УДК 681.5.015

РЕГРЕССИОННО-ЛОГИЧЕСКАЯ МОДЕЛЬ ДИАГНОСТИКИ ЗАБОЛЕВАНИЙ

Д.К. Тюмиков, С.А. Блашенцева, А.М. Субботин, Н.Н. Савченков

Статистические данные эксперимента обработаны тремя методами, вычислены коэффициенты корреляции, дисперсионные отношения (множественные, парные, дисперсионные отношения эффектов взаимодействия и дисперсионные отношения эффектов взаимосвязей) и информационные меры (множественные и парные). На основе их анализа выбраны доминантные переменные и предложена регрессионно-логическая модель диагностики заболевания.

Ключевые слова: коэффициент корреляции, дисперсионное отношение, информационная мера связи, доминантная переменная, регрессионно-логическая модель.

ВВЕДЕНИЕ

В медицинской практике все больше внимания уделяется математическим подходам. Они применяются для формализации описания заболеваний, решения диагностических задач, выбора методов лечения и оценки их эффективности и др. [1, 2]. Цель настоящей работы заключается в выявлении формальных признаков, вносящих наибольший вклад в формирование заключения о заболевании, и разработке математической модели, позволяющей установить диагноз заболевания.

1. ХАРАКТЕРИСТИКА ДАННЫХ И ПОСТАНОВКА ЗАДАЧИ

Имеется выборка объемом 156 наблюдений с числом предикторных переменных $n = 29$ (см. таблицу в Приложении) и одной зависимой переменной. Ряд переменных принимает ограниченное число значений: выходная переменная — {1, 2, 3}, пол — {0, 1} и др. Имеются и переменные, принимающие непрерывные значения. Диапазоны варьирования различны для различных переменных, поэтому в расчетах статистических мер применяется нормирование значений переменных для приведения их к одному диапазону [0; 1]. По данной информации (x, y), $x \in X^n \subset \Re^n$, $y \in Y^1 \subset \Re^1$, на основе элементов заданного множества функций $\Phi = \{f_1(x_i), ..., f_n(x_i)\}$, $i = \overline{1, n}$ с помощью бинарных операций $\Omega = \{+, —, *, /, **\}$ и множества коэффициентов $C = \{c_1, ..., c_k\}$ необходимо сконструировать многомерную зависимость $y(x)$, предположительно:

$$y = M[y|x] + \Xi = f(x_1, ..., x_n, c_1, ..., c_k) + \Xi,$$

где $M[y|x]$ — регрессия y по x, Ξ — помеха.

2. МЕТОД РАСЧЕТА

Исходная выборка была разбита на две части: обучающую, объемом 100 наблюдений, и контрольную, объемом 56 наблюдений. Обучающая выборка была обработана следующими мерами статистических связей: корреляционными (множественный и парные коэффициенты корреляции), дисперсионными [3] (множественное дисперсионное отношение, парные дисперсионные отношения, дисперсионные отношения эффектов взаимодействия, дисперсионные отношения эффектов взаимосвязей) и информационными [4] (многомерная информационная мера, парная информационная мера и парная условная информационная мера).

Основой подхода служит разложение множественного дисперсионного отношения, при анализе которого выбирается композиция многомерной

модели диагностики заболевания [3]. Введем обозначения: $n_{1,j} = C_n^{n-j}$, $j = \overline{0, n-1}$; $n_{2,j} = C_{n-j}^{n-j-i}$, $i = \overline{1, n-j}$; $X_{v,k} \subset X^n$ — v-е сочетание входных переменных размерности k, $0 \leqslant k \leqslant n$, например, $X_{0,2} = \{x_1, x_2\}$, $X_{1,2} = \{x_2, x_3\}$ и т. д.

Разложение дисперсии зависимой переменной имеет вид:

$$D[y] = M\Bigg[(y - M[y|X^n])^2 + \sum_{j=0}^{n-1}\sum_{v=1}^{n_{1,j}} M\Bigg[\Big(M[y|X_{v,n-j}] + \sum_{i=1}^{n-j}\sum_{v'=1}^{n_{2,j}} (-1)^i M[y|X_{v',\,n-j-i}]\Big)^2\Bigg]\Bigg] +$$

$$+ 2\sum_{j=0}^{n-1}\sum_{v=0}^{n_1}\sum_{j'=0}^{n'-1}\sum_{v''=1}^{n'_1} M\Bigg[\Big(M[y|X_{v,\,n-j}] + \sum_{i=1}^{n-j}\sum_{v'=1}^{n_2} (-1)^i M[y|X_{v',\,n-j-i}]\Big)\Big(M[y|X_{v'',\,n-j'}] +$$

