УДК 519.226

Бідюк П. І., Гожий О. П., Торовець Т. А.

БАЙЄСІВСЬКЕ ОЦІНЮВАННЯ ПАРАМЕТРІВ УЗАГАЛЬНЕНИХ ЛІНІЙНИХ МОДЕЛЕЙ

Наведено теоретичні відомості стосовно застосування апарату формування ймовірнісного байєсівського висновку, а також методи оцінювання якості вхідних даних і побудованих моделей. Застосування апарату формування байєсівського висновку до аналізу узагальнених лінійних моделей дає можливість використовувати множину різноманітних методів, що ґрунтуються на використанні функціональних форм спряжених розподілів. Отримані практичні результати підтверджують ефективність застосування байєсівського висновку при побудові узагальнених лінійних моделей та відкривають перспективи практичного застосування даної методики при аналізі фінансових, технічних, біологічних та інших процесів.

Ключові слова: *байєсівський висновок, узагальнені лінійні моделі, функціональні форми спряжених розподілів, оцінювання параметрів моделі.*

Приведены теоретические сведения относительно применения аппарата формирования байєсівського вывода, а также методы оценивания качества входных данных и построенных моделей. Применение аппарата формирования байесовского вывода для анализа обобщенных линейных моделей дает возможность использовать множество разнообразных методов, которые основываются на использовании функциональных форм сопряженных распределений. Полученные практические результаты подтверждают эффективность применения байесовского вывода при построении обобщенных линейных моделей и открывают перспективы практического применения данной методики при анализе финансовых, технических, биологических и других процессов.

Ключевые слова: *байесовский вывод, обобщенные линейные модели, функциональные формы сопряженных распределений, оценивание параметров модели.*

The paper provides theoretical information relative to concept and application of the vehicle of forming the Bayesian inference, and also the methods of evaluation of input data quality and constructed models. Application of the vehicle of forming the Bayesian inference for analysis of generalized linear models gives an opportunity to use a large number of various methods that are based on the use of functional forms of conjugate distributions. The practical results received confirm efficiency of application of Bayesian inference for construction of generalized linear models, and open the prospects of practical application of this methodology for the analysis of financial, engineering and biological processes.

Key words: *Bayesian inference, generalized linear models, functional forms of conjugate distributions, model parameter estimation.*

Вступ. Популярний клас моделей для математичного опису і прогнозування розвитку процесів довільної природи на основі статистичних даних складають узагальнені лінійні моделі (УЛМ). Ці моделі належать до байєсівського типу, тобто вони надають можливість прогнозувати апостеріорні розподіли змінних, що нас цікавлять у конкретному застосуванні. Їх можна застосовувати для опису процесів різної природи у різних галузях діяльності за умови наявності належних статистичних даних. Наприклад, відомі застосування у аналізі можливих втрат при страхуванні, кредитоспроможності клієнтів банку, прогнозуванні нещасних випадків, рівня довіри до компанії і напряму розвитку досліджуваних процесів та багато інших [1; 2]. Головними перевагами УЛМ є такі: (1) регресія не обмежується даними з нормальним розподілом, а розповсюджується на весь клас експоненціальних розподілів, що дає можливість описувати, наприклад, частотні або бінарні дані; (2) УЛМ дають можливість описувати адитивний вплив пояснюючих змінних на зміни (трансформування) середнього; (3) існує можливість використання в одній моделі змінних різних типів (статистичні дані і категорійні змінні). Хоча для застосування «стандартного» варіанту УЛМ необхідно мати некорельовані вибірки незалежних змінних, також існують можливості використання навіть високо корельованих змінних. До моделей такого підкласу відносять так звані маргінальні

моделі, які ґрунтуються на узагальнених оціночних рівняннях, і узагальнені лінійні змішані моделі [3]. Першочерговою задачею, яка виникає при застосуванні УЛМ, є коректне оцінювання параметрів моделей такого типу. Незважаючи на значні досягнення у цьому напрямі, сьогодні задача залишається актуальною.

Робота присвячена дослідженню практичного застосування методу формування байєсівського висновку для оцінювання параметрів узагальнених лінійних моделей. Розділи 2 і 3 присвячено теоретичному обґрунтуванню застосування методики байєсівського висновку для дослідження параметрів лінійної та логістичної регресії. У розділі 4 наведено результати експерименту, виконаного на реальних статистичних даних, що описують функціонування системи видачі кредитів. Окрім вивчення ефективності теоретичної методики, здійснюється дослідження інструментарію статистичного програмного забезпечення R і пропонуються шляхи удосконалення його окремих функцій.

