УДК 519.21

Ю.И. ПЕТУНИН, Д.А. КЛЮШИН, Е.Н. ГОЛУБЕВА, Л.А. НАЛЕСКИНА, Л.Н. КУНСКАЯ, В.Ф. ЧЕХУН

МЕТОДЫ РАСПОЗНАВАНИЯ ГРУППОВОЙ ПРИНАДЛЕЖНОСТИ, ОСНОВАННЫЕ НА ДОВЕРИТЕЛЬНЫХ ГРАНИЦАХ, И ИХ АПРОБАЦИЯ В КЛИНИЧЕСКОЙ ОНКОЛОГИИ

Ключевые слова: *метод линейной регрессии, доверительный интервал для математического ожидания, доверительный интервал для коэффициента корреляции, гомоцистеин, рак молочной железы.*

ВВЕДЕНИЕ

В практической деятельности исследователя часто возникает необходимость проанализировать и провести сопоставление двух групп объектов (людей, предметов, элементов и т. п.) по ряду изначально известных показателей. Задача усложняется, если на показателях нескольких объектов нужно установить, к какой исследуемой группе относится каждый объект. В некоторых случаях задача сводится к анализу каждой группы с последующим сравнением полученных результатов. Предположим, что некоторая совокупность показателей влияет на прогнозируемый параметр. Необходимо определить, какой из них имеет наибольшее влияние. Для решения данной задачи достаточно применить метод усеченной многомерной линейной регрессии к данным каждой группы, а затем, сравнив результаты, сделать соответствующие выводы.

Для анализа и сравнения двух групп математическая статистика предлагает множество методических подходов с уже готовыми реализованными алгоритмами, но иногда этого может быть недостаточно. В частности, для определения степени зависимости двух величин используется коэффициент корреляции, но если его значения у всех пар показателей в обеих группах невелики, например меньше 0,4, то сложно установить существование зависимости между рассматриваемыми показателями. Для анализа приведенных результатов эффективнее использовать не коэффициент корреляции, а его доверительные интервалы. Подход, описанный в настоящей статье, заключается в сравнении доверительных интервалов корреляции одних и тех же показателей в разных группах. Чем меньше длина пересечения интервалов, тем более значимой является связь между одинаковыми показателями в разных группах.

Для решения такой задачи был предложен метод распознавания, основанный на доверительных интервалах. В этом случае для каждой группы вычисляются доверительные интервалы математического ожидания конкретного показателя. Рассматриваемый элемент будет принадлежать определенной группе, если большее число значений показателей этого элемента попадает в соответствующие доверительные интервалы этой группы.

Изложим описанные подходы более детально.

ЗАДАЧИ И МЕТОДЫ ИССЛЕДОВАНИЯ

Рассмотрим следующие задачи.

1. Сравнение значимости показателей, влияющих на прогнозируемый показатель в двух группах.

2. Сравнение связей между парами показателей в двух группах.

3. Распознавание элемента в одной из двух групп по заданным значениям его показателей.

Для решения поставленных задач используем:

1) модель усеченной многомерной линейной регрессии;

2) доверительные интервалы для коэффициента корреляции;

3) метод распознавания, основанный на доверительных интервалах математического ожидания.

1. Модель усеченной многомерной линейной регрессии. Рассмотрим задачу сравнения значимости показателей, влияющих на прогнозируемый показатель, в двух группах. Пусть $X^{(1)}$ и $X^{(2)}$ — данные группы, каждая из которых характеризуется n показателями, а $y^{(1)}$, $y^{(2)}$ — прогнозируемые показатели соответствующих групп. Необходимо установить, какие показатели оказывают наибольшее влияние на прогнозируемый показатель в обеих группах, а также сравнить их значимость относительно данной группы.

Рассмотрим классическую модель многомерной линейной регрессии для k-й группы с известным (прогнозируемым) показателем

$$y^{(k)} = c_1^{(k)} x_1^{(k)} + c_2^{(k)} x_2^{(k)} + \ldots + c_n^{(k)} x_n^{(k)} + c_0^{(k)} + \xi^{(k)},\ k = 1,2, \qquad (1)$$

где $x_1^{(k)}, x_2^{(k)}, \ldots, x_n^{(k)}$ — неслучайные независимые величины, которые обычно интерпретируются как факторы; $c_1^{(k)}, c_2^{(k)}, \ldots, c_n^{(k)}, c_0^{(k)}$ — неизвестные коэффициенты линейной регрессии; $\xi^{(k)}$ — случайный шум (артефакт, искажающий истинное значение $y^{(k)}$). В дальнейшем для удобства будем опускать верхние индексы, обозначающие номер группы.

С помощью модели многомерной линейной регрессии можно определить значимость каждого из факторов $x_1, x_2, \ldots, x_n$, участвующих в формировании известных показателей y, и установить наиболее важные факторы в модели (1). Предположим, что при этом известны экспериментальные значения факторов $x_1^{(i)}, x_2^{(i)}, \ldots, x_n^{(i)}$ и соответствующее им значение $y^{(i)}$, полученные в результате i-го испытания (эксперимента) $(i = \overline{1, m})$. Число повторений эксперимента m должно значительно превышать число факторов n.

