ОГЛАВЛЕНИЕ

ЮНКЕРОВ
Виктор Иванович
ГРИГОРЬЕВ
Степан Григорьевич
РЕЗВАНЦЕВ
Михаил Владимирович

МАТЕМАТИКО-СТАТИСТИЧЕСКАЯ ОБРАБОТКА ДАННЫХ МЕДИЦИНСКИХ ИССЛЕДОВАНИЙ

Подписано в печать 24.12.2010 г. Формат 60x84 1/16.
Объем 20,0 п.л. Тираж 2000. Зак. 837
Отпечатано с готовых диапозитивов
в ООО «Литография»

ЧАСТЬ II

МНОГОМЕРНЫЕ МЕТОДЫ АНАЛИЗА ДАННЫХ И МОДЕЛИРОВАНИЯ МЕДИЦИНСКИХ СИСТЕМ

Вершиной медико-биологического исследования, закономерным его финалом очень часто является создание модели изучаемого явления, процесса. Наиболее объективными моделями являются модели, для создания которых используются математические методы.

Во второй части книги детально рассматриваются вопросы подготовки данных исследования к обработке с помощью многомерных методов математико-статистического моделирования, последовательность разработки, оценки качества и эксплуатации моделей на примерах материалов реальных исследований.

Глава 5. МНОГОФАКТОРНЫЙ КОРРЕЛЯЦИОННЫЙ И РЕГРЕССИОННЫЙ АНАЛИЗ ДАННЫХ МЕДИЦИНСКИХ ИССЛЕДОВАНИЙ

Задачи исследования сложных систем

Известно, что объекты исследования в медицине представляют собой сложные вероятностные (стохастические) системы. Сложные системы функционируют при воздействии на них множества входных факторов. Часть из них является контролируемыми X1, X2 ,..., Xk, измеряемыми количественно или оцениваемыми в баллах. Другая часть входных факторов относится к группе неконтролируемых, случайных факторов; они не поддаются измерению, но оказывают воздействие на систему, результатом которого является случайность ее функционирования. Состояние системы характеризуется множеством выходных параметров Y1, Y2 ,..., Yl, которые также измеряются количественно или в баллах и представляют собой случайные величины, следующие нормальному или иному закону распределения с соответствующими числовыми характеристиками. Наилучшие результаты многомерного статистического анализа данных медицинских исследований получают тогда, когда распределение входных факторов и выходных параметров нормальное или близкое к нему.

Наблюдавшиеся значения k факторов и l параметров для n объектов сводятся в матрицу наблюдений размером n×(k+l). По матрице наблюдений на ПК с помощью ППП Statistica 5.0 проводится:

- статистическое описание переменных;
- корреляционный анализ;
- канонический корреляционный анализ;
- регрессионный анализ.

В результате статистического описания устанавливают законы распределения переменных и определяют их числовые характеристики, строят графики основных зависимостей между факторами и параметрами.

Корреляционный анализ обеспечивает оценку связей всех переменных попарно.

Канонический корреляционный анализ дает оценку связи всего множества входных факторов со всеми выходными параметрами в совокупности.

На основе канонического корреляционного анализа можно судить о достаточности связи входных факторов, включенных в матрицу наблюдений, и выходных параметров, характеризующих состояние системы.

Моделирование каждого выходного параметра методами регрессионного анализа дает возможность построить линейные или нелинейные модели. Модели используются для решения основных задач системного анализа:

– изучения характера изменения выходных параметров при изменении входных факторов;

– оценки степени влияния факторов на параметры;

– прогнозирования параметров при заданных значениях факторов;

– поиска оптимальных уровней факторов для получения требуемых значений параметров;

– оценки информативности параметров при заданной совокупности воздействующих факторов.

Требования к базе данных для многомерного статистического анализа

Матрица наблюдений с n строками по числу наблюдавшихся объектов в выборке и (k+l) столбцами по числу наблюдавшихся k входных факторов и l выходных параметров должна содержать только количественные данные в натуральных единицах измерения или баллах.

При отсутствии данных по какому либо признаку его заменяют средним значением признака для всей выборки, хотя это приводит к искажению исходной информации. Следует также иметь в виду, что некоторые статистические пакеты не рассчитывают корреляционной матрицы, в случае если число переменных превышает число наблюдений. Надежное решение можно получить, если в матрицах наблюдений число строк n в 3-5 раз превышает число столбцов (k+l).

