УДК 616-07-053.5

ПРИМЕНЕНИЕ РЕГРЕССИОННОГО АНАЛИЗА В ЗАДАЧЕ ИАГНОСТИРОВАНИЯ СОСТОЯНИЯ ЗДОРОВЬЯ ДЕТЕЙ

Черкашина Ю.А.

Национальный исследовательский Томский политехнический университет, Томск, Россия (634050, Россия, г. Томск, проспект Ленина, 30), e-mail: cherr999y@mail.ru

Статья является результатом научных исследований сотрудников кафедры Прикладной математики Национального исследовательского Томского политехнического университета и сотрудников Лечебно-оздоровительного центра «Здоровая мама – крепкий малыш». Исследования посвящены применению регрессионного анализа для оценки состояния здоровья детей первого года жизни на основе медицинских данных гормонов и веществ, отвечающих за жизнедеятельность организма, измеренных с первого по двенадцатый месяцы жизни ребенка. В работе дана подробная характеристика исследуемых медицинских данных. Произведен краткий обзор моделей решения задачи. В статье подробно рассмотрена процедура бинарной логистической регрессии. Приведены результаты исследования. Представлена классификационная таблица прогнозируемых значений и фактических наблюдаемых значений, определена общая процентная доля правильного распознавания. Произведен расчет коэффициентов регрессионного уравнения, на основании которых записано общее уравнение регрессии.

Ключевые слова: регрессионный анализ, бинарная логистическая регрессия, классификационная таблица, диагностика.

APPLICATION OF REGRESSION ANALYSIS FOR SOLVING DIAGNOSIS PROBLEM OF CHILDREN'S HEALTH

Cherkashina Y.A.

National Research Tomsk Polytechnic University, Tomsk, Russia, 634050, Tomsk, Lenin Avenue, 30, e-mail: cherr999y@mail.ru

The article includes results of scientific researches achieved at department of Applied Mathematics at National Research Tomsk Polytechnic University and Medical and health center "Healthy mother-strong baby" at Siberian State Medical University. Researches are devoted the application of regression analysis to assess the health status of children in the first year of life, based on medical data hormones and substances responsible for vital functions of the body, measured from the first to the twelfth months of life. In this paper a detailed description of the studied medical data is given. Binary logistic regression procedure is discussed in the article. Classification table predicted values and factual observed values is presented, the overall percentage of correct recognition is determined. Regression equation coefficients are calculated based on them general regression equation is written.

Keywords: regression analysis, binary logistic regression, classification table, diagnostics.

Проблема диагностирования состояния здоровья детей привлекает внимание все большего числа исследователей. Эта проблема особенно актуальна в педиатрии, где главной задачей является выявление патологий и хронических заболеваний на ранних этапах развития.

Хорошо известно, что расположенность человека к различным заболеваниям закладывается в основном в первый год жизни, поэтому оценка состояния здоровья именно в этот период является актуальной [1].

Поэтому следует производить оценку состояния здоровья детей в период первого года жизни с целью еще на первоначальных этапах развития организма дать необходимые рекомендации во избежание предпатологических и патологических изменений.

Целью работы является использование статистических методов для оценки состояния здоровья детей.

Характеристика исследуемых данных

Для проведения экспериментальных исследований использовались данные, предоставленные врачами Сибирского государственного медицинского университета (СибГМУ) и лечебно-оздоровительного центра (ЛОЦ) «Здоровая мама – крепкий малыш».

Объектом исследования являются медицинские показатели детей первого года жизни (от рождения до 12 месяцев).

Организм ребенка находится в процессе созревания и взросления, а это происходит непрерывно в определенной закономерной последовательности.

Первый год жизни ребенка принято условно классифицировать на 2 периода: новорожденный (неонатальный) и грудной.

Для постановки диагноза доступны следующие количественные показатели гормонов и веществ, отвечающих за жизнедеятельность организма, измеренные в 1–12 месяцах: тиреотропный гормон (ТТГ), малоновый диальдегид (mda), трийодтиронин (Т3), тироксин (Т4), кортизол (kor), инсулин (ins), витамин Е (e).