Для решения поставленной задачи используется метод наименьших квадратов (МНК) [1]. Пусть $c_1^*, \ldots, c_n^*, c_0^*$ — оценки, полученные методом наименьших квадратов (МНК-оценки) на основании экспериментальных данных $y^{(i)}, x_1^{(i)}, \ldots, x_n^{(i)}$, $i = \overline{1, m}$. Обозначим s^2 ошибку модели (1),

$$s^2 = \sum_{i=1}^{m} \left(y^{(i)} - \sum_{l=1}^{n} c_l^* x_l^{(i)} - c_0^* \right)^2,$$

которая определяет точность модели (1). Для выявления важности фактора x_j необходимо исключить его из модели (1) и рассмотреть «усеченную» модель

$$y = c_1^{(j)} x_1 + \ldots + c_{j-1}^{(j)} x_{j-1} + c_{j+1}^{(j)} x_{j+1} + \ldots + c_n^{(j)} x_n + c_0^{(j)} + \xi, \qquad (2)$$

построенную по всем остальным факторам $x_1, \ldots, x_{j-1}, x_{j+1}, \ldots, x_n$. При этом ошибка модели имеет вид

$$s_j^2 = \sum_{i=1}^{m} \Big(y^{(i)} - \sum_{l=1,\ l \neq j}^{n} c_l^{(j)*} x_l^{(i)} - c_0^{(j)*} \Big)^2,$$

где $c_l^{(j)*}$, $l \neq j$, — МНК-оценки в модели (2), вычисленные на основании экспериментальных значений $y^{(i)}$, $x_1^{(i)}, \ldots, x_{j-1}^{(i)}, x_{j+1}^{(i)}, \ldots, x_n^{(i)}$, $i = \overline{1, m}$. Легко увидеть,

что ошибка s_j^2 возрастает: $s_j^2 \geq s^2$, причем для самого важного фактора x_j это возрастание будет наибольшим: $s_j^2 \geq s_l^2 \quad \forall\, l = \overline{1, n}$. Обоснование этого эффекта очевидно: если фактор x_j является наиболее важным, то его потеря, т. е. исключение из модели (1), приведет к наибольшему увеличению ошибки модели. Таким образом, на основании ошибки s_j^2, полученной в результате исключения фактора x_j из модели (1), можно судить о значимости этого фактора для показателя y. Если из ошибок s_j^2, $j = \overline{1, n}$, образовать вариационный ряд

$$s_{j_1}^2 \leq s_{j_2}^2 \leq \ldots \leq s_{j_n}^2,$$

то наиболее важным фактором является показатель x_{j_n}, затем $x_{j_{n-1}}$ и т.д., а наименее важным фактором будет показатель x_{j_1} с наименьшей ошибкой $s_{j_1}^2$.

Таким образом, сравнивая вариационный ряд, построенный для группы $X^{(1)}$, с рядом для группы $X^{(2)}$, можно оценить значимость показателей.

2. Доверительные интервалы для коэффициента корреляции. Рассмотрим задачу сравнения связи между парами показателей в двух группах. Пусть $X^{(1)}$ и $X^{(2)}$ — группы, каждая из которых характеризуется n показателями. Необходимо установить, как связаны между собой показатели x_i и x_j, $i \neq j$, в группах $X^{(1)}$ и $X^{(2)}$.

Степень зависимости между двумя величинами характеризуется коэффициентом корреляции

$$r(x, y) = \frac{K(x, y)}{\sigma(x)\sigma(y)},$$

где $K(x, y)$ — коэффициент ковариации, $\sigma(x)$ и $\sigma(y)$ — дисперсии [2].

При статистическом анализе экспериментальных данных используется выборочный аналог коэффициента корреляции

$$r^*(x, y) = \frac{K^*(x, y)}{\hat{s}(x)\hat{s}(y)}, \tag{3}$$

где

$$K^*(x, y) = \frac{1}{n-1}\sum_{k=1}^{n}(x_k - \bar{x})(y_k - \bar{y}), \tag{4}$$

$$\hat{s}(x) = \sqrt{\frac{1}{n-1}\sum_{k=1}^{n}(x_k - \bar{x})^2}, \quad \hat{s}(y) = \sqrt{\frac{1}{n-1}\sum_{k=1}^{n}(y_k - \bar{y})^2}, \quad \bar{x} = \frac{1}{n}\sum_{i=1}^{n}x_i, \quad \bar{y} = \frac{1}{n}\sum_{i=1}^{n}y_i.$$

Если выборочные коэффициенты корреляции для пары показателей x_i и x_j в группах $X^{(1)}$ и $X^{(2)}$ имеют незначительное различие, то необходимо перейти к доверительным интервалам.

Доверительным интервалом, содержащим основную распределенную массу значений генеральной совокупности G с доверительным уровнем β [3], называется интервал (a, b), для которого выполняется условие

$$p(x \in (a, b)) = \beta, \quad x \in G, \text{ где } \beta \geq 0{,}95.$$

Приближенный доверительный интервал для коэффициента корреляции $r(x, y)$ с уровнем значимости, не превышающим 0,05, имеет вид

$$J_r = \left(\frac{\bar{\gamma}}{s(x)s(y)} - \frac{3s_\gamma}{\sqrt{n}s(x)s(y)}, \frac{\bar{\gamma}}{s(x)s(y)} + \frac{3s_\gamma}{\sqrt{n}s(x)s(y)} \right),$$

где

$$\bar{\gamma} = \frac{1}{n-1}\sum_{i=1}^{n}\gamma_i;\ \gamma_i = (x_i - \bar{x})(y_i - \bar{y});\ s_\gamma = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(\gamma_i - \bar{\gamma})^2};$$

$$s(x) = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2};\ s(y) = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2};\ \bar{x} = \frac{1}{n}\sum_{i=1}^{n}x_i;\ \bar{y} = \frac{1}{n}\sum_{i=1}^{n}y_i.$$

Доверительный интервал построен по правилу $3s_1$, изложенному в [3]. По правилу $2s_1$ этот интервал имеет вид

$$J_r^{(1)} = \left(\frac{\bar{\gamma}}{s(x)s(y)} - \frac{2s_\gamma}{\sqrt{n}s(x)s(y)},\ \frac{\bar{\gamma}}{s(x)s(y)} + \frac{2s_\gamma}{\sqrt{n}s(x)s(y)}\right).$$

Для асимптотически нормальных оценок $r^*(x, y)$ более точным является интервал $J_r^{(1)}$, который используется в настоящей статье.

