Доклады БГУИР
2003 *июль–сентябрь* *Том 1, № 3*

УДК 519.248

ЧУВСТВИТЕЛЬНОСТЬ ЛОГИСТИЧЕСКОЙ НОРМАЛЬНОЙ МОДЕЛИ ДАННЫХ ПРИ ИСКАЖЕНИЯХ БИНАРНЫХ НАБЛЮДЕНИЙ

М.А. ПАШКЕВИЧ

Белорусский государственный университет
пр. Ф. Скорины, 4, Минск, 220050, Беларусь

Поступила в редакцию 22 октября 2003

Предложен метод, позволяющий оценить чувствительность логистической нормальной модели группированных бинарных данных к искажениям в исходных наблюдениях. Получены количественные оценки смещений параметров модели при известных уровнях искажений. Теоретические результаты иллюстрируются компьютерным моделированием.

Ключевые слова: группированные бинарные данные, регрессия, логистическая нормальная модель, искажения, чувствительность.

Введение

При описании стохастических свойств группированных бинарных данных традиционно используется регрессионная логистическая нормальная модель (ЛНМ), которая, в отличие от классической логистической регрессии, позволяет учесть межгрупповую корреляцию. Эта модель была предложена Хеагерти [1], после чего получила широкое применение на практике: в экономике, социологии, биометрике, медицине и других областях [2–5]. Идентификация параметров ЛНМ обычно проводится методом максимального правдоподобия [6]. Однако на практике наблюдаемые данные обычно искажены, в результате чего теоретическая модель наблюдений может оказаться неадекватной [7]. В работе Нойхауса [8] на частном случае ЛНМ показано, как подобные искажения могут приводить к неверным статистическим выводам в медицинских исследованиях. Поэтому актуальна задача анализа чувствительности оценки максимального правдоподобия (МП-оценки) параметров ЛНМ [6] к искажениям в наблюдаемых данных.

В данной работе предложен метод, позволяющий оценить чувствительность логистической нормальной модели группированных бинарных данных к искажениям в исходных наблюдениях и получить количественные оценки смещений параметров модели при известных уровнях искажений. Полученные теоретические результаты иллюстрируются результатами компьютерного моделирования.

Постановка задачи

Пусть результаты наблюдений описываются набором k бинарных векторов-строк $B = (B_1, B_2, \ldots, B_k)$, $B_i \in \{0,1\}^{n_i}$, где $B_i = (B_{i1}, B_{i2}, \ldots, B_{in_i})$ — результаты серии испытаний над i-м объектом, причем $B_{ij} = 1$, если в испытании j для объекта i случайное событие A имело место, и $B_{ij} = 0$ в противном случае. Объекту номер i в испытании номер j поставлен в соответствие некоторый m-вектор факторов $Z_{ij} \in R^m$, который имеет блочный вид: $Z_{ij}^T = (Z_i^T \mid X_{ij}^T)^T$, где

вектор $Z_i \in R^{m_1}$ описывает свойства объекта, а вектор $X_{ij} \in R^{m_2}$ характеризует условия, в которых производилось испытание. При этом предполагается, что описанная модель группированных бинарных данных (ГБД) обладает следующими свойствами.

С$_1$. Размеры серий испытаний $n_1, n_2, \ldots, n_k$ малы.

С$_2$. Объекты обладают свойством “слабой неоднородности” [3].

Для описания стохастических свойств рассматриваемых данных используется логистическая нормальная модель Хеагерти, основанная на следующих предположениях [1].

П$_1$. Для i-го объекта бинарные данные B_i связаны с соответствующими векторами факторов Z_{ij} моделью логистической регрессии [11]:

$$g\left(Z_{ij}|\mu_i,\gamma_i\right)=\mu_i+Z_{ij}^T\gamma_i,\ \ j=1,2,\ldots,n_i, \tag{1}$$

где $g(Z)=\ln\left(p(Z)/(1-p(Z))\right)$ — логистическое преобразование, $p(Z)$ — вероятность успеха при факторах Z.

П$_2$. Коэффициент μ_i в модели (1) имеет вид $\mu_i = \mu + u_i$, где μ — детерминированная скалярная величина, одинаковая для всех объектов, а u_i – случайный эффект, имеющий нормальное распределение $N(0, \sigma^2)$.