Все данные должны быть тщательно проверены: устраняются грубые ошибки, исключаются явно анормальные результаты наблюдения.

Выборка должна быть безусловно репрезентативной по отношению к исследуемой генеральной совокупности.

В соответствии с целью и задачами исследования в матрицу необходимо ввести дополнительные столбцы с группировочными признаками, например, группировочный признак G1 - контрольная группа с

кодом 1, опытная группа с кодом 2; группировочный признак пола G2 – мужчины с кодом 1, женщины с кодом 2 и т.п.

Задачи и содержание многомерного корреляционного анализа

Многомерный корреляционный анализ проводится для количественной оценки направления, силы и значимости линейной связи между всеми переменными базы данных попарно. Такая связь характеризуется коэффициентом корреляции Пирсона.

В результате решения по опциям Descriptive statistics и Correlations на экран выводятся следующие результаты:

- таблица числовых характеристик переменных;
- корреляционная матрица, содержащая коэффициенты корреляции и уровни их значимости для всех пар переменных.

По таблице числовых характеристик анализируется соответствие распределений каждой переменной нормальному закону.

По корреляционной матрице, представляющей собой квадратную симметричную таблицу с размером (k+l)×(k+l), судят о направлении, силе и значимости корреляционной связи переменных попарно, в особенности о связи входных факторов с выходными параметрами.

Вариант расчета числовых характеристик переменных и корреляционной матрицы, а также интерпретации результатов приведен в примере 5.1.

Назначение и содержание канонического корреляционного анализа

Канонический корреляционный анализ предназначен для изучения связи между входными факторами и выходными параметрами в их совокупности.

Для проведения канонического корреляционного анализа в исходной матрице наблюдений с размерами n×(k+l), где n - число наблюдавшихся объектов, k - число входных факторов и l - число выходных параметров, выделяют две группы переменных:

1. Left set - группа выходных параметров;
2. Right set - группа входных факторов.

Алгоритмом предусмотрено.

1.Определение ограниченного числа канонических переменных, обобщающих выходные параметры 1-ой группы, и такого же коли-

чества канонических переменных, обобщающих входные факторы 2-й группы. При этом первая пара канонических переменных обобщает наибольшую часть дисперсии переменных, вторая пара – большую долю из оставшейся части дисперсии и т.д. Количество пар канонических переменных зависит от размерности матрицы наблюдений. Практика показала, что 2-3 пар канонических переменных достаточно для надежного представления всей совокупности переменных.

2.Формирование полей рассеяния объектов в координатах первой, второй, третьей пары канонических переменных для 1-ой и 2-ой группы, а после их формирования - расчет канонических коэффициентов корреляции: Can r1 - по паре первых, Can r2 - по паре вторых, Can r3 - по паре третьих канонических переменных.

По величине канонических коэффициентов корреляции судят о силе связи между совокупностями входных факторов и выходных параметров.

Квадраты коэффициентов (Eigenvalue) характеризуют степень детерминации совокупности параметров совокупностью факторов для каждой пары канонических переменных.

Значимость канонических коэффициентов корреляции и детерминации оценивают по χ^2 - критерию Пирсона. Коэффициенты считают значимыми при вероятности равной и более 0,95 или при уровне значимости $p \le 0,05$;

3.Расчет факторной структуры канонических переменных (Factor structure), т.е. коэффициентов корреляции, характеризующих направление и силу корреляционной связи канонических переменных с наблюдавшимися входными факторами и выходными параметрами. В результате дается оценка важности входных факторов и информативности выходных параметров.

Такой анализ на начальном этапе исследования позволяет оценить достаточность связи между входными факторами и выходными параметрами с целью построения для них надежных моделей, а также выделить наиболее значимые факторы и информативные параметры-отклики на воздействия.

Назначение и содержание многомерного регрессионного анализа. Построение линейного уравнения регрессии

Многомерный регрессионный анализ (Multiple Regression) применяется для построения уравнения регрессии для параметра Y в зависимости от факторов X1 – Xk. Модель может быть линейной и нелинейной. Наиболее простой, содержащей только линейные эффекты факторов, является линейная модель.

$$\hat{y} = b_0 + b_1x_1 + b_2x_2 + ... + b_kx_k, \qquad (5.1)$$

где $\hat{y}$ - прогнозируемое значение выходного параметра;

b_0 – свободный член;

$b_1 b_2 ... b_k$ - коэффициенты регрессии;

$x_1 x_2 ... x_k$ - возможные значения факторов X1,X2,…Xk;

$b_1 x_1, b_2 x_2, ... b_k x_k$ – линейные эффекты факторов.