Покажем, что при больших значениях n уровень значимости доверительного интервала J_r для неизвестного коэффициента корреляции r не превосходит 0,05, поэтому его можно считать доверительным интервалом, содержащим основную распределенную массу оценок r. Начнем исследования с изучения оценок коэффициента ковариации $K^*(x, y)$ двух случайных величин x, y, с помощью которых вычисляются оценки r.

Теорема 1. Выборочный коэффициент ковариации (4)

$$K^*(x, y) = \frac{1}{n-1}\sum_{k=1}^{n}(x_k - \bar{x})(y_k - \bar{y})$$

является асимптотически нормальной случайной величиной, если выборки $(x_1, \ldots, x_n)$ и $(y_1, \ldots, y_n)$ получены в результате простого случайного выбора.

Доказательство. Рассмотрим несколько модифицированную оценку для $K(x, y)$:

$$\hat{K}(x, y) = \frac{1}{n}\sum_{k=1}^{n}(x_k - \bar{x})(y_k - \bar{y}),$$

$$K^*(x, y) = \frac{n}{n-1}\hat{K}(x, y).$$

Нетрудно показать, что

$$\hat{K}(x, y) = \tilde{K}(x, y) = \frac{1}{n}\sum_{k=1}^{n}x_k y_k - \bar{x}\bar{y}.$$

Предположим, что $z = xy$, $z_k = x_k y_k$, $u_k = z_k - m(z_k)$, $\sigma = \sigma(u_k) = \sigma(z)$, $\bar{z} = \frac{1}{n}\sum_{k=1}^{n}x_k y_k$, $D(\bar{z}) = \frac{D(z)}{n}$, $\sigma(\bar{z}) = \frac{\sigma(z)}{\sqrt{n}}$, $s_n^2 = \frac{1}{n}\sum_{k=1}^{n}(z_k - \bar{z})^2$, $\sigma_n = \sigma(\bar{z})$.

Легко увидеть, что z_k являются независимыми в совокупности одинаково распределенными случайными величинами, поэтому на основании центральной предельной теоремы [4]

$$\sup_n \left| \mathrm{P}\left(\frac{1}{\sigma\sqrt{n}}\sum_{k=1}^{n}u_k \le v\right) - \Phi(v)\right| \to 0 \text{ при } n \to \infty,$$

где $\Phi(v) = \frac{1}{2\pi}\int_{-\infty}^{v} e^{-x^2/2}dx$ — функция распределения нормированной нормально распределенной случайной величины, поэтому

$$\mathrm{P}\left(\frac{1}{\sigma\sqrt{n}}\sum_{k=1}^{n}u_k < v\right) \to \Phi(v) \text{ при } n \to \infty. \qquad (5)$$

В силу теоремы Колмогорова [5] случайная величина $\overline{xy}$ сходится почти наверное (с вероятностью единица) к величине $c = m(x)m(y)$ при $n \to \infty$:

$$\overline{xy} \xrightarrow{\text{п.н.}} c = m(x)m(y), \quad n \to \infty.$$

Отсюда на основании элементарной предельной теоремы следует [1]

$$\mathrm{P}\left(\frac{1}{\sigma\sqrt{n}} \sum_{k=1}^{n} u_k - \overline{xy} + m(\bar{x})m(\bar{y}) < v \right) \to \Phi(v), \quad n \to \infty. \tag{6}$$

Покажем, что из предельного соотношения (6) вытекает

$$\mathrm{P}\left(\frac{1}{n} \sum_{k=1}^{n} u_k + w_n < v \frac{\sigma}{\sqrt{n}} \right) \to \Phi(v) \quad \text{при} \quad n \to \infty, \tag{7}$$

где $w_n = -\overline{xy} + m(x)m(y)$. На основании (5)

$$\mathrm{P}\left(\frac{1}{n} \sum_{k=1}^{n} u_k < v \frac{\sigma}{\sqrt{n}} \right) \to \Phi(v), \quad n \to \infty. \tag{8}$$

Обозначим $a_n = \frac{1}{n} \sum_{k=1}^{n} u_k$, $b_n = a_n + w_n$, $\alpha_n = v \frac{\sigma}{\sqrt{n}}$, $F_n(\alpha_n) = \mathrm{P}(a_n < \alpha_n)$, $G_n(\alpha) = \mathrm{P}(b_n < \alpha)$. Если $a_n < \alpha_n$ и $w_n < 0$, то $b_n < \alpha_n$. Отсюда следует, что если $a_n < \alpha_n$, то или $b_n < \alpha_n$, или $w_n > 0$, поэтому

$$\mathrm{P}(a_n < \alpha_n) \le \mathrm{P}(b_n < \alpha_n) + \mathrm{P}(w_n > 0)$$

или

$$F_n(\alpha_n) \le G_n(\alpha_n) + \mathrm{P}(w_n > 0). \tag{9}$$

Так как w_n стремится почти наверное (т.е. с вероятностью, равной единице) к нулю, то вероятность события $w_n > 0$ для достаточно больших n будет меньше ε, поскольку $\lim\limits_{n\to\infty} \mathrm{P}(w_n > 0) = 0$. Таким образом, из (9) получаем

$$F_n(\alpha_n) < G_n(\alpha_n) + \varepsilon.$$

В силу предельного соотношения (8)

$$\lim_{n\to\infty} F_n(\alpha_n) = \Phi(v) \le \lim_{n\to\infty} G_n(\alpha_n) + \varepsilon, \tag{10}$$

а значит, $\Phi(v) \le \lim\limits_{n\to\infty} G_n(\alpha_n)$.

