О НЕКОТОРЫХ СВОЙСТВАХ МП-ОЦЕНКИ ПАРАМЕТРОВ ЛОГИСТИЧЕСКОЙ НОРМАЛЬНОЙ МОДЕЛИ С ИСКАЖЕНИЯМИ

Пашкевич М.А.

аспирант кафедры математического моделирования и анализа данных

Белорусский государственный университет, г. Минск

Введение

Логистическая нормальная модель (ЛНМ) традиционно используется для построения регрессионной зависимости вероятности успеха от условий испытаний в случае наблюдения группированных бинарных данных. Впервые она была предложена Хеагерти, после чего получила широкое применение на практике при решении прикладных задач статистического анализа данных в экономике, социологии, биометрике и других областях. Для идентификации параметров ЛНМ обычно используют метод максимального правдоподобия.

Однако на практике гипотетическая вероятностная модель наблюдений, как правило, оказывается неадекватной в силу искажений различных типов. Результаты, полученные Нойхаусом для частного случая ЛНМ, показали, что в случае искаженных результатов наблюдений оценки параметров ЛНМ могут быть смещены. Поэтому актуальна задача исследования робастности (устойчивости) оценки максимального правдоподобия (МП-оценки) параметров ЛНМ к искажениям в наблюдаемых данных.

В данной работе исследуется смещение МП-оценки параметров ЛНМ при аддитивных стохастических искажениях. Рассматриваемая модель искажений была предложена в и активно используется при анализе робастности различных моделей бинарных данных в силу своей практической значимости . В работе получены выражения для смещения МП-оценки в случае известных уровней искажений.

Математические модели и постановка задачи

Пусть определена некоторая совокупность из k объектов и некоторое случайное событие A. Над каждым объектом i этой совокупности производится

серия из n_i испытаний. Результаты испытаний 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-ым объектом, причем $B_{ij} = 1$, если в испытании j для объекта i случайное событие A имело место (успех), 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.

C_1. Размеры серий испытаний $n_1, n_2, \ldots, n_k$ малы.

C_2. Объекты обладают свойством "слабой неоднородности".

Логистическая нормальная модель Хеагерти для ГБД основана на следующих предположениях .

$П_1$. Для i-го объекта бинарные данные B_i связаны с соответствующими векторами факторов Z_{ij} моделью логистической регрессии:

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

где $g(Z) = \ln(p(Z)/(1 - p(Z)))$ – 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. $\gamma_i = \gamma \in R^m$.

Таким образом, параметрами рассматриваемой логистической нормальной модели являются $\mu \in R$, $\gamma \in R^m$ и $\sigma^2 \in R$, а сама модель и ее функция правдоподобия имеют следующий вид :

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

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

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 величины Бернулли. 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 ЛНМ в случае известных значений параметра σ^2 и уровней искажений ε_0, ε_1.

Смещение МП-оценки в случае искаженной выборки

Введем следующие обозначения для удобства изложения:

$$\widetilde{Z}^T = \left(1;\, Z^T\right)^T, \quad \widetilde{\gamma}^T = \left(\mu;\, \gamma^T\right)^T, \quad \mu^* = \mu^0 + \Delta\mu, \quad \gamma^* = \gamma^0 + \Delta\gamma, \quad \Delta\widetilde{\gamma}^T = \left(\Delta\mu;\, \Delta\gamma^T\right)^T,$$

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.

Будем говорить, что модель (2) является “ложной” моделью, т.к. не учитывает действующие искажения; в формулах будем отражать это индексом F:

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

Модель, учитывающую искажения уровней ε_0, ε_1, будем называть "истинной" и обозначать индексом T:

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

Нетрудно показать, что имеет место следующее соотношение

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

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

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

где $f(.)$ – плотность нормального распределения, $I(.)$ – информационный критерий Кулбана-Лейбнера, а $E_T\{.\}$ означает, что математическое ожидание берется с учетом истинной модели с искажениями (5). В следующей теореме получено асимптотическое разложение для уклонения МП-оценки, позволяющее определить ее смещение и доказать потерю состоятельности в случае искажений.

Теорема. Для описанной выше ЛНМ с искажениями (3), (4) имеет место следующее асимптотическое разложение для смещений МП-оценки $E\{\Delta\mu\}$, $E\{\Delta\gamma\}$:

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

Доказательство теоремы основано на асимптотическом анализе решения задачи (6) с учетом свойств логистического распределения. ■

Результаты компьютерных экспериментов

Для оценки точности асимптотического разложения (7) была проведена серия компьютерных экспериментов. В качестве номинальных параметров ЛНМ были выбраны следующие значения: $k = 1000$, $n_i = 10$, $\forall i$, $m = 2$, $m_1 = 1$, $m_2 = 1$, $\mu^0 = 0.1$, $\gamma^0 = (0.2, 0.3)^T$, причем Z_i и X_{ij} строились как реализации случайной величины с нормальным распределением вероятностей $N(1, 0.1)$. В процессе эксперимента генерировались 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).

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

Заключение

В работе получены выражения для смещения оценки максимального правдоподобия параметров логистической нормальной модели в случае известных уровней аддитивных стохастических искажений. Полученные результаты иллюстрируются результатами компьютерного моделирования. Данные исследования были частично поддержаны грантом БГУ для молодых ученых 628/30.