Тиреотропный гормон отвечает за нормальную работу щитовидной железы, стимулирует выработку гормонов щитовидной железы. Малоновый диальдегид используется для прогноза и контроля лечения ишемической болезни сердца, а также широкого спектра других заболеваний. Трийодтиронин и тироксин вырабатываются щитовидной железой. Они требуются для нормального внутриутробного развития таких органов и систем ребенка, как нервная система, сердечно-сосудистая система, половая система, опорно-двигательный аппарат и др. Кортизол – гормон стресса, который помогает организму справляться с отрицательными воздействиями внешней среды, стимулирует работу сердца и концентрирует внимание. Инсулин – гормон поджелудочной железы, который уменьшает количество глюкозы в крови, запасая ее в печени. Понижение уровня инсулина влечет развитие заболевания – диабета.

Бесспорно, гормоны оказывают огромное влияние на организм человека, а именно на все биохимические процессы – рост, развитие, размножение, обмен веществ. Для того чтобы организм мог нормально функционировать, необходимо наличие определенного соотношения гормонов в крови.

Обзор методов и моделей решения задачи

Рассматриваемая задача является задачей распознавания образов. Теория статистических решений исторически является наиболее выгодным математическим инструментом для постановки и решения задач распознавания образов. Начала теории статистических решений

лежат в основе построений алгоритмов распознавания. Эти алгоритмы, на основе опытных данных некоторого набора параметров, которые характеризуют этот объект, и данных, которые описывают классы анализируемых объектов, обеспечивают определение класса, к которому может быть отнесен конкретный объект. Позднее математический аппарат, применяющийся при решении задач распознавания образов, значительно обогатился за счет использования методов теории информации, алгебры логики математического программирования и некоторых других разделов прикладной математики.

К настоящему времени определенно выделились следующие подходы к решению задач распознавания образов: детерминистский [2], статистический [3], лингвистический [2], алгебраический [4], логический [5, 6].

Логистическая регрессия

Под регрессионным анализом понимают исследование влияния одной или нескольких независимых переменных X_1, X_2, …, X_p на зависимую переменную Y [7].

Задача регрессионного анализа заключается в построении математической модели, которая позволяет давать оценку значений зависимой переменной по значениям независимых переменных [8].

Основные цели регрессионного анализа:

- при помощи независимых переменных предсказывать значения зависимой переменной;
- определять вклад отдельных независимых переменных в изменение зависимой переменной.

Общее уравнение регрессии определяется зависимостью [9]:

$$Y = F(X_1, X_2, ..., X_p), \quad (1)$$

где F – неизвестная функция, подлежащая определению;

Y – зависимая переменная;

$X_1, X_2, \ldots, X_p$ – набор независимых переменных;

p – общее количество независимых переменных.

Различают 2 основных вида регрессии [9]:

- линейная;

- нелинейная (гиперболическая, показательная, полиномиальная, степенная, логарифмическая, логистическая).

Для проведения регрессионного анализа в работе используется логистическая регрессия.

Логистическая регрессия – это разновидность множественной регрессии, применяющаяся тогда, когда зависимая переменная может принимать только два взаимоисключающих значения, т.е. является бинарной (дихотомической). Метод бинарной логистической

регрессии позволяет изучить, как зависит дихотомическая переменная от независимых переменных, которые могут быть измерены в разных шкалах. Чаще всего при использовании дихотомических переменных, когда речь идёт о некотором событии, которое может наступить или не наступить (0 – событие не произошло, 1 – событие произошло), бинарная логистическая регрессия позволяет рассчитать вероятность наступления события в зависимости от значения независимых переменных. При этом используется следующее уравнение регрессии:

$$P(Y=1 \mid X_1,\ldots,X_p)=\frac{1}{1+\exp(-(\beta_0+X_1\beta_1+\ldots+X_p\beta_p))} \qquad (2)$$

где Y – зависимая переменная, принимающая значения 0 или 1;

$X_1,\ldots,X_p$ – набор зависимых переменных;

$\beta_0, \ldots,\beta_p$ – логистические регрессионные коэффициенты.