Аналогично, поменяв местами a_n и b_n, можно доказать, что

$$G_n(\alpha_n) < F_n(\alpha_n) + \varepsilon. \tag{11}$$

Из (10) и (11) следует равенство

$$\lim_{n\to\infty} G_n(\alpha_n) = \Phi(v);$$

следовательно, предельное соотношение (7) доказано.

Легко увидеть, что

$$K(x, y) = m\{[x - m(x)][y - m(y)]\} = m(xy) - m(x)m(y),$$

поэтому $\hat{K}(x, y) - K(x, y) = \frac{1}{n} \sum_{k=1}^{n} x_k y_k - \overline{xy} + m(x)m(y)$.

В силу (7)

$$\mathrm{P}(\hat{K}(x, y) - K(x, y) < v \frac{\sigma}{\sqrt{n}}) \to \Phi(v) \text{ при } n \to \infty; \tag{12}$$

следовательно, оценки $\hat{K}(x, y)$ и $K^*(x, y)$ являются асимптотически нормальными случайными величинами.

Теорема доказана.

Напомним [3], что интервалы $(a_k(x_1,\dots,x_k), b_k(x_1,\dots,x_k))$, $k=1,2,\dots$, называются асимптотическими доверительными для показателя p, отвечающими уровню значимости β, если

$$\underline{\lim}_{k\to\infty} \mathrm{P}(p \in (a_k(x_1,\dots,x_k), b_k(x_1,\dots,x_k))) = 1-\beta,$$

а $a_k(x_1,\dots,x_k), b_k(x_1,\dots,x_k)$ называются асимптотическими доверительными границами.

Следствие. Асимптотический уровень значимости доверительного интервала для коэффициента ковариации $K(x,y)$, имеющего вид

$$J = \left(\bar{\gamma} - 2\frac{s_\gamma}{\sqrt{n}}, \bar{\gamma} + 2\frac{s_\gamma}{\sqrt{n}}\right),$$

где $s_\gamma^2 = \frac{1}{n-1}\sum_{i=1}^{n}(\gamma_i - \bar{\gamma})^2$, $\bar{\gamma} = \frac{1}{n-1}\sum_{i=1}^{n}\gamma_i$, а $\gamma_i = (x_i - \bar{x})(y_i - \bar{y})$, $i=1,2,\dots,n$, не превосходит 0,05 [3].

Доказательство. Очевидно, что $\bar{\gamma} = K^*(x,y)$. Из формулы (12) вытекает, что

$$\mathrm{P}\left(|K^*(x,y) - K(x,y)| \le \upsilon\frac{\sigma}{\sqrt{n}}\right) = \mathrm{P}\left(|\bar{\gamma} - K(x,y)| \le \upsilon\frac{\sigma}{\sqrt{n}}\right) =$$

$$= \mathrm{P}\left(-\upsilon\frac{\sigma}{\sqrt{n}} \le \bar{\gamma} - K(x,y) \le \upsilon\frac{\sigma}{\sqrt{n}}\right) =$$

$$= \mathrm{P}\left(\bar{\gamma} - K(x,y) < -\upsilon\frac{\sigma}{\sqrt{n}}\right) - \mathrm{P}\left(\bar{\gamma} - K(x,y) < -\upsilon\frac{\sigma}{\sqrt{n}}\right) \to \Phi(\upsilon) - \Phi(-\upsilon) \text{ при } n\to\infty.$$

В частности, при $\upsilon = 2$ имеем

$$\mathrm{P}\left(|\bar{\gamma} - K(x,y)| \le 2\frac{\sigma}{\sqrt{n}}\right) \to \Phi(2) - \Phi(-2) \approx 0{,}9772 - 0{,}0228 \approx 0{,}95432.$$

Поэтому для доверительного интервала $J^*\left(\bar{\gamma} - 2\frac{\sigma}{\sqrt{n}}, \bar{\gamma} + 2\frac{\sigma}{\sqrt{n}}\right)$ асимптотический уровень значимости не превосходит 0,05. Можно показать, что случайная величина $s_\gamma^2 = s_n^2$ сходится почти наверное к величине $\sigma^2 = \sigma^2(z)$ при $n\to\infty$ и асимптотический уровень значимости доверительного интервала J также не превосходит 0,05.

Следствие доказано.

Как известно [3], коэффициент корреляции

$$r(x,y) = \frac{K(x,y)}{\sigma(x)\sigma(y)}$$

отличается от коэффициента ковариации $K(x,y)$ лишь множителем $\lambda = \frac{1}{\sigma(x)\sigma(y)} > 0$.

В связи с этим доверительные границы для коэффициента корреляции $r(x,y)$, полученные на основе правила $2s_1$, можно вычислить с помощью следующих простых аналитических преобразований:

$$\mathrm{P}\left(K(x,y) \in \left(\bar{\gamma} - \frac{2s_\gamma}{\sqrt{n}}, \bar{\gamma} + \frac{2s_\gamma}{\sqrt{n}}\right)\right) = \mathrm{P}\left(\lambda K(x,y) \in \left(\lambda\bar{\gamma} - \frac{2\lambda s_\gamma}{\sqrt{n}}, \lambda\bar{\gamma} + \frac{2\lambda s_\gamma}{\sqrt{n}}\right)\right) =$$

$$= \mathrm{P}\left(r(x,y) \in \left(\frac{\bar{\gamma}}{\sigma(x)\sigma(y)} - \frac{2}{\sqrt{n}}\frac{s_\gamma}{\sigma(x)\sigma(y)}, \frac{\bar{\gamma}}{\sigma(x)\sigma(y)} + \frac{2}{\sqrt{n}}\frac{s_\gamma}{\sigma(x)\sigma(y)}\right)\right) \ge 0{,}95.$$

Если заменить $\sigma(x)$ и $\sigma(y)$ их оценками $s(x)$ и $s(y)$, то получим доверительный интервал для неизвестного коэффициента корреляции $r(x, y)$

$$J_r = \left(\frac{\bar{\gamma}}{s(x)s(y)} - \frac{2}{\sqrt{n}} \frac{s_\gamma}{s(x)s(y)},\ \frac{\bar{\gamma}}{s(x)s(y)} + \frac{2}{\sqrt{n}} \frac{s_\gamma}{s(x)s(y)} \right),$$

уровень значимости которого приблизительно равен 0,05.