Узагальнені лінійні моделі та їх окремі випадки

Нормальна лінійна регресія

Для розв'язання задачі оцінювання параметрів звичайної множинної регресії необхідно виразити відхилення залежної змінної від середнього y через k незалежних змінних $x_1,\ldots,x_k$. Середнє значення залежної змінної y_i для i-го спостереження описується так:

$$E(y_i \mid \beta, X) = x_i \beta,$$

де $x_i = (x_{i1},\ldots,x_{ik})$ вектор-рядок незалежних змінних для i-го спостереження, $\beta = (\beta_1,\ldots,\beta_k)$ – вектор-стовпчик коефіцієнтів регресії; $\{y_i\}$ – умовно незалежні від значень параметрів і змінних. При оцінюванні звичайної лінійної регресії припускається що відхилення однакові, де $\mathrm{var}(y_i \mid \theta, X) = \sigma^2$. Вектор невідомих параметрів можна представити так: $\theta = (\beta_1, \beta_2, ..., \beta_k, \sigma^2)$. За припущення, що похибки $\varepsilon_i = y_i - E(y_i \mid \beta, X)$ незалежні і нормально розподілені із нульовим середнім значенням і дисперсією σ^2 у матричній нотації модель матиме вигляд:

$$y \mid \beta, \sigma^2, X \sim N_n(X\beta, \sigma^2 I),$$

де $N_k(\mu, A)$ – нормальний багатовимірний розподіл розмірності k з вектором середніх μ та коваріаційною матрицею A. Для завершення байєсівського формулювання моделі припустимо що параметри (β, σ^2) мають типовий неінформативний апріорний розподіл [4; 5]:

$$g(\beta, \sigma^2) \propto \frac{1}{\sigma^2}.$$

Спряжений розподіл параметрів (β, σ^2) можна представити у вигляді добутку:

$$g(\beta, \sigma^2 \mid y) = g(\beta \mid \sigma^2, y)\, g(\sigma^2 \mid y).$$

Тобто у правій частині апостеріорний розподіл вектора параметрів регресії β, умовний по дисперсії похибки σ^2; $g(\beta \mid y, \sigma^2)$ – нормальний багатовимірний розподіл із середнім $\hat{\beta}$ та коваріаційною матрицею $V_{\beta\sigma^2}$, де

$$\hat{\beta} = (\mathrm{X}^T \mathrm{X})^{-1} \mathrm{X}^T \mathrm{y},\ \mathrm{V}_\beta = (\mathrm{X}^T \mathrm{X})^{-1}.$$

Якщо позначити обернену гама-щільність (a,b) пропорційною до $y^{-a-1} \exp\{-b/y\}$, то маргінальний апостеріорний розподіл σ^2 буде зворотнім гама-розподілом із параметрами $(n-k)/2, S/2)$, де $S = (\mathrm{y} - \mathrm{X}\hat{\beta})^T (\mathrm{y} - \mathrm{X}\hat{\beta})$ – сума квадратів похибок моделі.

Припустимо, що необхідно спрогнозувати майбутнє спостереження $\tilde{y}$ на основі вектора параметрів і значень змінних x^*. Відповідно до моделі, побудованої за вибіркою, маємо, що вектор $\tilde{y}$ умовний по β та σ^2 має розподіл $N(x^*\beta, \sigma^2)$. Апостеріорна прогнозна щільність $\tilde{y}$, $p(\tilde{y} \mid y)$ може бути представлена композицією вибіркових щільностей $p(\tilde{y} \mid \beta, \sigma^2)$, усереднених по апостеріорному розподілу параметрів β та σ^2:

$$p(\tilde{y} \mid y) = \int p(\tilde{y} \mid \beta, \sigma^2)\, g(\beta, \sigma^2 \mid y)\, d\beta\, d\sigma^2.$$

Оцінювання параметрів нормальної лінійної регресії

Оцінювання параметрів моделі

Апостеріорний та прогнозний розподіли, представлені у явному вигляді, дають можливість створювати ефективні алгоритми імітаційного моделювання [6]. Для моделювання спряженого апостеріорного розподілу вектора коефіцієнтів регресії β і дисперсії похибки σ^2 потрібно виконати такі кроки:

– оцінити дисперсію похибки σ^2 із маргінальної апостеріорної щільності $g(\sigma^2 \mid y)$;

– оцінити значення β із умовної апостеріорної щільності $g(\beta \mid \sigma^2, y)$.

