t,\theta))^2}{2\sigma^2}} dz = \Phi\left(\frac{a_k - F(X_t;\theta)}{\sigma}\right) - \Phi\left(\frac{a_{k-1} - F(X_t;\theta)}{\sigma}\right),$$

где $k \in \mathrm{K}$, $\delta = (\theta^T, \sigma^2)^T \in R^{m+1}$, $\Phi(\cdot)$ – функция распределения стандартного нормального закона распределения. В силу независимости $\{\nu_t\}_{t=1}^n$ логарифмическая функция правдоподобия имеет вид:

$$l(\delta) = \sum_{t=1}^n \ln P_{X_t}(\nu_t;\delta).$$

Максимизируя функцию $l(\delta)$ по δ, найдем оценки максимального правдоподобия [5]:

$$\hat{\delta}: \quad l(\hat{\delta}) = \max_{\delta} l(\delta).$$

Информационная матрица Фишера для выборки $\{\nu_t\}_{t=1}^n$ будет иметь вид $\Gamma_n(\delta) = \sum_{t=1}^n \Gamma_{n,t}(\delta)$, где $\Gamma_{n,t}(\delta) = (\gamma_{X_t}^{i,j}(\delta))_{i,j=1}^{m+1}$, $\gamma_{X_t}^{i,j}(\delta) = \mathrm{E}_{X_t,\delta}\{\frac{\partial \ln P_{X_t}(\nu_t;\delta)}{\partial \delta_i}\frac{\partial \ln P_{X_t}(\nu_t;\delta)}{\partial \delta_j}\}$.

Теорема 1. Пусть ОМП $\hat{\delta}$ является состоятельной оценкой вектора параметров δ^0; для любого фиксированного $\delta = (\theta^T, \sigma^2)^T \in R^{m+1}$ функции $F(X_t;\theta)$, $\frac{\partial F(X_t;\theta)}{\partial \theta_i}$, $\frac{\partial^2 F(X_t;\theta)}{\partial \theta_i \partial \theta_j}$, $\frac{\partial^3 F(X_t;\theta)}{\partial \theta_i \partial \theta_j \partial \theta_s}$, $i, j, s = 1, \dots, m$, ограничены на $\mathrm{X} \subseteq R^N$; $\frac{1}{n}\Gamma_n(\delta) \succ 0$, $\underline{\lim}_{n\to\infty} |\frac{1}{n}\Gamma_n(\delta)| = b > 0$. Тогда ОМП $\hat{\delta}$ асимптотически нормально распределена:

$$L\{(\Gamma_n(\delta^0))^T(\hat{\delta} - \delta^0)\} \xrightarrow[n\to\infty]{} N(0_{m+1}, I_{(m+1)\times(m+1)}).$$

Проверка простой нулевой гипотезы и сложной альтернативной гипотезы

Определим две гипотезы: простую нулевую и сложную альтернативную

$$\mathrm{H}_0: \delta^0 = \overline{\delta},$$
$$\mathrm{H}_1: \delta^0 \neq \overline{\delta}.$$

Статистика отношения правдоподобия для проверки сложных гипотез H_0, H_1 примет вид $\Lambda = \Lambda_X(H) = \dfrac{P_X(H,\overline{\delta})}{\sup_{\delta\in\Omega} P_X(H,\delta)} \in [0,1]$, где $P_X(H,\delta) = \prod_{t=1}^n P_{x_t}(\nu_t;\delta)$ [5].

Пусть $F_{\chi^2_m}$ – функция распределения вероятностей χ^2_m с m степенями свободы. Обозначим $(A^{-1})_{i,j}$ – i, j элемент матрицы A^{-1}.

Теорема 2. Пусть выполнены условия теоремы 1. Тогда для любого наперед заданного ε, $\varepsilon \in [0, 1]$, существует $c^* = F_{\chi^2_m}{}^{-1}(1-\varepsilon)$, такое, что предел при $n \to \infty$ размера решающего правила

$$d^* = d^*_X(H) = \begin{cases} 0, & -2\ln \Lambda_X(H) < c^*; \\ 1, & -2\ln \Lambda_X(H) \geq c^* \end{cases}$$

не превосходит ε. Решающее правило $d^*_X(H)$ обладает максимальной мощностью среди всех решающих правил $\tilde{d} = \tilde{d}_X(H)$, для которых $\mathbf{P}_{X,\bar{\gamma}}\{\tilde{d} = 1\} \leq \mathbf{P}_{X,\bar{\gamma}}\{d^* = 1\}$.

Теорема 3. Пусть выполнены условия теоремы 1. Тогда для любого заданного ε, $\varepsilon \in [0,1]$, существует $c^* = \max_{i=1,\dots,n} \sqrt{((\Gamma_n)^{-1})_{i,j}}\, \Phi^{-1}(1-\frac{\varepsilon}{2})$, такое, что предел при $n \to \infty$ размера решающего правила

$$d^* = d^*_X(H) = \begin{cases} 0, & \max_{i=1,\dots,n} |\hat{\delta}_i - \overline{\delta}_i| < c^*; \\ 1, & \max_{i=1,\dots,n} |\hat{\delta}_i - \overline{\delta}_i| \geq c^* \end{cases}$$

не превосходит ε. Решающее правило $d^*_X(H)$ обладает максимальной мощностью среди всех решающих правил $\tilde{d} = \tilde{d}_X(H)$, для которых $\mathbf{P}_{X,\bar{\gamma}}\{\tilde{d} = 1\} \leq \mathbf{P}_{X,\bar{\gamma}}\{d^* = 1\}$.