Коэффициенты модели получают методом наименьших квадратов по исходной матрице наблюдений $n \times (k+1)$, где: n – число строк в матрице, равное числу наблюдаемых объектов, k+1 – число столбцов, равное числу независимых переменных (k факторов X1,X2,…Xk) и одной зависимой переменной (моделируемый параметр Y).

Значимость коэффициентов оценивают по t-критерию Стьюдента. При построении модели в ответственных случаях, например, для прогноза параметра Y, в модели сохраняют только значимые коэффициенты с доверительной вероятностью больше или равной 0,95 или с уровнем значимостью $p \le 0,05$. В поисковом исследовании с целью изучения характера изменения параметра Y, при изменении факторов и степени влияния их на параметр, допускают сохранение в модели эффектов с коэффициентами, при уровне их значимости $p \le 0,30$ (доверительной вероятностью равной или больше 0,70).

Стандартный алгоритм регрессионного анализа предусматривает расчет:

– числовых характеристик переменных;

– корреляционной матрицы;

– коэффициентов модели с оценками их значимости;

– результатов дисперсионного анализа модели и оценки коэффициентов множественной корреляции и детерминации, средней квадратичной ошибки прогноза параметра Y по модели;

– графика линии регрессии с указанием 95%-го доверительного интервала для прогноза значений параметра Y.

Вариант многофакторного регрессионного анализа с целью построения линейного уравнения регрессии дан в примере 5.1.

Сущность пошагового регрессионного анализа

Стандартный алгоритм многомерного регрессионного анализа обеспечивает получение коэффициентов модели для всех независимых переменных X1 – Xk. Исходя из уровней значимости, исследователь решает, какие коэффициенты должны быть включены в модель, как значимые, достоверные. Для автоматического включения значимых эффектов в модель и исключения незначимых предлагается пошаговый регрессионный анализ в двух вариантах:

– Forward – поочередное включение в модель наиболее значимых эффектов;

– Backward – поочередное исключение из полной модели наименее значимых эффектов.

Отбор значимых эффектов реализуется по критерию F – Фишера.

В ответственных исследованиях для получения коэффициентов с уровнем значимости $p \le 0{,}05$ задается значение критерия F=3-4. В поисковых исследованиях значение F=1-2 обеспечивает включение в модель коэффициентов с уровнем значимости $p \le 0{,}30$.

В примере 5.1 методом пошагового регрессионного анализа при F=1 получены коэффициенты модели, приведенные в машинограмме 5.3. Исключенным из модели оказался эффект фактора X1.

Дисперсионный анализ и оценка эффективности модели

Дисперсионный анализ модели выполняется для оценки ее эффективности. Под эффективностью модели понимают ее информативность и значимость (достоверность). Модель считают информативной, если ее коэффициент детерминации $R^2>0{,}5$; значимой, достоверной при уровне значимости по F – критерию $p \le 0{,}05$ (достоверности $\ge 0{,}95$).

В примере 5.1 в машинограмме 5.4. дан дисперсионный анализ модели, из которого следует, что модель достоверна (p=1,93E–0,5); из машинограммы 5.3. видно, что модель информативна, т.к. $R^2=0{,}819$ (значительно больше 0,5), а стандартная ошибка прогноза возможных значений параметра S_0=26,528 нКи/кг и среднеожидаемых значений параметра:

$$m_{\hat{y}} = \frac{S_0}{\sqrt{n}} = \frac{26{,}528}{\sqrt{20}} = 5{,}9 \text{нКи / кг}.$$

Такую модель можно применять для решения задач исследования.

Оценка степени влияния факторов на моделируемый параметр

Степень влияния факторов на параметр Y рассчитывается по величине стандартизованных коэффициентов регрессии BETA по формуле (5.2.):

$$K_j = \frac{100 \times BETA_j}{\sum_{(j)} |BETA_j|} \times R^2, \text{в\%}. \qquad (5.2)$$

По данным примера 5.1 оценка степени влияния факторов X5, X3, X2 и X4 дана в таблице 5.2.

Прогноз по модели и оценка его точности и надежности

Прогноз среднеожидаемых значений параметра может быть дан по модели (5.1.) или графику линии регрессии (рис.5.1).