Програмна реалізація цих алгоритмів не становить труднощів, оскільки обидва компоненти розподілів (зворотній гамма-розподіл та багатовимірний нормальний) мають зручні функціональні форми.

Після моделювання спряженого апостеріорного розподілу можна отримати вибірку з маргінального апостеріорного розподілу будь-якої функції $h(\beta, \sigma)$. Нехай x^* – вектор-рядок окремих значень регресорів. Припустимо, що необхідно визначити середній відгук для x^*:

$$E(y \mid x^*) = x^* \beta.$$

Якщо β^* оцінено за маргінальним апостеріорним розподілом β, то $x^*\beta^*$ буде оцінено за маргінальним апостеріорним розподілом $x^*\beta$ [7].

Аналогічно для представлення апостеріорного прогнозного розподілу значень майбутнього відгуку можна запропонувати простий алгоритм моделювання. Припустимо, що $\tilde{y}$ – значення майбутнього відгуку, який відповідає вектору-рядку незалежних змінних x^*. Для моделювання окремого значення $\tilde{y}$ необхідно:

– оцінити (β,σ) за спряженим апостеріорним розподілом для заданих значень y;

– оцінити $\tilde{y}$ за вибірковою щільністю для заданих значень параметрів β і σ:

$$\tilde{y} \sim N(x^*\beta,\sigma).$$

Визначеним апостеріорним прогнозним розподілом можна скористатись для перевірки якості підгонки моделі [8].

Перевірка моделі

Припустимо, що значення $\tilde{y}_1,...,\tilde{y}_n$ оцінені за апостеріорним прогнозним розподілом для тих же незалежних змінних $x_1,..., x_n$, що були використані для моделювання даних. Для того щоб оцінити чи консистентне окреме значення відгуку y_i до підібраної моделі, необхідно порівняти місцезнаходження y_i, у гістограмі оцінених значень $\tilde{y}_i$, з відповідним значенням прогнозного розподілу. Якщо y_i знаходяться у хвості розподілу, це свідчить про те, що дане спостереження – кандидат на вилучення. Інший підхід до перевірки моделі ґрунтується на «байєсівських залишках» [9]. У традиційному регресійному аналізі оцінювання відповідності підібраної моделі здійснюється через перевірку стандартизованих залишків за виразом:

$$r_i = \frac{y_i - x_i\hat{\beta}}{\hat{\sigma}\sqrt{1-h_{ii}}},$$

де $\hat{\beta}$ і $\hat{\sigma}$ – звичайні оцінки вектора регресії та середньоквадратичного відхилення похибок, а h_{ii} – i-й елемент матриці оцінок.

З байєсівської точки зору можна розглядати розподіл параметричних залишків моделі:

$$\{\varepsilon_i = y_i - x_i\beta\}.$$

Перед отриманням даних спостережень припускається, що параметричні залишки – випадкова вибірка з розподілу $N(0,\sigma)$. Припустимо, що i-те спостереження є кандидатом на вилучення якщо $|\varepsilon_i| > k\sigma$, де k – попередньо задана константа, скажімо 2 чи 3. Попередня ймовірність того, що окреме спостереження виявиться кандидатом на вилучення дорівнює $2\Phi(-k)$, де $\Phi(z)$ – функція стандартного нормального розподілу. Після отримання даних спостережень y можна обчислити апостеріорну ймовірність того, що будь-яке спостереження буде кандидатом на вилучення. Визначимо функції z_1 та z_2 так:

$$z_1 = (k - \varepsilon_i/\sigma)/\sqrt{h_{ii}},\ z_2 = (-k - \varepsilon_i/\sigma)/\sqrt{h_{ii}},$$

де $\hat{\varepsilon}_i = y_i - x_i\hat{\beta}$.

Тоді апостеріорна ймовірність того, що i-е спостереження буде кандидатом на вилучення дорівнює:

$$p_i = P(|\varepsilon_i| > k\sigma,\ y) = \int (1 - \Phi(z_1) + \Phi(z_2))\, g(\sigma^2 \mid y)\, d\sigma^2.$$

На практиці p_i можна обчислити і порівняти з апріорною ймовірністю $2\Phi(-k)$.

