УДК 430.6 : 519.92

ПРОГРАММА ПОСТРОЕНИЯ И АНАЛИЗА РЕГРЕССИОННОЙ МОДЕЛИ ДЛЯ РАСПОЗНАВАНИЯ СПОСОБА ТРАВМИРУЮЩЕГО ВОЗДЕЙСТВИЯ В РЕЗУЛЬТАТЕ ДТП

В. А. Голуб, Н. В. Огаркова, А. С. Попов

Воронежский государственный университет

Описывается программное обеспечение, предназначенное для оценки эффективности использования регрессионного метода в судебно-медицинской практике, с целью распознавания вида травмирующего воздействия, возникшего вследствие дорожно-транспортного происшествия, по признакам и характеру разрушения длинных трубчатых костей. Приводятся примеры работы программы.

1. ВВЕДЕНИЕ

Определение способа травмирующего воздействия по признакам и характеру разрушения длинных трубчатых костей является актуальной задачей судебной медицины [1, 2]. Одним из подходов к решению такой задачи является использование регрессионного метода анализа признаков разрушения. Исследование возможности использования этого метода в судебно-медицинской практике, а также оценка его эффективности требует разработки соответствующего программного комплекса, который должен обеспечивать построение математической модели и оценку ее эффективности.

В статье рассматривается программа, предназначенная для построения регрессионной модели и оценки ее качества, с целью исследования возможности использования такой модели для решения задачи определения вида травмирующего воздействия по признакам разрушения длинных трубчатых костей в судебно-медицинской практике.

2. РЕГРЕССИОННАЯ МОДЕЛЬ

С помощью регрессионного анализа может быть исследована зависимость одного признака (результирующего) от набора независимых (факторных) признаков [3]. Разделение признаков на результирующий и факторные осуществляется на основе содержательных представлений об изучаемом процессе. Обычно рассматриваются количественные признаки, но допускается и использование дихотомических признаков, принимающих лишь два значения, например, 0 и 1. Именно такая ситуация имеет место в рассматриваемой задаче, так как распознаваемое травмирующее воздействие может относиться к одному из двух возможных видов: «Удар» или «Давление» [1, 2].

В таких случаях применима логистическая регрессия [3], представляющая разновидность множественной регрессии, общее назначение которой состоит в анализе связи между несколькими независимыми переменными (факторными признаками) $X_1,\ldots, X_m$, называемыми также регрессорами или предикторами, и бинарной зависимой переменной, принимающей только два значения, например, 0 и 1. В рассматриваемой задаче определение вида травмирующего воздействия в этом случае эквивалентно определению того, что случайная величина Y, представляющая зависимую переменную, примет значение 1. Вероятность $P\,(Y=1)$ в этом случае является вероятностью наступления прогнозируемого исхода.

Функция распределения $F(y)$ и плотность распределения вероятностей $f(y)$ случайной величины, имеющей логистическое распределение, представляются следующими соотношениями:

$$F\,(y)=\frac{e^{y}}{1+e^{y}}=\frac{1}{1+e^{-y}},-\infty<y<\infty \qquad (1)$$

$$f(y)=\frac{e^{y}}{(1+e^{y})^{2}},\ -\infty<y<\infty\,. \qquad (2)$$

Представим множество m анализируемых признаков как вектор значений независимых переменных $\mathbf{x}=(x_0, x_1, x_2, .., x_m)$, где $x_0=1$, а множество коэффициентов регрессии как вектор $\mathbf{a}=(a_0, a_1, .a_2, ., a_m)$.

Рассмотрим регрессионную модель с биномиальной зависимой переменной и логистическим распределением — *логит*.

Для такой модели вероятность того, что случайная величина Y, значения которой $y = \mathbf{ax}$, примет единичное значение равна

$$F(\mathbf{ax}) = \frac{e^{\mathbf{ax}}}{1+e^{\mathbf{ax}}} = \frac{1}{1+e^{-\mathbf{xa}}} = P\,(Y=1) \qquad (3)$$

Для расчета коэффициентов логистической регрессии может использоваться, например, метод наименьших квадратов [4] или метод максимального правдоподобия [5].

3. ОЦЕНКА КАЧЕСТВА МОДЕЛИ

Пусть n — количество рассматриваемых случаев. Под очередным случаем будем понимать пару значений $(\mathbf{x}_i, y_i)$, где $\mathbf{x}_i = (x_0, x_1, x_2, .., x_m)$ вектор значений независимых переменных, а y_i фактическое значение результирующего признака i-го случая. Обозначим через $\hat{y}_i$ расчетные значения для i-го случая, где $\hat{y}_i$.определяется по формуле:

$$\hat{y}_i = F\left(\mathbf{ax}_i\right) = \frac{1}{1+e^{-\mathbf{x}_i\mathbf{a}}}, \qquad (4)$$