Если вычисленные доверительные интервалы практически полностью перекрываются: $J_r^{(1)} \approx J_r^{(2)}$, то это означает, что коэффициенты корреляции показателей мало отличаются один от другого. Если же доверительные интервалы не перекрываются или имеют незначительное пересечение, причем $r^{*(1)}(x_i, x_j) > r^{*(2)}(x_i, x_j)$, то зависимость показателей x_i и x_j в первой группе является более выраженной и эта зависимость значительно превосходит ее во второй группе.

3. Метод распознавания, основанный на доверительных интервалах математического ожидания. Рассмотрим задачу распознавания элемента группы по заданным значениям его показателей. Предположим, что $X^{(1)}$ и $X^{(2)}$ — две группы, каждая из которых характеризуется n показателями. Необходимо установить, к какой группе принадлежит элемент $x = (x_1, x_2, \ldots, x_n)$.

Для решения этой задачи необходимо сначала вычислить доверительные интервалы математического ожидания каждого показателя в каждой группе. Приближенный доверительный интервал для неизвестного математического ожидания $m(x)$, построенный по правилу $3s_1$, имеет вид

$$J = \left(\bar{x} - 3\frac{s}{\sqrt{n}},\ \bar{x} + 3\frac{s}{\sqrt{n}} \right),$$

где $\bar{x}$ — выборочное математическое ожидание [6].

Таким образом, имеем $(J_1^{(1)}, J_2^{(1)}, \ldots, J_n^{(1)})$ и $(J_1^{(2)}, J_2^{(2)}, \ldots, J_n^{(2)})$, где $J_i^{(k)}$ — доверительный интервал математического ожидания i-го показателя в k-й группе, $i = \overline{(1, n)}$, $k = (1, 2)$. Если пары интервалов $(J_i^{(1)}, J_i^{(2)})$, $i = \overline{(1, n)}$, пересекаются, то из каждого интервала исключается их пересечение.

Далее рассматривается индикаторная функция

$$\varphi(x) = \sum_{i=1}^{n} \varphi_i, \text{ где } \varphi_i = \begin{cases} 1, & \text{если } x_i \in J_i^{(1)}, \\ -1, & \text{если } x_i \in J_i^{(2)}, \\ 0, & \text{если } x_i \notin J_i^{(1)},\ x_i \notin J_i^{(2)}. \end{cases}$$

Таким образом, элемент $x = (x_1, x_2, \ldots, x_n)$ принадлежит группе $X^{(1)}$, если $\varphi(x) > 0$; x принадлежит группе $X^{(2)}$, если $\varphi(x) < 0$. При $\varphi(x) = 0$ имеем случай непринятия решения.

ПРАКТИЧЕСКОЕ ПРИМЕНЕНИЕ

Рассмотренные подходы были использованы для статистического анализа связей между уровнем гомоцистеина в крови и различными клиническими показателями в двух сопоставляемых группах пациентов с онкологическими заболеваниями — с отягощенным и неотягощенным онкологической патологией анамнезом. Исследуемая группа состояла из 100 женщин, прооперированных по поводу рака молочной железы, с гистологически подтвержденным диагнозом. Из них 47 пациентов имеют родственников с онкологическими заболеваниями и составляют группу пациентов с отягощенным анамнезом, остальные пациенты (53 человека) составляют группу лиц с неотягощенным анамнезом [7]. У каждого пациента измерялся уровень гомоцистеина (мкмоль/ л) в крови [8, 9]. Кроме того, на каждого паци-

Таблица 1

Обозначение элемента	Исследуемые факторы
y	Содержание гомоцистеина, мкмоль/л
x_1	Возраст, лет
x_2	Индекс массы тела
x_3	Размеры опухоли, см
x_4	Наличие метастазов: 0 — нет, 1 — есть
x_5	Наличие предшествующих доброкачественных изменений в молочной железе: 0 — нет, 1 — есть
x_6	Наличие сердечно-сосудистых заболеваний: 0 — нет, 1 — есть
x_7	Наличие заболеваний органов пищеварения: 0 — нет, 1 — есть
x_8	Наличие заболеваний органов репродуктивной системы: 0 — нет, 1 — есть
x_9	Наличие онкологических заболеваний: 0 — нет, 1 — есть
x_{10}	Наличие заболеваний эндокринной системы: 0 — нет, 1 — есть
x_{11}	Курение: 0 — нет, 1 — да
x_{12}	Предпочтение кофе: 0 — нет, 1 — да
x_{13}	Преобладание пищи животного происхождения: 0 — нет, 1 — да
x_{14}	Преобладание пищи растительного происхождения: 0 — нет, 1 — да
x_{15}	Полноценная пища по содержанию витаминов и микронутриентов: 0 — нет, 1 — да
x_{16}	Употребление витамина группы В: 0 — нет, 1 — да
x_{17}	Употребление поливитаминов: 0 — нет, 1 — да
x_{18}	Употребление гормональных препаратов: 0 — нет, 1 — да
x_{19}	Употребление медикаментов: 0 — нет, 1 — да
x_{20}	Митозы: 0 — нет, 1 — единичные, 2 — умеренное количество, 3 — большое количество
x_{21}	Степень злокачественности опухоли: 1 — I степень, 2 — II степень, 3 — III степень
x_{22}	Продолжительность климакса, лет
x_{23}	Количество родов
x_{24}	Количество абортов
x_{25}	Количество родственников со злокачественными новообразованиями

ента заполнена анкета с учетом основных факторов, которые с различной долей вероятности могут оказывать влияние на уровень гомоцистеина. Принципы биоэтики в исследовании соблюдены. В табл. 1 приведен перечень рассматриваемых факторов.