Вибір моделі з використанням попереднього розподілу Зелнера

Арнольд Зелнер [10] запропонував простий спосіб додавання суб'єктивної інформації при розв'язанні задач регресійного аналізу за допомогою часткового розподілу. Цей частковий розподіл називають g-апріорним, він дає змогу легко обрати одну із множини регресійних моделей. Для g-апріорного розподілу припускається, що вектор коефіцієнтів регресії β, умовний по σ, має багатовимірний нормальний попередній розподіл із середнім β^0 та коваріаційною матрицею $c\,\sigma^2(\mathrm{X}^T\mathrm{X})^{-1}$, а для σ^2 – стандартний неінформативний попередній розподіл, пропорційний до $1/\sigma^2$. Для того щоб отримати можливість використання цього попереднього розподілу необхідно визначити дві величини: гіпотетичний вектор β^0 для вектора регресії та константу c, що відображає величину інформації у даних, яка пов'язана з апостеріорним розподілом. Якщо припущення щодо попереднього розподілу свідчить про наявність суттєвої інформації, слід обирати невелике значення c, і навпаки, вибір великого значення c матиме ефект, подібний до вибору стандартного неінформативного розподілу для (β, σ^2).

Однією з переваг g-апріорного розподілу є відносно проста функціональна форма апостеріорного розподілу. Спряжену апостеріорну щільність (β, σ^2) можна представити так:

$$g(\beta, \sigma^2 \mid y) = g(\beta \mid y, \sigma^2)\, g(\sigma^2 \mid y).$$

Апостеріорний розподіл вектора регресії β, умовний по σ^2, є багатовимірним нормальним із середнім β^1 і коваріаційною матрицею V_1, де

$$\beta^1 = \frac{c}{c+1}\left(\frac{\beta^0}{c} + \hat{\beta}\right),\ V_1 = \frac{\sigma^2 c}{c+1}\left(\mathrm{X}^T\mathrm{X}\right)^{-1}.$$

Маргінальний апостеріорний розподіл σ^2 – це зворотній гамма-розподіл (a_1, b_1), де

$$a_1 = n/2,\ b_1 = \frac{S}{2} + \frac{1}{2(c+1)}\left(\beta_0 - \hat{\beta}\right)^T \mathrm{X}^T\mathrm{X}\left(\beta_0 - \hat{\beta}\right).$$

Зелнерівський клас апріорних *g*-розподілів використовують для вибору кращої моделі регресії. Нехай існує k потенційних незалежних змінних для прогнозування значень залежної змінної y. Тоді кількість можливих моделей, що відповідають включенню або виключенню змінних з моделі становить 2^k. Позначимо через β вектор параметрів повної моделі, що включає усі незалежні змінні. Визначимо для β попередній розподіл з $\beta_0 = 0$ і «великим» значенням c, скажімо $c = 100$, що відповідає невизначеній апріорній інформації стосовно знаходження β. Тоді, якщо β^P – регресійна модель, що містить підмножину із P незалежних змінних, поставимо їй у відповідність *g*–апріорне такої ж функціональної форми.

Різні моделі порівнюються з використанням апріорної прогнозної щільності. Якщо вибіркова щільність залежної змінної задається функцією $f(y \mid \beta, \sigma^2)$, а для вектора параметрів (β, σ^2) задана апріорна щільність $g(\beta, \sigma^2)$, то апріорна прогнозна щільність визначається інтегралом:

$$m(y) = \int f(y \mid \beta, \sigma^2) g(\beta, \sigma^2) d\beta \, d\sigma^2 .$$

Якщо використати логарифмічне перетворення σ^2: $\eta = \log \sigma$, то інтеграл по (β, η) можна апроксимувати за методом Лапласа з високим рівнем точності. На практиці виникає необхідність обчислення прогнозної щільності для множини правдоподібних моделей. Припустимо, що 2^k моделей мають приблизно однакові апріорні розподіли. Тоді апостеріорна ймовірність моделі M_j задається так [10]:

$$P(M_j \mid y) = \frac{m_j(y)}{\sum_{i=1}^{2^k} m_i(y)} .$$

Приклад оцінювання параметрів з використанням фактичних даних

Для оптимального вибору параметрів моделі з використанням g-апріорного оцінювання розподілу параметрів лінійної та логістичної регресії використано фактичні дані про 6500 клієнтів компанії з видачі кредитів, зібрані протягом 12 місяців. Як незалежні змінні $x_1, \ldots, x_k$ використано дані, що характеризують стан клієнта. Як залежну змінну, у випадку моделювання з використанням лінійної регресії, використано рівень прибутку, отриманого від клієнта, а у випадку моделювання з використанням логістичної регресії – бінарна змінна, що вказує на те, чи був отриманий прибуток від клієнта. Для виконання обчислень, моделювання значень апріорних та апостеріорних розподілів та вибору параметрів моделі використано пакет Learn Bayes [11] оболонки для статистичних розрахунків R [12].

Вибір оптимальних параметрів моделі

Функція Bayes.model.selection, реалізована в пакеті LearnBayes, використовує алгоритм, описаний в розділі 3.3 для обчислення прогнозної щільності кожної моделі.

Функція має три вхідні параметри: y – вектор залежних змінних, X – матриця, що містить усі незалежні змінні, і c – значення константи c для g-prior. Функція повертає матрицю логарифмічних прогнозних щільностей усіх моделей. Функція також повертає апостеріорні ймовірності моделей, виходячи з припущення, що усі моделі мають однакову апріорну ймовірність. Для вибору оптимальної множини параметрів використано 12 незалежних змінних і перебрано 4096 різних комбінацій моделей. Нижче наведено результати для 5 кращих комбінацій незалежних змінних, які упорядковані за спаданням апостеріорних ймовірностей.

Таблиця 1

Результати перебору комбінацій незалежних змінних та відповідні значення апостеріорних ймовірностей

X1	X2	X3	X4	X5	X6	X7	X8	X9	X10	X11	X12	log, m	Prob
false	true	true	true	true	true	true	true	true	false	false	false	–51143,8	0,507
true	true	true	true	true	true	true	true	true	false	false	false	–51145,3	0,117
false	true	true	true	true	true	true	true	true	false	true	false	–51145,5	0,096
false	true	true	true	true	true	true	true	true	false	false	true	–51145,6	0,087
false	true	true	true	true	true	true	true	true	false	false	false	–51146	0,060

Таким чином, для побудови лінійної регресії обрано 8 змінних: X2 – X9. Перші чотири змінні X2 – X5 – неперервні, а решта – приймають бінарні значення. Для вибору параметрів логістичної регресії у системі *R* використовується команда lm. Змінні X2 та X3 мають найбільший вплив на результат. Маленькі значення змінної X2 та великі значення змінної X3 приводять до збільшення прибутку. Функція blinreg використовується для генерування значень із спряженого апостеріорного розподілу β та σ^2. Вхідними значеннями функції є вектор залежних змінних y, матриця плану X лінійної регресії та кількість спостережень m: > *theta. sample = blinreg* (*fity, fitx,5000*). Функція повертає дві компоненти: beta – матриця симульованих значень із маргінального апостеріорного розподілу β та *sigma* – вектор симульованих значень із маргінального апостеріорного розподілу σ. На рисунку 2 зображено гістограми значень кожного коефіцієнта регресії і дисперсії похибок, оцінені за апостеріорним розподілом.