$$+ \sum_{i'=1}^{n-j'}\sum_{v'''=1}^{n'_2} (-1)^{i'} M[y|X_{v''',\,n-j'-i'}]\Big)\Bigg], \qquad (1)$$

$$X_{v',\,n-j-i} \subset X_{v,\,n-j}.$$

Если поделить левую и правую части разложения (1) на $D[y]$ и переобозначить соответствующие слагаемые, то получим относительные величины, например, для двух переменных:

$$1 = \theta_{yx_1x_2} + \eta_{y(x_1x_2)} + \eta_{yx_1} + \eta_{yx_2} + 2(\eta_{y(x_1x_2)(x_1)} + \eta_{y(x_1x_2)(x_2)} + \eta_{y(x_1)(x_2)}),$$

где $\theta_{yx_1x_2}$ — доля (дисперсионное отношение) остаточной дисперсии, $\eta_{y(x_1x_2)}$ — дисперсионное отношение эффекта взаимодействия двух переменных; η_{yx_1}, η_{yx_2} — парные дисперсионные отношения; ($\eta_{y(x_1x_2)(x_i)}$ — дисперсионные отношения эффектов влияния взаимосвязей переменных x_1, x_2 и x_i, $i = 1, 2$.

Расчет коэффициентов корреляции используется не только как «отслеживание» линейной статистической связи, но и для определения знака связи (прямая или отрицательная), что важно для интерпретации феноменологических явлений при различных заболеваниях. Расчет информационных мер привлечен для контроля возможностей дисперсионных характеристик в оценивании статистических связей при наличии функционально неоднозначных зависимостей.

Поскольку авторы не ставили своей целью полное исследование взаимосвязей между всеми переменными, было принято решение остановиться на применении программы обработки, дающей значения мер связи для трех предикторных переменных и одной выходной. Расчет дисперсионных отношений и информационных мер при большем количестве переменных приводит к резкому возрастанию времени обработки, при этом, как будет видно далее, не давая возможности значительно улучшить конечную модель. Для получения пересечения эффектов взаимодействия и взаимосвязей переменных использовалась скользящая индексация: $\{x_1, x_2, x_3, y\}$, $\{x_2, x_3, x_4, y\}$, ..., $\{x_{29}, x_1, x_2, y\}$.

На основании изучения парных дисперсионных отношений (ПДО), парных коэффициентов корреляции (ПКК) и парных информационных мер (ПИМ) из заданных признаков определили те, которые имеют наибольшее значение для построения модели; *SN* (см. таблицу в Приложении) — номер предикторной переменной при ранжировании по степени их значимости для построения модели, сокращенно — номер значимости.

Доминантные (существенные) переменные выбраны на основе ПДО по формуле [3]:

$$Q = \min_m Q(m) = \min_m \left(\frac{1}{n-m}\sum_{i=m+1}^{n} \eta_{yx_i} \Big/ \frac{1}{m}\sum_{i=1}^{m} \eta_{yx_i}\right),$$

где m — число доминантных переменных, $1 \leqslant m \leqslant n$.

Содержательная интерпретация выбора заключается в разбиении ранжированного ряда парных дисперсионных отношений на две группы путем максимизации расстояния между центрами (средними значениями) этих групп.

Из таблицы видно, что ПДО и ПИМ достаточно согласованны. В большей части их согласованность поддерживают ПКК. Также видно отсутствие полного мажорирования ПДО над ПКК. Это связано с ограниченным числом уровней вариации предикторных переменных для оценивания дисперсионных отношений, в то время, как коэффициенты корреляции вычислялись без категоризации.

О предикторах 27 и 29, выделенных информационной мерой, можно сказать, что они являются аргументами, порождающими неоднозначность зависимой функции. Такую зависимость распознает только информационная мера [5, 6]. Применение ПИМ позволило выделить две переменные, отбрасываемые при корреляционных и дисперсионных мерах. Дополнительно было выяснено, что при построении модели диагноза без этих переменных около 5 % обследуемых будет отнесено не к той группе заболевания. Также было выяснено, что ис-