Точность и надежность прогнозируемого значением параметра оценивается 95%-м доверительным интервалом:

$$Y = \hat{y} \pm t_{95} \times m_{\hat{y}}. \qquad (5.3)$$

В примере 5.1 для заданных значений факторов получен прогноз y=68,4 нКи/кг и его 95%-й доверительный интервал от 54 до 83 нКи/кг.

Достаточно большой доверительный интервал характерен для поискового исследования из-за приблизительной оценки значений факторов и параметра в матрице наблюдений.

Особенности нелинейного регрессионного анализа

При нелинейной зависимости моделируемого параметра от входных факторов линейное уравнение регрессии (5.1.) может оказаться неинформативным. В таких случаях следует построить нелинейное уравнение регрессии по операции Fixed non-linear модуля Multiple Regression или по модулю Nonlinear Estimation.

Вид нелинейного уравнения можно предположить, проведя предварительно графическое исследование зависимости параметра Y от факторов X1 – Xk путем построения соответствующих графиков. Наибо-

Глава 10. ЛОГИСТИЧЕСКАЯ РЕГРЕССИЯ

Назначение и содержание логистической регрессии

Построение моделей для показателей состояния многомерных медицинских систем в зависимости от воздействующих на них факторов является важной задачей статистического анализа, выполняемого исследователями с применением современной информационной технологии.

По модели решают основные задачи исследования, среди которых изучение характера изменения показателя при изменении действующих на систему факторов; оценка степени влияния факторов на величину показателя-отклика; прогнозирование показателя-отклика для заданных уровней факторов; определение оптимальных уровней факторов для получения требуемых или желаемых значений показателей состояния системы. Построение таких моделей проводится на основе выборочного наблюдения, по результатам которого формируется исходная база данных, представляющая собой матрицу наблюдений с числом строк, равным числу наблюдавшихся объектов и числом столбцов, равным числу контролируемых факторов и моделируемого показателя-отклика на воздействующие факторы.

В условиях количественного определения факторов и показателя-отклика для построения модели показателя – уравнения регрессии – применяют многомерный регрессионный анализ. Коэффициенты модели при этом определяются методом наименьших квадратов. В основу метода наименьших квадратов заложен принцип минимизации суммы квадратов отклонений прогнозируемых значений показателя по модели от наблюдавшихся значений в выборке.

В условиях наблюдения качественных оценок показателя – отклика всего на двух уровнях, например, выживание больного при тяжелой травме (код 1) и летальный исход (код 0), для построения модели вероятности благоприятного исхода применяют логистическую регрессию, представляющую собой нелинейную функцию распределения вероятностей. Коэффициенты уравнения логистической регрессии определяются методом максимального правдоподобия, в основу которого положен принцип максимизации вероятности соответствия, адекватности прогнозируемых по моделям уровней показателя–отклика с наблюдавшимися значениями этого показателя в выборке.

Для случая, когда исследуется некоторый положительный эффект при воздействии на объект только одного фактора, например, в эксперименте «доза-эффект», уравнение логистической регрессии имеет вид:

$$\hat{y} = \frac{\exp(b_0 + bx)}{1 + \exp(b_0 + bx)}, \qquad (10.1)$$

где $\hat{y}$ - вероятность положительного эффекта ($0 \le \hat{y} \le 1$);

b_0 – константа;

b – коэффициент фактора X;

x – текущее значение фактора X.

Кривая функции логистической регрессии, соответствующая уравнению (10.1), показана на рисунке 10.1. Она представляет собой кривую, непрерывно возрастающую от 0 до 1.

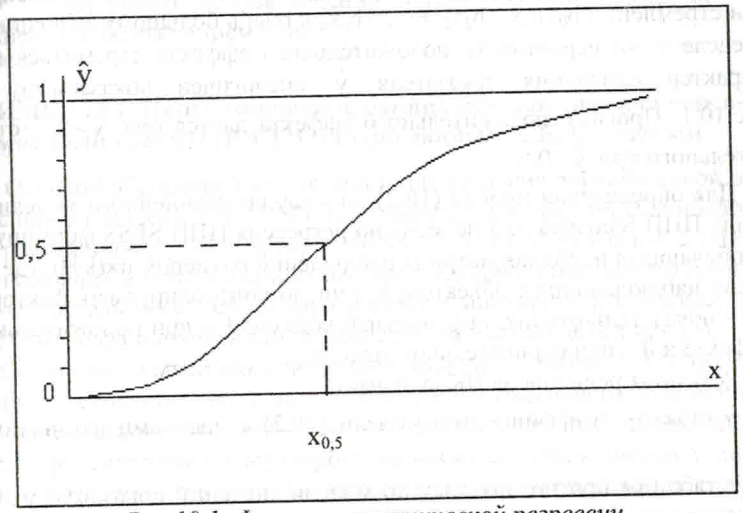

Рис.10.1. Функция логистической регрессии.