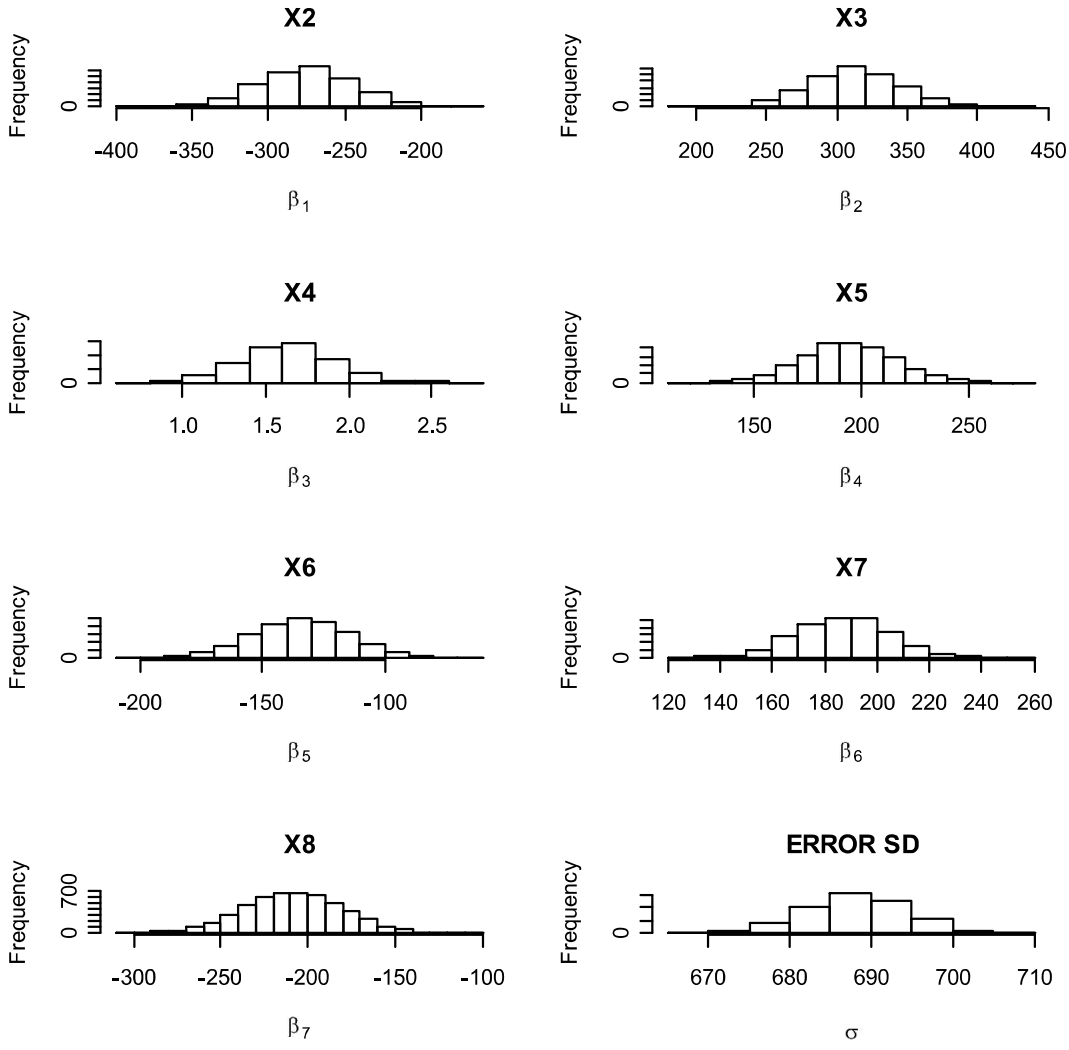

Рис. 2. Гістограми значень коефіцієнтів лінійної регресії

Отримані значення кожного з індивідуальних параметрів можна підсумувати, обчисливши 5-й, 10-й та 95-й процентилі для кожної вибірки оцінених значень:

```
>apply(theta.sample$beta,2,quantile,c(.05,.5,.95))
>quantile(theta.sample$sigma,c(.05,.5,.95)).
```

Таблиця 2

Значення процентилів оцінених значень індивідуальних параметрів

	константа	**X2**	**X3**	**X4**	**X5**	**X6**	**X7**	**X8**	**sigma**
5%	77,50	-325,06	257,757	1,176	159,990	-166,69	158,424	-254,13	678.53
50%	130,14	-275,54	311,769	1,625	193,594	-134,07	187,530	-208,33	688.31
95%	182,90	-226,58	366,174	2,054	229,007	-100,38	216,396	-163,64	698.35

Медіани апостеріорного розподілу мають значення, подібні до оцінок, отриманих при звичайному регресійному оцінюванні. Вони практично не відрізняються, оскільки було використано неінформативний апріорний розподіл для β; будь-які незначні відмінності були спричинені маленькими похибками, що мали місце при моделюванні.