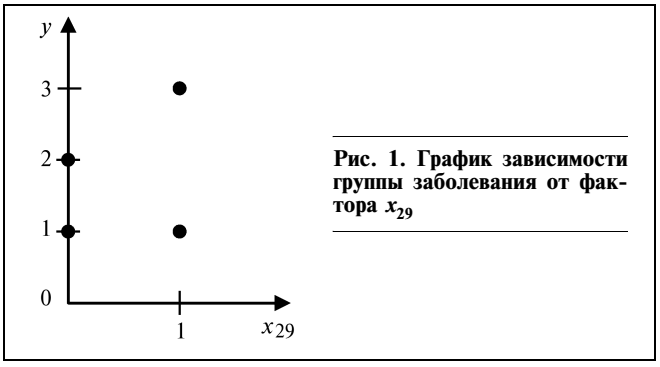


Рис. 1. График зависимости группы заболевания от фактора x_{29}

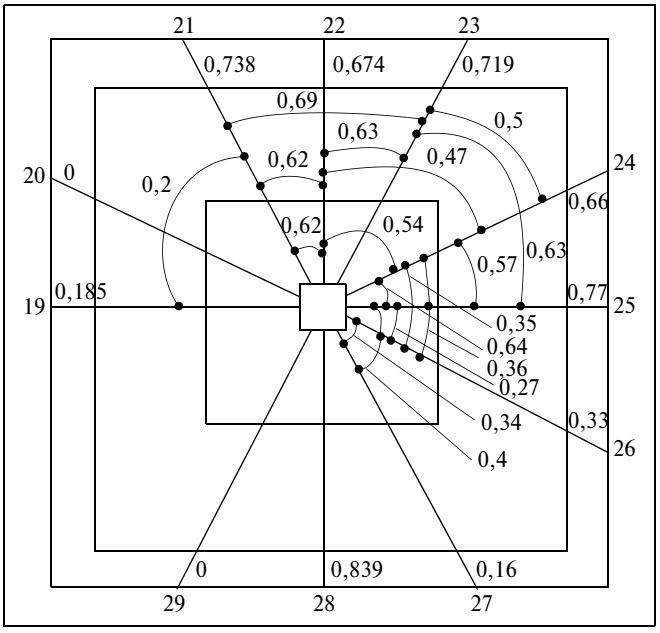


Рис. 2. Эффекты взаимодействия и взаимосвязей для признаков 19—29

пользование одной из этих переменных позволяет однозначно отнести больного ко второй или третьей группе, при этом первая группа заболевания надежно идентифицировалась на основе регрессионной модели.

Тот факт, что данную переменную удалось выделить только с помощью ПИМ, можно объяснить, обратившись к рис. 1. У всех больных второй группы данная переменная имеет значение, равное 0, тогда как у всех больных третьей группы значение данной переменной равно 1. У больных первой группы данный параметр может принимать оба значения. Очевидна неоднозначность зависимости как по аргументу, так и по функции.

Полный набор использованных при анализе дисперсионных отношений рассмотрим на переменных 19—29 (см. таблицу в Приложении). Эти переменные имеют наиболее выраженные эффекты взаимодействия и взаимосвязи, что представлено в виде диаграммы на рис. 2.

На диаграмме значения дисперсионных отношений эффектов взаимодействия отражены во внутреннем квадрате, эффектов взаимосвязи — в среднем, а парные дисперсионные отношения в крайнем поле (наружном квадрате). Вне поля проставлены номера входных переменных (предикторы) в общей выборке.

Видно, что наибольшее значение ПДО имеют переменные 28, 25, 21, 22, 23, 24 {0,84, ..., 0,66}. Переменные 28 и 23 использовались на начальном этапе анализа данных и при конструировании модели диагностики будут исключены, так как их значения определяются врачом-профессионалом.

Анализ дисперсионных отношений эффекта взаимодействия показывает, что набольшее взаимодействие наблюдается между переменными: (24, 25) → 0,64; (21—22) → 0,62; (22—24) → 0,54; 24—26 → 0,35; (24 —25—26) → 0,36; (25—26) → 0,27; (26—27) → 0,34; (25—26—27) → 0,4.

Значения дисперсионных отношений эффектов влияния взаимосвязей для этого набора переменных: (19)(21) → 0,2; (21)(22) → 0,62; (21)(23) → 0,69; (22)(23) → 0,63; (22)(24) → 0,47; (23)(24) → 0,5; (23)(25) → 0,63; (24)(25) → 0,57.

В ходе исследования были выявлены взаимодействие и взаимосвязь между рядом предикторов. Как оказалось, эти предикторы отражали, в целом, одну характеристику — состояние кислотопродуцирующей функции желудка, что, в общем, согласуется с современными представлениями о механизмах развития атрофического гастрита.