При малых дозах фактора X вероятность положительного эффекта незначительна; при больших дозах, например, при $x > x_{0,5}$ вероятность положительного эффекта возрастает от 0,5 до 1.

При прогнозе вероятности положительного эффекта по (10.1) принимают:

-положительный эффект при $\hat{y} > 0,5$;

-отрицательный результат при $\hat{y} \le 0{,}5$.

Для случая, когда исследуется положительный эффект при воздействии на объект множества контролируемых факторов X1, X2, … Xk, уравнение логистической регрессии принимает вид:

$$\hat{y} = \frac{\exp(b_0 + b_1x_1 + b_2x_2 + ... + b_kx_k)}{1 + \exp(b_0 + b_1x_1 + b_2x_2 + ... + b_kx_k)}, \qquad (10.2)$$

где $\hat{y}$ - вероятность положительного эффекта ($0 \le \hat{y} \le 1$);

b_0 – константа;

$b_1, b_2, \ldots b_k$ – коэффициенты X1, X2, … Xk факторов;

$x_1, x_2, \ldots x_k$ – текущие значения k факторов.

Из (10.2) следует, что при стремлении величины показателя экспоненты $b_0+b_1x_1+b_2x_2+ \ldots + b_kx_k$ к очень малому значению (в пределе $-\infty$) вероятность положительного эффекта стремиться к 0, и наоборот, при стремлении $b_0+b_1x_1+b_2x_2 + \ldots +b_kx_k$ к очень большому значению (в пределе $+\infty$) вероятность положительного эффекта стремиться к 1. Характер изменения показателя $\hat{y}$ аналогичен показанному на рис.10.1. Прогноз, положительного эффекта дается при $\hat{y} > 0{,}5$, отрицательного при $\hat{y} \le 0{,}5$.

Для определения модели (10.2) по модулю нелинейного моделирования ППП Statistica или по модулю регрессия ППП SPSS формируется обучающая исходная матрица наблюдений размером $n \times (k+1)$, где n – число наблюдавшихся объектов; k – число контролируемых факторов и 1 – показатель-отклик, выражаемый кодами: 1 – при положительном эффекте и 0 – при отрицательном исходе.

Результат решения на ПК включает:

– таблицу коэффициентов модели (10.2) с оценками их значимости;

– таблицы прогнозируемых по модели значений показателя $\hat{y}$, наблюдавшихся значений y и разностей $\hat{y}$-y для всех объектов обучающей матрицы наблюдений;

– классификационную матрицу, характеризующую абсолютные величины и относительные частоты правильных прогнозов положительных и отрицательных эффектов по модели;

– уровень значимости модели по критерию хи-квадрат;

* оценку соответствия распределения остатков нормальному закону.

Модель признается значимой при уровне значимости $p \le 0{,}05$ (достоверности $1-p \ge 0{,}95$).

Следует отметить, что в процедуре Logistic Regression ППП SPSS возможен пошаговый отбор значимых коэффициентов для включения в модель 10.2. С этой целью задается критическое значение критерия Фишера $F \ge 4$, обеспечивающее уровень значимости коэффициентов $p \le 0{,}05$.

Примеры применения логистической регрессии:

* в экспериментах «доза-эффект»;
* в медицинской диагностике (есть заболевание или нет заболевания у обследуемого);
* в профотборе (пригодность или непригодность кандидата);
* при оценке исхода лечения больного (выжил, умер; без осложнения, с осложнением и др.).
*

ПРИМЕР 10.1. Прогнозирование ранних исходов тяжелой черепно-мозговой травмы (РИ ТЧМТ) (по данным Н.Б. Клименко).

Основой обучающей информации для создания логистической регрессионной модели РИ ТЧМТ по признаку выписан из стационара - умер в стационаре стали истории болезней 300 пострадавших с ТЧМТ, поступивших на стационарное лечение в клинику Российского научно-исследовательского нейрохирургического института им. профессора А.Л. Поленова и прошедших лечение до определившегося исхода.