П$_3$. Коэффициент γ_i в модели (1) является детерминированным и одинаковым для всех объектов, т.е. $\gamma_i = \gamma \in R^m$.

Параметрами ЛНМ являются $\mu \in R$, $\gamma \in R^m$ и $\sigma^2 \in R$, а сама модель и ее функция правдоподобия имеют следующий вид [6]:

$$g\left(Z_{ij}\mid\mu,\gamma,u_i\right)=\mu+Z_{ij}^T\gamma+u_i,\ \ i=1,2,\ldots,k,\ j=1,2,\ldots,n_i, \tag{2}$$

$$L\left(\mu,\gamma,\sigma^2\right)=\prod_{i=1}^{k}\left(\int_{-\infty}^{+\infty}\left(\prod_{j=1}^{n_i}\frac{e^{b_{ij}\left(\mu+Z_{ij}^T\gamma+u_i\right)}}{1+e^{\mu+Z_{ij}^T\gamma+u_i}}\right)\frac{1}{\sqrt{2\pi}\,\sigma}e^{-\frac{u_i^2}{2\sigma^2}}du_i\right).$$

Предположим, что данные B подвержены аддитивным стохастическим искажениям, и наблюдаются искаженные данные $\widetilde{B}$:

$$\widetilde{B}_{ij}=B_{ij}\oplus\eta_{ij},\quad i=1,2,\ldots,k,\quad j=1,2,\ldots,n_i, \tag{3}$$

где $\oplus$ — операция сложения по модулю два, а $\{\eta_{ij}\}$, $i = 1, 2, \ldots, k$, $j = 1, 2, \ldots, n_i$, — независимые случайные бинарные величины. При этом для каждого i, j имеет место следующая зависимость случайной величины η_{ij} от B_{ij}:

$$P\left\{\eta_{ij}=1\middle|B_{ij}=0\right\}=\varepsilon_0,\quad P\left\{\eta_{ij}=1\middle|B_{ij}=1\right\}=\varepsilon_1, \tag{4}$$

где ε_0, ε_1 — известные уровни искажений. Необходимо исследовать влияние искажений (3), (4) на свойства МП-оценки параметров μ, γ в случае известных значений параметра σ^2 и уровней искажений ε_0, ε_1.

Чувствительность МП-оценки к искажениям

Введем следующие обозначения: $\widetilde{Z}^T=\left(1;\,Z^T\right)^T$, $\widetilde{\gamma}^T=\left(\mu;\,\gamma^T\right)^T$, $\mu^*=\mu^0+\Delta\mu$, $\gamma^*=\gamma^0+\Delta\gamma$, $\Delta\widetilde{\gamma}^T=\left(\Delta\mu;\,\Delta\gamma^T\right)^T$, где μ^0, γ^0 — неизвестные истинные значения соответствующих параметров, μ^*, γ^* – классические МП-оценки, $\Delta\mu$, $\Delta\gamma$ — уклонения МП-оценок параметров при уровнях искажений ε_0, ε_1. Модель (2), не учитывающую искажения, будем обозначать индексом F:

$$P_{F,i}\left(b,Z,\Delta\tilde{\gamma}\right)=P\left\{b\middle|Z,i,\Delta\tilde{\gamma}\right\}=\frac{e^{b\cdot\tilde{Z}^T\tilde{\gamma}^*}}{1+e^{\tilde{Z}^T\tilde{\gamma}^*}},\quad P_{F,i}^0\left(b,Z\right)=\frac{e^{b\cdot\tilde{Z}^T\tilde{\gamma}^0}}{1+e^{\tilde{Z}^T\tilde{\gamma}^0}}.$$

Модель, учитывающую искажения уровней ε_0, ε_1, будем обозначать индексом T:

$$P_{T,i}\left(b,Z,\varepsilon_0,\varepsilon_1\right)=P\left\{b\middle|Z,i,\varepsilon_0,\varepsilon_1\right\}. \tag{5}$$

Нетрудно показать, что имеет место следующее соотношение:

$$P_{T,i}\left(1,Z,\varepsilon_0,\varepsilon_1\right)=P\left\{b=1\middle|Z,i,\varepsilon_0,\varepsilon_1\right\}=\left(1-\varepsilon_0-\varepsilon_1\right)P_{F,i}^0\left(1,Z\right)+\varepsilon_0.$$