Компьютерное моделирование

Для компьютерного моделирования в качестве функции нелинейной регрессии использовалась производственная функция Кобба – Дугласа [6]:

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

Рассматривались гипотезы H_0, H_1:

$$H_0 : \delta^0 = (2.248, 0.404, 0.804, 1)^T,$$

$$H_1 : \delta^0 \neq (2.248, 0.404, 0.804, 1)^T.$$

При моделировании предполагались $A_1 = (-\infty, 11]$, $A_2 = (11, 16]$, $A_3 = (16, 19]$, $A_4 = (19, +\infty)$. Значения регрессоров $\{X_{t,1}, X_{t,2}\}_{t=1}^n$ представляют собой узлы равномерной сетки на $[0,10]\times[0,10]$. Проводились две серии экспериментов. В первом случае для моделирования использовался вектор параметров $\delta^0 = (2.248, 0.404, 0.804, 1)^T$, во втором – $\delta^0 = (2, 0.9, 0.5, 2)^T$. По методу Монте – Карло для каждого значения объема выборки n проводилось $Q = 100$ экспериментов и строились статистические тесты $d^*_X(H_1^q)$ и $d^*_X(H_2^q)$ по первому и второму методам для проверки гипотез H_0, H_1. Вычислялись статистики

$$\hat{\alpha} = \sum_{q=1}^{Q} d^*_X(H_1^q),\ \hat{w} = 1 - \sum_{q=1}^{Q} d^*_X(H_2^q).$$

Статистика $\hat{\alpha}$ является оценкой ошибки I рода построенного статистического теста, статистика $\hat{w}$ – оценка мощности построенного статистического теста. На рис. 1 и 2 изображены графики зависимости $\hat{\alpha}$ и $\hat{w}$ от объема выборки n соответственно.

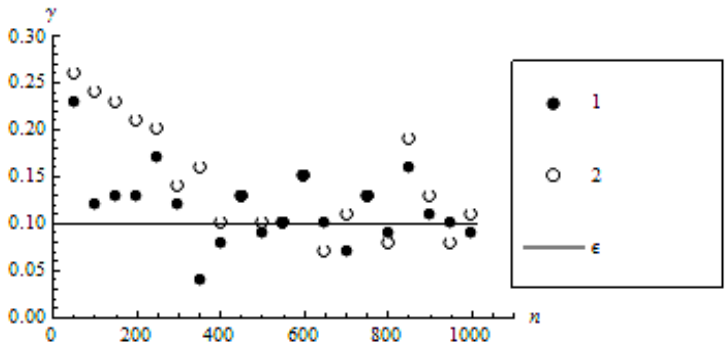

Рис. 1. График зависимости $\hat{\alpha}$ от n

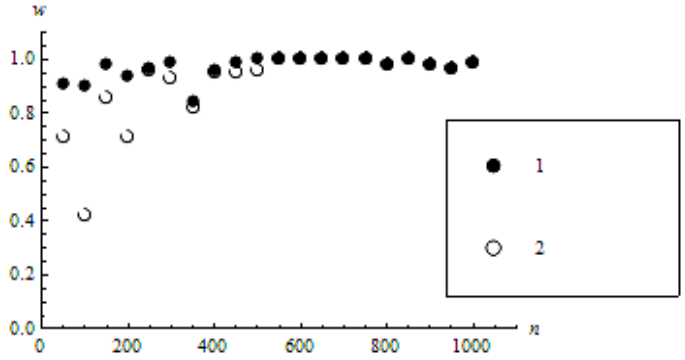

Рис. 2. График зависимости $\hat{w}$ от n

Заключение

Рассмотрена регрессионная модель, в которой зависимые данные наблюдаются неполностью: вместо точных значений известны только номера классов, в которые они попадают. Построены статистические тесты для проверки простой нулевой и сложной альтернативной гипотез для множественной регрессии при наличии классификации наблюдений. Проведены численные эксперименты, иллюстрирующие теоретические результаты.

Библиографические ссылки

1. *Dempster A. P., Rubin D. B.* Rounding error in regression: the appropriateness of Sheppard corrections // J. Roy. Statist. Soc. 1983. Ser. V. № 45. C. 51–59.
2. *Heitjan D. F.* Inference from Grouped Continuous Data: A Review // Statistical Science. 1989. V. 4. № 2. C. 164–183.
3. *Koul H., Susarla V., Ryzin V. J.* Regression analysis with randomly right-censored data // Ann. Statist. 1981. V. 9. № 6. C. 1276–1288.
4. *Nelson W., Hahn G. J.* Linear estimation of a regression relationship from censored data (part I). Technometrics. 1972. V. 14. P. 247–269.
5. *Боровков А. А.* Математическая статистика. М.: Наука, 1984.
6. *Бородич С. А.* Эконометрика. Минск: Новое знание, 2001.
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. Минск : БГУ, 2008.
9. *Хьюбер Дж. П.* Робастность в статистике. М. : Мир, 1984.
10. *Агеева Е. С., Харин Ю. С.* Состоятельность оценки максимального правдоподобия параметров множественной регрессии по классифицированным наблюдениям // Доклады НАН Беларуси. 2012. Т. 56. № 5. С. 11–19.