ЗАДАЧИ ИССЛЕДОВАНИЯ

1. Сравнение значимости факторов, влияющих на уровень гомоцистеина, в группах пациентов с отягощенным и неотягощенным анамнезом.

2. Сравнение связей между парами факторов в обеих группах.

3. Определение принадлежности пациента к группе больных с отягощенным или неотягощенным анамнезом по значениям, наиболее характерным для этих групп факторов.

Рассмотрим каждую из поставленных задач.

1. Сравнение значимости факторов, влияющих на уровень гомоцистеина, в группах пациентов с отягощенным и неотягощенным анамнезом. Прежде чем использовать в исследовании модель многомерной линейной регрессии, необходимо убедиться в независимости рассматриваемых факторов $x_1, x_2, \ldots, x_n$. Для этого подсчитываются коэффициенты корреляции между всеми факторами и исключаются из рассмотрения те, у которых абсолютные значения коэффициента корреляции являются достаточно большими: $|r^*(x, y)| > 0{,}4$.

Подсчитанные коэффициенты корреляции показали, что в рассматриваемой нами группе, состоящей из 100 женщин, которые больны раком молочной железы, возраст прямо пропорционален продолжительности климакса ($r^*(x_1, x_{22}) = 0{,}61$), наличие сердечно-сосудистых заболеваний коррелирует с преобладанием пищи животного происхождения в рационе пациента ($r^*(x_6, x_{13}) = -0{,}44$), степень злокачественности опухоли прямо пропорциональна количеству митозов ($r^*(x_{20}, x_{21}) = 0{,}55$). Таким образом, из рассмотрения в исследованной группе больных были исключены факторы x_{13} (преобладание пищи животного происхождения), x_{20} (количество митозов) и x_{22} (продолжительность климакса).

Таблица 2

Исключенные факторы	Ошибка модели, SS
Присутствуют все	1178,366
x_1	1183,586
x_2	1191,84
x_3	1199,476
x_4	1209,599
x_5	1183,644
x_6	1180,65
x_7	1193,085
x_8	1179,847
x_9	1182,247
x_{10}	1182,552
x_{11}	1189,21
x_{12}	1183,8
x_{14}	1180,191
x_{15}	1191,238
x_{16}	1191,042
x_{17}	1180,311
x_{18}	1183,734
x_{19}	1198,078
x_{21}	1179,059
x_{23}	1192,661
x_{24}	1192,904
x_{25}	1185,334

Таблица 3

Исключенные факторы	Ошибка модели, SS	
	Отягощенный анамнез	Неотягощенный анамнез
Присутствуют все	354,8173	468,485
x_1	359,9798	509,0228
x_2	369,1878	469,2073
x_3	379,2588	470,0018
x_4	371,3992	483,0893
x_5	355,4097	490,2465
x_6	394,1127	468,6959
x_7	364,1483	468,5654
x_8	357,1034	472,7677
x_9	374,2444	472,5644
x_{10}	354,8314	468,489
x_{11}	356,3341	469,4806
x_{12}	359,3726	472,9062
x_{14}	363,6511	480,9804
x_{15}	356,0098	469,0198
x_{16}	355,2347	492,2281
x_{17}	356,8427	512,9461
x_{18}	356,8427	506,0835
x_{19}	360,762	477,0975
x_{21}	369,4536	468,5122
x_{23}	370,1534	473,1543
x_{24}	355,2668	493,996
x_{25}	358,4712	–

Для оставшихся 22 факторов была построена модель многомерной линейной регрессии, где $x_1,\ldots,x_{12},x_{14},\ldots,x_{19},x_{21},x_{23},\ldots,x_{25}$ — рассматриваемые факторы, *SS* — остаточная сумма квадратов, характеризующая ошибку модели. Исключив по одному из этих показателей, получим новые модели усеченной многомерной регрессии. Ошибки этих моделей даны в табл. 2.

После ранжировки найденных ошибок имеем наиболее важные факторы, влияющие на уровень гомоцистеина: x_4 (наличие метастазов), x_3 (размеры опухоли), x_{19} (употребление медикаментов), x_7 (наличие заболеваний органов пищеварения).

Далее были построены модели многомерной линейной регрессии по тем же 22 факторам, но с разделением группы на пациентов с отягощенным анамнезом (47 человек) и по 21 фактору для группы лиц с неотягощенным анамнезом (53 человека). Во второй группе исключен показатель x_{25} (родственники со злокачественными новообразованиями), поскольку он равен нулю для всех пациентов. В табл. 3 получены ошибки моделей. Наиболее важные факторы, влияющие на уровень гомоцистеина в группах пациентов с отягощенным и неотягощенным анамнезом, представлены в табл. 4.