Проілюструємо методи перевірки відповідності спостережень і отриманої моделі. Перший метод ґрунтується на апостеріорному прогнозному розподілі. Позначимо через y^* щільність логарифму майбутнього прибутку для вектора незалежних змінних x_i. Використовуючи функцію binregpred, можна симулювати значення апостеріорних прогнозних щільностей для логарифму кожного $y_i ... y_n$, використавши матрицю плану fit$x як аргумент. Для кожної прогнозної щільності необхідно обчислити 5-й та 95-й

процентилі. Необхідно співставити відповідність значень залежної змінної і отриманих прогнозних щільностей. Кожна точка, що знаходиться поза 90-відсотковою інтервальною смугою є потенційним кандидатом на вилучення. У вибірці є 411 точок, що перевищують значення 90-відсоткового квантиля.

```
> pred.draws = blinregpred(fit$x, theta.sample)
> pred.sum = apply(pred. draws,2, quantile, c(.05, .95))
> out = (Profit > pred.sum[2,])
> table(out)
out.
FALSE TRUE: 6019 411
```

Інший метод пошуку кандидатів на вилучення грунтується на використанні байєсівських залишків $\varepsilon_i = y_i - x_i\beta$. Ймовірності $P(|\varepsilon_i| > k \mid y)$ можуть бути обчислені за допомогою функції bayesresiduals. Вхідними параметрами є структура змодельованої логістичної регресії fit, матриця симульованих значень theta.sample і константа k. Використовуючи команду identify можна визначити спостереження, що мають ймовірність бути кандидатами на вилучення більше 0,4. У даному випадку отримаємо 331 потенційний кандидат на вилучення:

```
>prob.out = bayesresiduals (fit, theta. sample, 2)
> out2 = (prob.out > 0.4)
>table(out2)
out2
FALSE TRUE: 6099 331.
```

Аналогічний метод отримання апостеріорного розподілу коефіцієнтів регресії можна застосувати використати при оцінюванні параметрів логістичної регресії [13]. Для оцінювання параметрів логістичної регресії замість неперервних значень змінних X1-X15 використано категоріальні змінні, отримані при поділі змінних на інтервали [14; 15]. У якості залежної змінної використано бінарну змінну gbProfit, що дорівнює 1 у разі якщо значення змінної Profit більше 300, і 0 – якщо менше. Для оцінювання параметрів логістичної регресії використана стандартна функція glm пакету R для роботи з узагальненими лінійними моделями.

```
>mylogit<-
glm(gbSumOfProfit~X1+X2+X3+X4+X5+X6+X7+X8+X
9+X10+X11+X12+X13+X14+X15,
Family = binomial (link = "logit"), data=datatrain)
>summary(mylogit)
Call
glm(formula = gbSumOfProfit ~
X1+X2+X3+X4+X5+X6+X7+X8+X9+X10+X11+X12+
X13+X14+X15,
family = binomial(link = "logit"), data = datacdtrain)
```

Результат роботи програми видається у формі:

```
Deviance Residuals:
Min  1Q  Median  3Q  Max
-2,2210 -1,1344 0,6573 0,9619 2,3230
```

Значення оцінок коефіцієнтів та статистики їх якості:

```
Estimate Std. Error z-value  Pr(>|z|)
(Intercept)    -0,64741 0,17184 -3,767  0,000165
X11            0,12712  0,08405  1,512  0,130426
X12            0,21298  0,12047  1,768  0,077078
…
X15            -0,22385 0,10375  -2,158 0,030954
AIC: 526,2.
Number of Fisher Scoring iterations: 4
```

Припустимо, що вектор коефіцієнтів регресії $\beta = (\beta_0 ... \beta_{15})$ має рівномірний неінформативний апріорний розподіл. Апостеріорну моду і коваріаційну матрицю β було знайдено з використанням функції *laplace*. Як вхідні параметри функції використано створену функцію logisticpost2, оцінений вектор параметрів регресії mylogit$coef та матрицю незалежних змінних, об'єднану з вектором залежних змінних. Отриману матрицю коваріації loglaplace$var використано як вхідний параметр функції rwmetrop для генерування значень з апостеріорного розподілу β. Оцінювання квантилів розподілу свідчить про близькість медіан розподілу до значень, отриманих при прямому оцінюванні коефіцієнтів логістичної регресії. Оскільки значення параметрів регресії набували дискретних значень, то отримана щільність розподілу значно відрізняється від результатів, отриманих при оцінюванні щільності параметрів лінійної регресії.

Висновки. Застосування апарату формування байєсівського висновку до аналізу узагальнених лінійних моделей дає можливість використовувати множину різноманітних методів, що грунтуються на використанні функціональних форм спряжених розподілів. У роботі наведено теоретичні відомості щодо поняття та застосування апарату формування байєсівського висновку, а також методи оцінювання якості вхідних даних і побудованих моделей. Використання апріорного розподілу Зелнера дає можливість здійснювати вибір оптимальної множини параметрів лінійної регресії. Методику обчислення байєсівських залишків використано для пошуку кандидатів на вилучення у вхідних даних.