Таким образом, формула (2) описывает вероятность того, что событие наступит, т.е. независимая переменная Y примет значение, равное 1, в зависимости от независимых переменных $X_1,\ldots,X_p$.

Если значение P получится меньше 0,5, то можно считать, что событие не наступит, в противоположном случае можно полагать наступление события.

Зависимость, связывающая вероятность события и величину Y, которая называется логистической кривой, показана на рисунке 1.

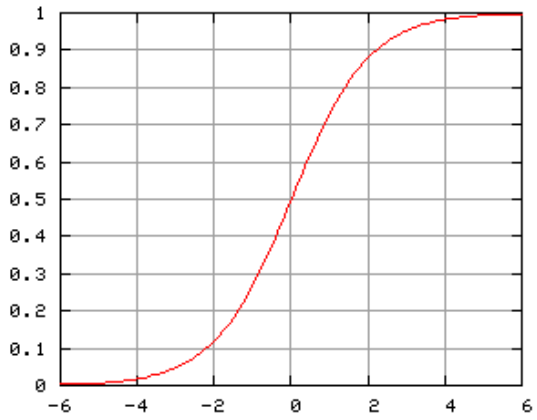

Рис. 1. Логистическая кривая

Из формулы (2) и рисунка 1 можно увидеть, что в этой модели независимо от коэффициентов регрессии $\beta_0, \ldots, \beta_p$ или величин $X_1,\ldots,X_p$, спрогнозированные значения Y всегда будут находиться в диапазоне от 0 до 1.

Для нахождения коэффициентов логистической регрессии используется метод максимального правдоподобия. Суть этого метода заключается в том, что процесс оценки коэффициентов регрессии сводится к максимизации вероятности появления конкретной выборки (при заданных наблюдаемых значениях).

Анализ результатов исследования

Экспериментальные исследования проводились с целью получить результат, отражающий либо наличие заболевания у ребенка, либо его отсутствие, и сравнить этот результат с диагнозом, поставленным врачом.

По результатам обследования 198 детей формируется выборка, которая разделена на 2 класса (диагноза): здоровые (154 детей) и больные (44 ребенка).

Вычисления проводились с использованием статистического пакета SPSS Statistics [10].

С помощью регрессионного анализа было установлено, что только 13 количественных переменных влияют на постановку диагноза: T41, kor1-3, kor6, ins5-12. Остальные переменные, такие как kor4-12 не вошли в уравнение регрессии.

В таблице 1 представлена классификационная таблица, в которой предсказываемые значения зависимой переменной, посчитанные по уравнению регрессии, сравниваются с фактическими наблюдаемыми значениями.

Таблица 1

Таблица классификации

			Предсказанные		
			Состояние		Процент правильных
	Наблюденные		Здоров	Болен	
Шаг 1	Состояние	Здоров	148	6	96,1
		Болен	25	19	43,2
	Общая процентная доля				84,3

Для определения прогнозируемой величины для каждого объекта вычисляется вероятность, на основании которой текущему объекту присваивается одно из двух значений бинарной переменной. В случае если вероятность оказалась менее 0,5, то диагноз – «здоров» (значение переменной «диагноз» равно 0), в противоположном случае – «болен» (значение переменной «диагноз» равно 1). Из таблицы 1 можно увидеть, что для 84,3 **%** объектов исследования предсказанные результаты оказались верными. Это хороший результат, но процент правильного распознавания больных детей очень мал, что на практике может привести к отрицательным результатам.

В Таблице 2 приводится порядок включения независимых переменных в уравнение регрессии на каждом шаге его построения.