Сравним значимость вышеуказанных факторов каждой группы. Так, x_{17} (употребление поливитаминов) в группе пациентов с неотягощенным анамнезом по значимости занимает первое место, в то время как этот же фактор в группе лиц с отягощенным анамнезом занимает только шестнадцатое место; x_1 (возраст пациента) в группе с неотягощенным анамнезом занимает второе место, а в группе с отягощенным — одиннадцатое; x_{18} (употребление гормональных препаратов) у пациентов с неотягощенным анамнезом занимает третье, а с отягощенным — пятнадцатое; x_{24} (аборты) в группе с неотягощенным анамнезом — четвертое место, а с отягощенным —

Таблица 4

Обозначение элемента	Исследуемые факторы
Неотягощенный анамнез	
x_{17}	Употребление поливитаминов: 0 — нет, 1 — да
x_1	Возраст, лет
x_{18}	Употребление гормональных препаратов: 0 — нет, 1 — да
x_{24}	Количество абортов
Отягощенный анамнез	
x_6	Наличие сердечно-сосудистых заболеваний: 0 — нет, 1 — есть
x_3	Размеры опухоли, см
x_9	Наличие онкологических заболеваний: 0 — нет, 1 — есть
x_4	Наличие метастазов: 0 — нет, 1 — есть

Таблица 5

Обозначение элемента	Коэффициенты регрессии	
	Отягощенный анамнез	Неотягощенный анамнез
x_1	0,04665	0,12242
x_3	1,08407	–0,33610
x_6	–3,44814	0,20090
x_9	3,66469	3,41251
x_{17}	–0,62940	2,34255
x_{18}	–5,39385	3,92065
x_{24}	0,05080	–0,43027

двадцатое; x_6 (наличие сердечно-сосудистых заболеваний) у пациентов с отягощенным анамнезом занимает первое место по значимости, а с неотягощенным — только восемнадцатое; x_3 (размеры опухоли) — второе место с отягощенным анамнезом и четырнадцатое — с неотягощенным; x_9 (наличие онкологических заболеваний) занимает третье место с отягощенным анамнезом и тринадцатое — с неотягощенным; x_4 (наличие метастазов) занимает четвертое место у пациентов с отягощенным анамнезом и седьмое — с неотягощенным.

Все рассмотренные факторы (кроме x_4) имеют значительные различия в степени своей значимости в группах пациентов с отягощенным и неотягощенным анамнезом. Таким образом, значимость факторов, оказывающих влияние на уровень гомоцистеина в крови пациента, зависит от группы, к которой пациенты относятся: с отягощенным или неотягощенным анамнезом. Для того чтобы уточнить, как именно связаны факторы с прогнозируемым показателем (в данном случае с гомоцистеином), необходимо рассмотреть коэффициенты линейной многомерной регрессии (табл. 5).

Если пациент имеет родственников с онкологическими заболеваниями, т.е. он относится к группе с отягощенным анамнезом, то увеличение количества гомоцистеина в его крови может быть связано с наличием онкологических заболеваний, отсутствием сердечно-сосудистой патологии и большими размерами опухоли. Если у пациента нет родственников с онкологическими заболеваниями, т.е. он относится к группе с неотягощенным анамнезом, то увеличение количества гомоцистеина в его крови может зависеть от возраста, употребления поливитаминов и гормональных препаратов и отсутствия родов.

Для нахождения точности построенных моделей линейных многомерных регрессий вычислим коэффициент детерминации R^2, который определяет долю объясняемой изменчивости соответствующих переменных.

Модель линейной многомерной регрессии, построенная для всех больных раком молочной железы в целом, имеет очень низкий коэффициент детерминации ($R^2 = 0{,}2$), в то время как модели регрессии, построенные для групп с отягощенным и неотягощенным анамнезом, — достаточно высокий ($R^2 = 0{,}45$ и $R^2 = 0{,}42$ соответственно). Это означает, что зависимость уровня гомоцистеина от рассматриваемых факторов наиболее точно прослеживается при делении пациентов на две группы: с отягощенным и неотягощенным анамнезом.

2. Сравнение связей между парами факторов в группах. Воспользуемся доверительными интервалами (a, b) для коэффициента корреляции $r^*(x, y)$, чтобы установить детальное различие между группами пациентов с отягощенным и нео-

тягощенным анамнезом. Подсчитав доверительные интервалы для всех пар показателей, выделим те, для которых пересечение доверительных интервалов составляет менее 10% (от меньшего интервала) (табл. 6).

Таблица 6

Обозначение элемента	$r^*(x,y)$	Доверительные интервалы (a,b)	$r^*(x,y)$	Доверительные интервалы (a,b)	Показатель пересечения, %
	Неотягощенный анамнез		Отягощенный анамнез		
x_3 и x_{10}	–0,43099	(–0,6759, –0,18611)	0,14416	(–0,1369, 0,425216)	0
x_8 и x_{21}	–0,15155	(–0,4282, 0,125106)	0,354609	(0,086104, 0,623115)	7
x_4 и x_{10}	–0,19337	(–0,363, –0,02379)	0,230475	(–0,0585, 0,519447)	10

На основании приведенных данных можно сделать следующие обобщения:

1) у пациентов с неотягощенным анамнезом наличие заболеваний эндокринной системы коррелирует с размерами опухоли; с отягощенным анамнезом такая связь не установлена;

2) у пациентов с отягощенным анамнезом наличие заболеваний органов репродуктивной системы коррелирует со степенью злокачественности опухоли; с неотягощенным анамнезом такая связь не установлена;

3) у пациентов с отягощенным анамнезом наличие заболеваний эндокринной системы коррелирует с наличием метастазов; с неотягощенным анамнезом корреляция не обязательна.

3. Установление принадлежности пациента к группе больных с отягощенным или неотягощенным анамнезом по значению его факторов. Доверительные интервалы (a, b) для неизвестного математического ожидания $\bar{x}$ каждого показателя в соответствующей группе, построенные по правилу $3s_1$, представлены в табл. 7.