Основной задачей моделирования является прогноз РИ ТЧМТ по данным первичного врачебного осмотра пострадавшего в приемном отделении стационара или в любых других условиях, например, на месте происшествия на автодороге, на производстве, в боевых условиях и т.п. По сути, такая модель является экспресс-прогнозом, т.к. строится на основании минимально достаточного числа наиболее простых и всегда исследуемых неврологических симптомов и синдромов, не требующих высокой квалификации врачебного персонала и не включающих применения специальных дополнительных инструментальных методов исследования.

В качестве прогнозируемого показателя-отклика определен исход травмы (благоприятный - больной выписан из стационара - 1 и неблагоприятный - больной умер в стационаре - 0).

В качестве признаков, предшествующих исходу травмы, и включаемых в модель как независимые факторы-причины, определена совокупность клинических признаков, достоверно связанных с исходами и определяемых у больных на ранних этапах оказания медицинской помощи. В исходную обучающую матрицу было включено 59 признаков, получаемых анамнестически и с помощью непосредственного врачебного обследования и регистрируемых в приемном отделении. После логического анализа и оценки связей исходных данных с помощью корреляционного анализа для дальнейшего исследования в обучающей матрице осталось 25 признаков, которые имели сильную ($r>0,7$) или умеренную ($0,27<r<0,7$) и статистически значимую ($p<0,05$) корреляционную связь с РИ ТЧМТ.

Решение задачи логистического регрессионного анализа может быть реализовано с помощью процедуры Logistic Regression из пакетов прикладных программ по статистической обработке данных Statistica или SPSS. По нашему мнению, преимущество следует отдать ППП SPSS, так как он обеспечивает пошаговый отбор в модель статистически значимых факторов с заданным порогом значимости.

По итогам расчетов с помощью модуля Logistic Regression ППП SPSS, в модель включены 5 признаков, обладающих статистической надежностью не менее 80%. Перечень этих признаков и их коэффициенты приведены в табл.10.1.

Полученная методом логистического регрессионного анализа статистически значимая ($p<0,0001$) модель, имеет вид:

$$\hat{y} = \exp(7,54-0,59x_1-0,26x_2-0,79x_3+0,56x_4-1,12x_5)/ \quad (10.3)$$
$$(1 + \exp(7,54-0,59x_1-0,26x_2-0,79x_3+0,56x_4-1,12x_5)).$$

Расчеты прогноза по этой модели могут быть произведены на ПЭВМ или на программируемом микрокалькуляторе.

Любая синтезированная модель, логистическая в том числе, требует подтверждения - на сколько она соответствует наблюдавшимся данным. С этой целью используются:

* классификация данных обучающей информации с помощью полученной модели и оценка соответствия этой классификации с наблюдаемой в опыте (табл. 10.2);
* сравнение опытных и прогнозируемых значений для каждого конкретного наблюдения (табл. 10.3);
* оценка остатков (разности наблюдаемых величин и величин прогнозируемых с помощью модели) (рис. 10.2-10.3).

Таблица 10.1

Признаки, включенные в логистическую регрессионную модель прогноза РИ ТЧМТ

№ пп	Наименования и градации симптомов	Коды	Коэффициенты модели	Уровень значимости, p
1	Возраст: 15-24 года – 1, 25-34 года – 2, 35-44 года – 3, 45-54 года – 4, 55-64 года – 5, 65-74 года – 6, 75 лет и старше – 7.	X1	-0,59	0,000
2	Систолическое АД: 110-140 мм рт.ст. – 1, 141-180 мм рт.ст. – 2, 90-109 мм рт.ст. – 3, 181мм рт.ст. и более – 4, 89 мм рт.ст. и менее – 5.	X2	-0,26	0,059
3	Уровень сознания: ясное – 1, легкое оглушение – 2, умеренное оглушение – 3, сопор – 4, кома I – 5, кома II – 6, кома III – 7.	X3	-0,79	0,000
4	Окулоцефалический рефлекс: отсутствует – 1, вызывается – 2.	X4	0,56	0,179
5	Иннервация зрачков: не нарушена – 1, нарушена – 2.	X5	-1,12	0,072
	Константа.		7,54	0,000

Результаты классификации исходов, полученные с помощью логистической регрессионной модели по данным обучающей информации,