Зазначені вище методи застосовано для аналізу параметрів лінійної та логістичної регресії; побудовано та проаналізовано модель на основі реальних даних клієнтів кредитної організації. Здійснено підбір параметрів моделі, виконано оцінювання кандидатів на вилучення, побудована щільність розподілу коефіцієнтів регресії. Також створено додаткову реалізацію функції статистичної оболонки R для аналізу моделей з багатьма вхідними змінними. Отримані практичні результати підтверджують ефективність застосування процедури формування байєсівського ймовірнісного висновку при побудові узагальнених лінійних моделей та відкривають перспективи практичного застосування даної методики при аналізі фінансових, технічних та біологічних процесів.

Надалі планується дослідити ефективність застосування методу формування байєсівського висновку для інших видів узагальнених лінійних моделей.

ЛІТЕРАТУРА

1. Haberman S. Generalized linear models and actuarial science / S. Haberman, A. E. Renshaw // The Statistician. – 1996. – Vol. 45. – № 4. – P. 407–436.
2. Modern actuarial risk theory / [R. Kaas, M. Gooverts, J. Dhaene, M. Denuit]. – Dordrecht : Kluwer Academic Publishers. – 2002. – 238 p.
3. Dey D. K. GLM – A Bayesian perspective / D. K. Dey, S. K. Ghosh, B. K. Mallick. – New York : Marcel Dekker, Inc., 2000. – 442 p.
4. Congdon P. Applied Bayesian modeling / P. Congdon. – New York : Wiley & Sons Ltd, 2003. – 472 p.
5. Bedrick E. A new perspective on priors for generalized linear models / E. Bedrick, R. Christensen, W. Johnson // Journal of the American Statistical Association, 1996. – Vol. 91. – P. 1450–1460.
6. Rossman A. J. Workshop Statistics : Discovery with Data, a Bayesian Approach. – Emeryville, CA : Key College, 2001. – 12 p.
7. Dey D. K. (1998). Simulation based model checking for hierarchical models/ D. K. Dey, A. E. Gelfand, T. Swartz, P. K. Vlachos. – Test, 7, 1998. – P. 325–246.
8. Albert J. Bayesian computations with R. Springer / J. Albert. – 2009. – P. 205–234.
9. Chaloner K. A Bayesian Approach to Outlier Detection and Residual Analyses / K. Chaloner, R. Brant // Biometrica, 1988. – Vol. 75, Issue 4. – P. 651–659.
10. Зельнер А. Байесовские методы в эконометрии / А. Зельнер ; [пер. с английского]. – Москва : «Статистика», 1980. – 438 с.
11. Albert J. Package «LearnBayes» [Electronic resours]/ Albert J. – Acces mode : http://cran.r-project.org/web/packages/LearnBayes/LearnBayes.pdf, 2012. – 74 p.
12. Verzani J. Using R for Introductory Statistics / J. Verzani. – Boca Raton, FL : Chapman and Hall, 2004. – 148 p.
13. Koch K. R. Introduction to Bayesian Statistics / K. R. Koch. – Berlin : Springer, 2007. – 258 p.
14. Congdon P. Bayesian Models for Categorical Data / P. Congdon. – Chichester : John Wiley and Sons, 2005. – 447 p.
15. Fitting Generalized Linear Models/ R documentation. – Acces mode : http://web.njit.edu/all_topics/Prog_Lang_Docs/html/library/base/html/glm.html. – 2009. – 77 p.

Рецензенти: Мусієнко М. П., д.т.н., професор;
Батрак Ю. А., к.т.н., доцент.

Дата надходження статті до редколегії 10.05.2013 р.

ГОЖИЙ Олександр Петрович, декан факультету комп'ютерних наук, кандидат технічних наук, доцент, Чорноморський державний університет імені Петра Могили.

БІДЮК Петро Іванович, д.т.н., професор кафедри математичних методів системного аналізу, Навчально-науковий комплекс «Інститут прикладного системного аналізу», Національний технічний університет України «Київський політехнічний інститут».

ТОРОВЕЦЬ Тетяна Анатоліївна, аспірантка Інституту прикладного системного аналізу, Навчально-науковий комплекс «Інститут прикладного системного аналізу», Національний технічний університет України «Київський політехнічний інститут».