Таблица 7

Обозначение элемента	$\bar{x}$	Доверительные интервалы (a,b)	$\bar{x}$	Доверительные интервалы (a,b)
	Неотягощенный анамнез		Отягощенный анамнез	
x_1	56,45283	(52,5296, 60,37599)	52,14893	(47,80642, 56,49144)
x_2	26,73396	(24,6296, 28,83823)	28,3	(26,12835, 30,47164)
x_3	2,118867	(1,82291, 2,414819)	2,404255	(1,989321, 2,819189)
x_4	0,094339	(–0,0272, 0,215944)	0,191489	(–0,005022, 0,388000)
x_5	0,226415	(0,05230, 0,400526)	0,404255	(0,187185, 0,621325)
x_6	0,698113	(0,50712, 0,889100)	0,723404	(0,525545, 0,921263)
x_7	0,547169	(0,34008, 0,754254)	0,702127	(0,499841, 0,904413)
x_8	0,339622	(0,14260, 0,536644)	0,382978	(0,167958, 0,597999)
x_9	0,018867	(–0,0377, 0,075471)	0,042553	(–0,046729, 0,131835)
x_{10}	0,264150	(0,08073, 0,447568)	0,404255	(0,187185, 0,621325)
x_{11}	0,132075	(–0,0087, 0,272930)	0,106382	(–0,029998, 0,242764)
x_{12}	0,226415	(0,05230, 0,400526)	0,255319	(0,062447, 0,448191)
x_{14}	0,396226	(0,19274, 0,599709)	0,255319	(0,062447, 0,448191)
x_{15}	0,207547	(0,03882, 0,376266)	0,212765	(0,031738, 0,393793)
x_{16}	0,792452	(0,62373, 0,961172)	0,829787	(0,663552, 0,996021)
x_{17}	0,056603	(–0,0395, 0,152740)	0,127659	(–0,019949, 0,275268)
x_{18}	0,471698	(0,26401, 0,679377)	0,319148	(0,112960, 0,525337)
x_{19}	0,094339	(–0,0272, 0,215944)	0,042553	(–0,046729, 0,131835)
x_{21}	0,132075	(–0,0087, 0,272930)	0,042553	(–0,046729, 0,131835)
x_{23}	1,226415	(0,88697, 1,565856)	1,340425	(0,983792, 1,697058)
x_{24}	2,509433	(2,27212, 2,746744)	2,425531	(2,188564, 2,662499)

Все пары интервалов пересекаются, поэтому из каждого интервала исключается их общее пересечение, затем вычисляется индикаторная функция для пациентов, для которых необходимо найти принадлежность к определенной группе. Исследование было проведено у 100 пациентов. У 63% больных такая принадлежность была определена правильно, у 13% — неправильно и у 24% — непринятие решения.

ВЫВОДЫ

1. С помощью модели усеченной многомерной линейной регрессии установлена значимость каждого исследованного клинического показателя по отношению к уровню гомоцистеина в крови больных раком молочной железы.

2. Вычислены коэффициенты корреляции между уровнем гомоцистеина и каждым клиническим показателем. Построены доверительные интервалы для коэффициентов корреляции и математического ожидания.

3. Предложен и апробирован на клиническом материале больных раком молочной железы с отягощенным и неотягощенным онкологической патологией анамнезом метод распознавания групповой принадлежности пациентов по этому признаку, основанный на построении доверительных интервалов и вычислении индикаторных функций.

4. Предложенные математические методы и модели позволили провести оценку зависимостей некоторых клинических показателей и уровня гомоцистеина в крови больных раком молочной железы и с высокой достоверностью выявить статистически значимые закономерности.

СПИСОК ЛИТЕРАТУРЫ

1. Ван дер Варден Б.Л. Математическая статистика. — М.: ИЛ, 1960. — 435 с.
2. Феллер В. Введение в теорию вероятностей и ее приложения. — М.: Мир, Т.1, 1964. — 484 с.
3. Клюшин Д.А., Петунин Ю.И. Доказательная медицина. Применение статистических методов. — М.: ООО «И. Д. Вильямс», 2008. — 320 с.
4. Петров В.В. Суммы независимых случайных величин. — М.: Наука, 1972. — 414 с.
5. Лоев М. Теория вероятностей. — М.: ИЛ, 1962. — 719 с.
6. Петунин Ю.И., Клюшин Д.А., Кулик Г.И., Юрченко О.В., Тодор И.Н., Чехун В.Ф. Стратификационный анализ морфометрических показателей популяции раковых клеток с фенотипом лекарственной резистентности // Кибернетика и системный анализ. — 2005. — № 6. — С. 158–167.
7. Налєскіна Л.А., Поліщук Л.З., Анікусько Н.Ф., Лук'янова Н.Ю., Юрченко О.В., Півнюк В.М., Чехун В.Ф. Клініко-генеалогічна характеристика та морфологічні особливості пухлин хворих на рак молочної залози залежно від обтяженості родоводів на онкологічну патологію // Онкологія. — 2008. — **10**, № 1. — С. 44–50.
8. Цыбиков Н.Н., Цыбикова Н.М. Роль гомоцистеина в патологии человека // Успехи современной биологии. — 2007. — **127**, № 5. — С. 471–481.
9. Чехун В.Ф., Налєскіна Л.А., Призимирська Т.В., Лук'янова Н.Ю., Юрченко О.В., Лозовська Ю.В., Кулик Г.І., Смоланка І.І. Корекція порушень обміну гомоцистеїну у хворих на рак молочної залози (методичні рекомендації). — Київ: ТОВ ДІА, 2008. — 22 с.

Поступила 14.11.2008