3. ОПРЕДЕЛЕНИЕ СТРУКТУРЫ МНОГОМЕРНОЙ ЗАВИСИМОСТИ НА ОСНОВЕ МЕР СТАТИСТИЧЕСКИХ СВЯЗЕЙ

При исследовании статистических связей доминантных переменных было выяснено, что среди набора дисперсионных отношений значительно преобладают парные дисперсионные отношения. Это позволяет выдвинуть гипотезу аддитивной многомерной зависимости [3]:

$$y = \sum_{i=1}^{m} f_i(x_i),$$

где m — число доминантных переменных, $f_i(x_i)$ — некоторая функция одной переменной.

Учитывая близость по значениям парных дисперсионных отношений и парного коэффициента

корреляции, можно предложить гипотезу линейной аддитивной композиции:

$$y = \sum_{i=1}^{m} c_i x_i.$$

Коэффициенты зависимости определялись на обучающей выборке методом наименьших квадратов.

На основе выявленных ведущих признаков была построена регрессионная модель, которая позволяет с высокой точностью отнести пациента к определенной группе:

$$y_p = \sum_{i=1}^{8} c_i x_i.$$

Непрерывный диапазон значений отклика регрессионной модели был разбит на поддиапазоны. Номер группы пациента определялся в зависимости от того, в какой поддиапазон попало значение, рассчитанное по модели диагностики: группа хронического гастрита без атрофии (№ 1), атрофического гастрита без геликобактериоза (№ 2) или атрофического гастрита с геликобактериозом (№ 3).

Учитывая вид взаимосвязи зависимой переменной с предиктором x_{29}, было решено ввести этот предиктор в модель операциями логики. В результате была получена регрессионно-логическая модель диагностики заболевания с безошибочной диагностикой на заданной выборке, включая контрольную:

$$y_д = \begin{cases} 1 \text{ при } y_p \in [0{,}5;\ 1{,}5), \\ 2 \text{ при } (y_p \in [1{,}5;\ 3{,}5)) \wedge (x_{29} = 0), \\ 3 \text{ при } (y_p \in [1{,}5;\ 3{,}5)) \wedge (x_{29} = 1), \end{cases}$$

где $\wedge$ — знак конъюнкции.

Описанная модель для практического применения врачами-гастроэнтерологами представлена в виде компьютерной программы. Спорным моментом до сих пор считалась необходимость проведения быстрого уреазного теста (предиктор x_{29}) у больных с хроническим гастритом и его диагностическое значение. Исследования показали, что введение в модель принятия решения о диагностике группы значения уреазного теста, однозначно определяет 2-ю или 3-ю группу. Первая группа пациентов диагностируется с высокой степенью вероятности без уреазного теста на основании построенной регрессионной модели.

ЗАКЛЮЧЕНИЕ

На основе статистических данных о пациентах при априорной неопределенности о виде модели диагностики заболевания были вычислены три группы статистических мер связей: корреляционные, дисперсионные и информационные. Анализ парных дисперсионных характеристик позволил выбрать доминантные переменные, уменьшив размерность вектора предикторных переменных с 29 до 9 (8 + 1). Дисперсионные отношения позволили вскрыть феноменологическую природу развития атрофического гастрита и постулировать аддитивную структуру многомерной зависимости. Информационные меры обнаружили неоднозначность зависимости выходной переменной от одного из предикторов, чем и была вызвана необходимость ввода этого предиктора в модель диагностики с помощью логических операций. Модель показала безошибочную диагностику на представленной выборке.

ПРИЛОЖЕНИЕ

Результаты исследования изучаемых признаков с учетом ПДО, ПКК и ПИМ

№	Наименование переменных	*SN*	ПДО	ПКК	ПИМ
1	Пол	0	0	−0,1	0
2	Возраст	**9**	0,5	0,67	0,61
3	Боли в эпигастрии	0	0	0	0
4	Дискомфорт в эпигастрии	0	0	0,16	0
5	Наследственность по раку желудка	0	0,123	0,35	0,135
6	Стресс	0	0	−0,13	0
7	Нарушение питания	0	0,13	−0,35	0,18
8	Длительность заболевания	0	0,35	0,21	0,5
9	Частота обострения в год	0	0,11	−0,1	0,17