Поскольку МП-оценки параметров ЛНМ строятся на основании модели (2), то в соответствии с результатом Уайта [12] уклонения оценок могут быть найдены как решения следующей оптимизационной задачи:

$$I\left(\Delta\tilde{\gamma}\right)=E_T\left\{ln\left(\frac{\prod_{i=1}^{k}\int_{-\infty}^{+\infty}\prod_{j=1}^{n_i}P_{T,i}\left(\tilde{b}_{ij},Z_{ij},\varepsilon_0,\varepsilon_1\right)f\left(u_i\right)du_i}{\prod_{i=1}^{k}\int_{-\infty}^{+\infty}\prod_{j=1}^{n_i}P_{F,i}\left(\tilde{b}_{ij},Z_{ij},\Delta\tilde{\gamma}\right)f\left(u_i\right)du_i}\right)\right\}\to\min_{\Delta\tilde{\gamma}}, \tag{6}$$

где $f(.)$ — плотность нормального распределения; $I(.)$ — информационный критерий Кулбака-Лейбнера, а $E_T\{.\}$ означает, что математическое ожидание берется с учетом модели с искажениями (5). В результате можно показать, что смещения МП-оценки $E\{\Delta\mu\}$, $E\{\Delta\gamma\}$ в случае искаженных данных при уровнях искажений ε_0, ε_1 определяются как

$$E\left\{\begin{pmatrix}\Delta\mu\\ \Delta\gamma\end{pmatrix}\right\}=\left(E_F^0\left\{C\left(\tilde{B}\right)\right\}\right)^{-1}\begin{pmatrix}E_{\varepsilon_0}\left\{D_0\left(\tilde{B}\right)\right\}\\ E_{\varepsilon_1}\left\{D_0\left(\tilde{B}\right)\right\}\end{pmatrix}\begin{pmatrix}\varepsilon_0\\ \varepsilon_1\end{pmatrix}+1_{m+1}\left(o\left(\varepsilon_0\right)+o\left(\varepsilon_1\right)\right), \tag{7}$$

где математические ожидания $E_F^0\left\{h\left(\tilde{B}\right)\right\}$, $E_{\varepsilon_0}\left\{h\left(\tilde{B}\right)\right\}$, $E_{\varepsilon_1}\left\{h\left(\tilde{B}\right)\right\}$ для случайной величины $h\left(\tilde{B}\right)$ вычисляются следующим образом:

$$E_F^0\left\{h\left(\tilde{B}\right)\right\}=\sum_{B\in\mathrm{B}}\left(h\left(\tilde{B}\right)\prod_{i=1}^{k}\prod_{j=1}^{n_i}\pi_{0,ij}^{\tilde{b}_{ij}}\left(1-\pi_{0,ij}\right)^{1-\tilde{b}_{ij}}\right),\quad \pi_{0,ij}=\left(1+e^{-\left(\tilde{Z}_{ij}^T\tilde{\gamma}^0\right)}\right)^{-1},$$

$$E_{\varepsilon_0}\left\{h\left(\tilde{B}\right)\right\}=\sum_{B\in\mathrm{B}}\left(h\left(\tilde{B}\right)\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(\prod_{\substack{t=1\\ t\neq i}}^{k}\prod_{\substack{l=1\\ l\neq j}}^{n_i}\pi_{0,tl}^{\tilde{b}_{tl}}\left(1-\pi_{0,tl}\right)^{1-\tilde{b}_{tl}}\right)s\left(\tilde{b}_{ij}\right)\left(1-\pi_{0,ij}\right)\right),$$

$$E_{\varepsilon_0}\left\{h\left(\tilde{B}\right)\right\}=-\sum_{B\in\mathrm{B}}\left(h\left(\tilde{B}\right)\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(\prod_{\substack{t=1\\ t\neq i}}^{k}\prod_{\substack{l=1\\ l\neq j}}^{n_i}\pi_{0,tl}^{\tilde{b}_{tl}}\left(1-\pi_{0,tl}\right)^{1-\tilde{b}_{tl}}\right)s\left(\tilde{b}_{ij}\right)\pi_{0,ij}\right),\quad s\left(b\right)=\begin{cases}1, b=1\\ -1, b=0\end{cases},$$