Таблица 2

Переменные в уравнении

		B	Стандартное отклонение	Критерий Вальда	Знач.	Exp
Шаг 1[a]	T41	-0,13	0,03	15,12	0	0,87
	kor1	22,41	24,67	0,82	0,36	$5,41*10^{9}$

	kor2	-55,57	61,66	0,81	0,36	0
	kor3	36,69	41,09	0,79	0,37	$8{,}60*10^{15}$
	kor6	-3,51	4,10	0,73	0,39	0,03
	ins5	116,59	78,30	2,21	0,13	$4{,}33*10^{50}$
	ins6	-80,53	141,27	0,32	0,56	0
	ins7	-206,24	81,78	6,37	0,01	0
	ins8	39,23	80,44	0,23	0,62	$1{,}09*10^{17}$
	ins9	179,46	65,78	7,44	0,01	$8{,}68*10^{77}$
	ins10	34,04	78,39	0,18	0,66	$6{,}09*10^{14}$
	ins11	-69,56	77,06	0,81	0,36	0
	ins12	-13,53	50,07	0,07	0,78	0
	Константа	14,34	3,77	14,42	0	$1{,}69*10^{6}$

В таблице используются следующие обозначения:

В – коэффициенты регрессионного уравнения;

Стандартное отклонение – мера изменчивости коэффициентов регрессионного уравнения В;

Критерий Вальда – критерий значимости коэффициентов регрессионного уравнения для соответствующего прогностического параметра. Чем выше значение, тем выше значимость.

Exp – e^{B}.

С учетом найденных коэффициентов, уравнение регрессии запишется в виде формулы:

$$P(Y) = \frac{1}{1+e^{-(14{,}3-0{,}1X_1+22{,}4X_2-55{,}6X_3+36{,}7X_4-3{,}5X_5+116{,}6X_6-80{,}5X_7-206{,}2X_8+39{,}2X_9+179{,}5X_{10}+34X_{11}-69{,}6X_{12}+13{,}5X_{13})}} \quad (3)$$

Для вычисления вероятности того, что ребенок болен, необходимо в уравнение регрессии подставить значения переменных $X_1, X_2, ..., X_{13}$, соответствующие каждому объекту (состоянию ребенка).

Заключение

Логистическая регрессия имеет большое значение и практическое применение в медицине. Метод был апробирован на реальных медицинских данных, предоставленных медицинскими работниками. Качество распознавания, 84,3 % детей был правильно отнесен к соответствующему классу, можно считать приемлемым.

Список литературы

1. Лбов Г.С. Методы обработки разнотипных экспериментальных данных. – Новосибирск: Наука, 1981. – 160 с.

2. А. Фор Восприятие и распознавание образов. – М.: Машиностроение, 1989. – 320 с.

3. Дж. Ту, Р. Гонсалес Принципы распознавания образов. – М.: Мир, 1987. – 431 с.

4. Мазуров В. Д. Математические методы распознавания образов в решении задач планирования и управления. – Свердловск: Урал, 1977. – 148 с.
5. Колесникова С.И., Янковская А.Е. К вопросу вычисления весовых коэффициентов признаков в интеллектуальных системах поддержки принятия решений при большой размерности признакового пространства // Вестник Томского государственного университета. – 2006. – № 18. – С. 223.
6. Гедике А.И. Выявление закономерностей в знаниях и принятие решений в интеллектуальных логико-комбинаторных распознающих системах: дисс. … канд. техн. наук / А. И. Гедике. – Томск, 1998. – 228 с.
7. Елисеева И.И., Юзбашев М.М. Общая теория статистики. – М.: Финансы и статистика, 2004. – 656с.
8. Соколов Г.А., Сагитов Р.В. Введение в регрессионный анализ и планирование регрессионных экспериментов в экономике. – М.: Инфра-М, 2005. – 208 с.
9. Ефимова М.Р., Петрова Е.В., Румянцев В.Н. Общая теория статистики. – М.: Инфра-М, 2004. – 416 с.
10. Официальный русскоязычный сайт SPSS IBM. [Электронный ресурс]. Режим доступа: http://www.predictivesolutions.ru// (дата обращения: 01.04.2015).

Рецензенты:

Берестнева О.Г., д.т.н., профессор, кафедра Прикладной математики, Национальный исследовательский Томский политехнический университет, Институт кибернетики, г. Томск;

Кривоногова Т.С., д.м.н., профессор, Лечебно-оздоровительный центр «Здоровая мама – крепкий малыш», г. Томск.