№	Наименование переменных	*SN*	ПДО	ПКК	ПИМ
10	Пальпация (болезненность в эпигастрии)	0	0	−0,12	0
11	Норма	0	0	−017	0
12	Эндоскопические признаки эритематозной и застойной гастропатии	**6**	0,7	−0,85	0,8
13	Эндоскопические признаки атрофии антрума	0	0,36	0,6	0,35
14	Эндоскопические признаки атрофии антрума и тела	0	0,16	0,4	0,18
15	Эндоскопические признаки атрофии тела	0	0	0,14	0,11
16	Норма при морфологическом исследовании слизистой желудка	0	0	0	0
17	Хронический гастрит без атрофии по данным морфологии	**2**	0,8	−0,9	1
18	Атрофия антрума по данным морфологии	0	0,29	0,5	0,285
19	Атрофия антрума по данным эндоскопии	0	0,19	0,43	0,21
20	Атрофия антрума и тела по данным морфологии	0	0	0,14	0,11
21	pH тела	**4**	0,74	0,73	0,93
22	pH антрума	**7**	0,67	0,6	0,91
23	Заключение по результатам pH-метрии	**5**	0,72	0,8	0,86
24	Гастрин 17	**8**	0,66	−0,59	0,9
25	Гастрин 17 (стимул)	3	0,77	−0,8	0,92
26	Пепсиноген 1	0	0,33	−0,4	0,48
27	H. pylori	0	0,16	0	0,72
28	Заключение по результатам серологического исследования	**1**	0,84	0,8	1
29	Уреазный тест на H. pylori	0	0	0	0,6

ЛИТЕРАТУРА

1. *Леонов В.П.* Применение статистики в статьях и диссертациях по медицине и биологии. Ч. IV. Наукометрия статистической парадигмы экспериментальной биомедицины. URL: http://www/mediasphera.ru/mjmp/2002/3/r3-02-1.htm (дата обращения 29.11.2010).
2. *Блашенцева С.А., Субботин А.М., Ефимова Е.И.* Прогнозирование развития атрофического гастрита с использованием математического моделирования // Российский журнал гастроэнтерологии, гепатологии, колопроктологии. — № 5. — 2009. — С. 23.
3. *Тюмиков Д.К.* Идентификация многомерных нелинейных объектов на основе дисперсионных отношений. — Самара: изд-во СНЦ РАН, 2008. — С. 162.
4. *Савченков Н.Н., Тюмиков Д.К.* Информационные меры статистической связи для идентификации многомерных по входу объектов // Изв. Самарского науч. центра РАН. — 2007. — Т. 9, № 3 (21). — С. 650—653.
5. *Дисперсионная* идентификация / Под ред. Н.С. Райбмана. — М.: Наука, 1981. — 336 с.
6. *Савченков Н.Н., Тюмиков Д.К.* Энтропийная мера для идентификации нелинейных статистических связей // Математика. Компьютер. Образование: Сб. науч. тр. — М. — Ижевск, 2007. — Т. 2. — 392 с.

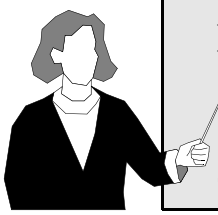
Статья представлена к публикации членом редколлегии В.Н. Новосельцевым.

Тюмиков Дмитрий Кондратович — канд. техн. наук, доцент, Самарский государственный университет путей сообщения, ✉ dktyumikov@mail.ru,

Блашенцева Светлана Александровна — д-р мед. наук, профессор, Институт последипломного образования Самарского государственного медицинского университета, ☎ (846) 333-71-44, ✉ svalb63@yandex.ru,

Субботин Александр Михайлович — зав. отделом, МЛПУ «Городская больница № 13», г. Нижний Новгород, ☎ (831) 294-33-16, ✉ subbotinam@rambler.ru,

Савченков Николай Николаевич — вед. инженер, Самарский государственный университет путей сообщения, ✉ nsavchen@rambler.ru.



Читайте в следующем номере

- ✓ ***Добровидов А.В., Кулида Е.Л., Рудько И.М.*** Управление движением объекта в конфликтной среде
- ✓ ***Зуев А.С., Федянин Д.Н.*** Модели управления мнениями агентов в социальных сетях
- ✓ ***Корноушенко Е.К.*** Линейный подход к управлению равновесными состояниями нелинейных нормированных моделей
- ✓ ***Жирабок А.Н., Бобко Е.Ю., Варнаков А.И., Писарец А.М.*** Учет неуправляемости системы при решении задачи функционального диагностирования
- ✓ ***Безгинов А.Н., Трегубов С.Ю.*** Многокритериальный подход к оценке расписания занятий на основе нечёткой логики