где B — множество всех возможных ГБД соответствующей размерности, а случайные величины $C\left(\tilde{B}\right)$, $D_0\left(\tilde{B}\right)$ имеют следующий вид:

$$C(\tilde{B}) = \sum_{i=1}^{k} \left(\frac{c_{6,i}(\tilde{b}_{ij}, Z_{ij}) - c_{5,i}(\tilde{b}_{ij}, Z_{ij}) \int_{-\infty}^{+\infty} \left(\prod_{j=1}^{n_i} P_{F,i}^{0}(\tilde{b}_{ij}, Z_{ij}) \sum_{j=1}^{n_i} \frac{c_{1,i}(\tilde{b}_{ij}, Z_{ij})}{P_{F,i}^{0}(\tilde{b}_{ij}, Z_{ij})} f(u_i) du_i \right)}{\int_{-\infty}^{+\infty} \left(\prod_{j=1}^{n_i} P_{F,i}^{0}(\tilde{b}_{ij}, Z_{ij}) f(u_i) du_i \right)} \right),$$

$$c_{6,i} = \int_{-\infty}^{+\infty} c_{3,i} f(u_i) du_i, \quad c_{5,i} = \int_{-\infty}^{+\infty} c_{4,i} f(u_i) du_i, \quad c_{4,i} = \sum_{j=1}^{n_i} \left(\prod_{\substack{l=1 \\ l \neq j}}^{n_i} P_{F,l}^{0}(\tilde{b}_{il}, Z_{ij}) c_{1,i}^{T}(\tilde{b}_{ij}, Z_{ij}) \right),$$

$$c_{3,i} = \prod_{j=1}^{n_i} P_{F,i}^{0}(\tilde{b}_{ij}, Z_{ij}) \sum_{j=1}^{n_i} \left(\frac{c_{2,i}(\tilde{b}_{ij}, Z_{ij})}{P_{F,i}^{0}(\tilde{b}_{ij}, Z_{ij})} - \frac{c_{1,i}^{T}(\tilde{b}_{ij}, Z_{ij})}{\left(P_{F,i}^{0}(\tilde{b}_{ij}, Z_{ij})\right)^2} \right) + c_{4,i} \sum_{j=1}^{n_i} \frac{c_{1,i}(\tilde{b}_{ij}, Z_{ij})}{P_{F,i}^{0}(\tilde{b}_{ij}, Z_{ij})},$$

$$c_{1,i} = (-1)^{\tilde{b}_{ij}+1} \tilde{Z}_{ij} \frac{e^{\tilde{Z}_{ij} \cdot \tilde{\gamma}_0 + u_i}}{\left(1 + e^{\tilde{Z}_{ij} \cdot \tilde{\gamma}_0 + u_i}\right)^2}, \quad c_{2,i} = (-1)^{\tilde{b}_{ij}+1} \tilde{Z}_{ij} \tilde{Z}_{ij}^{T} \frac{e^{\tilde{Z}_{ij} \cdot \tilde{\gamma}_0 + u_i} \left(1 - e^{\tilde{Z}_{ij} \cdot \tilde{\gamma}_0 + u_i}\right)}{\left(1 + e^{\tilde{Z}_{ij} \cdot \tilde{\gamma}_0 + u_i}\right)^3},$$

$$D_0(\tilde{B}) = \sum_{i=1}^{k} \frac{\int_{-\infty}^{+\infty} \left(\prod_{j=1}^{n_i} P_{F,i}^{0}(\tilde{b}_{ij}, Z_{ij}) \sum_{j=1}^{n_i} \frac{c_{1,i}(\tilde{b}_{ij}, Z_{ij})}{P_{F,i}^{0}(\tilde{b}_{ij}, Z_{ij})} f(u_i) du_i \right)}{\int_{-\infty}^{+\infty} \left(\prod_{j=1}^{n_i} P_{F,i}^{0}(\tilde{b}_{ij}, Z_{ij}) f(u_i) du_i \right)}.$$

Результаты компьютерных экспериментов

Для оценки точности выражения (7) была проведена серия компьютерных экспериментов. В качестве номинальных параметров ЛНМ были выбраны следующие значения: $k = 1000$, $n_i = 10$, $m = 2$, $m_1 = 1$, $m_2 = 1$, $\mu^0 = 0.1$, $\gamma^0 = (0{,}2;\ 0{,}3)^T$, причем Z_i и X_{ij} строились как реализации случайной величины с нормальным распределением вероятностей $N(1, 0.1)$. В процессе эксперимента генерировали 100 случайных выборок B, подчиняющейся ЛНМ с приведенными выше параметрами. Для каждой выборки производилось искажение данных, согласно (3, 4), при этом уровни искажений совпадали ($\varepsilon_0 = \varepsilon_1 = \varepsilon$) и изменялись в пределах от 0,00 до 0,04 с шагом 0,01. Затем для каждой выборки строилась МП-оценка параметров ЛНМ, и вычислялись уклонения оценки параметров от истинных значений. При фиксированном уровне искажений по полученным экспериментальным значениям уклонений строился 95-процентный доверительный интервал. Наконец, для каждого уровня искажений вычислялись теоретические смещения МП-оценки при помощи выражения (7).

Сравнение экспериментальных доверительных интервалов с теоретическими смещениями

ε	$\Delta\mu^-$	$\Delta\mu^+$	$\Delta\mu$	$\Delta\gamma_1^-$	$\Delta\gamma_1^+$	$\Delta\gamma_1$	$\Delta\gamma_2^-$	$\Delta\gamma_2^+$	$\Delta\gamma_2$
0,00	–0,0019	0,0032	0,0000	–0,0034	0,0043	0,0000	0,0013	0,0081	0,0000
0,01	–0,0060	–0,0004	–0,0035	–0,0060	0,0001	–0,0019	–0,0067	0,0002	–0,0035
0,02	–0,0050	0,0005	–0,0070	–0,0044	0,0030	–0,0038	–0,0113	–0,0045	–0,0071
0,03	–0,0147	–0,0095	–0,0105	–0,0095	–0,0023	–0,0057	–0,0158	–0,0095	–0,0106
0,04	–0,0154	–0,0095	–0,0140	–0,0130	–0,0051	–0,0076	–0,0175	–0,0091	–0,0141

Результаты компьютерного моделирования приводятся в таблице. Из нее следует, что выражение (7) достаточно точно аппроксимирует зависимость уклонений МП-оценки от уровня искажений, а именно, попадает в 95-процентный экспериментальный доверительный интервал.

Заключение

В результате проведенных исследований установлено, что в случае искаженных бинарных наблюдений оценка максимального правдоподобия логистической нормальной модели группированных бинарных данных является существенно смещенной. При этом для количественной оценки величины смещения может быть использовано полученное в работе асимптотическое разложение (7). Проведенные компьютерные эксперименты показали высокую точность предложенного приближенного метода.

SENSITIVITY OF THE LOGISTIC NORMAL DATA MODEL IN CASE OF DISTORTIONS OF BINARY OBSERVATIONS

M.A. PASHKEVICH

Abstract

A technique for assessing the sensitivity of the logistic normal model of grouped binary data in the case of distorted observations is proposed. An estimation of the parameter bias in the case of known distortion levels is obtained. Computer simulation verified the theoretical results.

Литература

1. *Heagerty P.* // Biometrics. 1999. Vol. 55. P. 688–698.
2. *Santos Silva J.M.C., Murteria J.M.R.* // Econometric Society World Congress Contributed Papers. 2000. Paper № 1121.
3. *Agresti A., Booth J.G., Hobert J.P., and Caffo B.* // Sociological Methodology. 2000. Vol. 30. P. 27–80.
4. *Colombo R., Weina J.* // Journal of Interactive Marketing. 1999. Vol. 13. P. 2–12.
5. *Neuhaus J.M.* // Statistical Methods in Medical Research. 1992. Vol. 1. P. 249–273.
6. *Aitchison J., Shen S.M.* // Biometrika. 1980. Vol. 67. P. 261–272.
7. *Kharin Yu.* Robustness in Statistical Pattern Recognition. Kluwer Academic Publishers, Dordrecht, 1996.
8. *Neuhaus J.M.* // Biometrics. 2002. Vol. 58. P. 675–683.
9. *Copas J.B.* // Journal of Royal Statistical Society. 1988. Vol. 50B. P. 225–265.
10. *Bianco A.M., Johai V.J.* // Robust Statistics, Data Analysis, and Computer Intensive Methods. Lecture Notes in Statistics. Springer Verlag: New York. 1996. Vol. 109. P. 17–34.
11. *Hosmer D.W., Lemeshow S.* Applied logistic regression. New York. 2000.
12. *White H.* // Econometrica. 1982. Vol. 50(1). P. 1–26.