где $\mathbf{a} = (a_0, a_1, .a_2, ., a_m)$ — коэффициенты регрессии. Через $\overline{y}$ обозначим среднее фактических значений результирующего признака:

$$\overline{y} = \frac{\sum_{i=1}^{n} y_i}{n}. \qquad (5)$$

Оценка качества построенной модели осуществляется по следующим показателям:

1) количество правильно/неправильно классифицированных случаев;

2) значение среднеквадратической ошибки, которое определяется формулой

$$E = \sqrt{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}\ ; \qquad (6)$$

3) значение множественного коэффициента корреляции R, определяемого формулой

$$R = 1 - \frac{Q_{\text{ост}}}{Q_{\text{общ}}} = \frac{Q_{\text{объясн}}}{Q_{\text{общ}}}, \qquad (7)$$

где $Q_{\text{ост}}$ — остаточная сумма квадратов (сумма квадратов отклонений фактических значений от расчетных)

$$Q_{\text{ост}} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2, \qquad (8)$$

$Q_{\text{объясн}}$ — объясненная сумма квадратов (сумма квадратов отклонений расчетных значений от среднего)

$$Q_{\text{объясн}} = \sum_{i=1}^{n} (\hat{y}_i - \overline{y})^2, \qquad (9)$$

а $Q_{\text{общ}}$ — общая сумма квадратов [6]

$$Q_{\text{общ}} = \sum_{i=1}^{n} (y_i - \overline{y})^2.. \qquad (10)$$

4) ROC-кривая (Receiver Operator Characteristic) [3].

ROC-кривая — кривая, которая наиболее часто используется для представления результатов бинарной классификации. Поскольку рассматривается случай, когда имеется два класса — «Удар» и «Давление», один из них называется классом с положительными исходами, второй — с отрицательными исходами, причем, какой исход считать положительным, а какой — отрицательным, зависит от конкретной задачи. Например, если при анализе признаков разрушения кости вследствие ДТП травмирующее воздействие определяется как удар, то именно класс «Удар» может считаться положительным исходом, тогда как класс «Давление» будет считаться отрицательным исходом.

ROC-кривая показывает зависимость количества верно классифицированных положительных примеров от количества неверно классифицированных отрицательных примеров. Сравнение различных регрессионных моделей с целью определения лучшей на основе ROC-кривых осуществляется путем оценки площади под кривыми следующим образом: чем больше площадь под ROC-кривой, тем лучше соответствующая модель [3].

4. ВЫБОР ФАКТОРНЫХ ПРИЗНАКОВ ДЛЯ ЛОГИСТИЧЕСКОЙ РЕГРЕССИИ

При решении задачи определения вида травмирующего воздействия в качестве факторных признаков, могут использоваться различные признаки разрушения костной ткани [1, 2].

Для определения того, какие именно признаки наиболее целесообразно использовать для построения математической модели процедуры распознавания вида травмирующего воздействия, возможны два подхода:

1) экспертное определение набора признаков, которые будут рассматриваться в качестве факторных признаков, а затем использованы для расчета коэффициентов модели и оценки ее качества;

2) пошаговый отбор, при котором отбираются только те признаки, которые обеспечивают максимальное значение множественного коэф-

фициента корреляции: признак включается в уравнение модели только в том случае, если его включение существенно увеличивает значение множественного коэффициента корреляции. Такой поход позволяет последовательно отбирать только те факторы, которые оказывают существенное влияние на результирующий признак даже в условиях мультиколлинеарности системы признаков. При этом первым в уравнение включается фактор, имеющий наибольшую корреляцию с Y, вторым включается фактор, который в паре с первым из отобранных дает максимальное значение множественного коэффициента корреляции, и т.д. Существенно, что на каждом шаге значение множественного коэффициента увеличивается на некоторое значение, на основе которого можно определить вклад каждого отобранного фактора в регрессионную модель.

5. ПРОГРАММА ПОСТРОЕНИЯ РЕГРЕССИОННОЙ МОДЕЛИ И ОЦЕНКИ ЕЕ КАЧЕСТВА

Для расчета коэффициентов $\mathbf{a} = (a_0, a_1, .a_2, ., a_m)$ регрессионной модели (5) и оценки ее качества в соответствии с описанными выше показателями разработана специальная программа, примеры работы которой представлены на рис. 1 и рис. 2.

На рис. 1 представлено окно, в котором отражены коэффициенты построенного регрессионного уравнения, числовые характеристики для оценки качества модели, ROC-кривая, а также оценки, полученные для каждого отдельного случая. При этом пользователю здесь же предоставляется возможность изменить метод расчета коэффициентов регрессии, что значительно экономит время работы.

На рис. 2 показано окно, используемое в случае пошагового отбора признаков. В левой части окна отображаются включенные в модель признаки и значение коэффициента множественной корреляции R при их отборе. В правой части окна отображаются показатели качества модели (среднеквадратическая ошибка, количество правильно классифицированных случаев, количество неправильно классифицированных и коэффициент множественной корреляции R), в случае включения в нее ранее не использовавшегося признака.

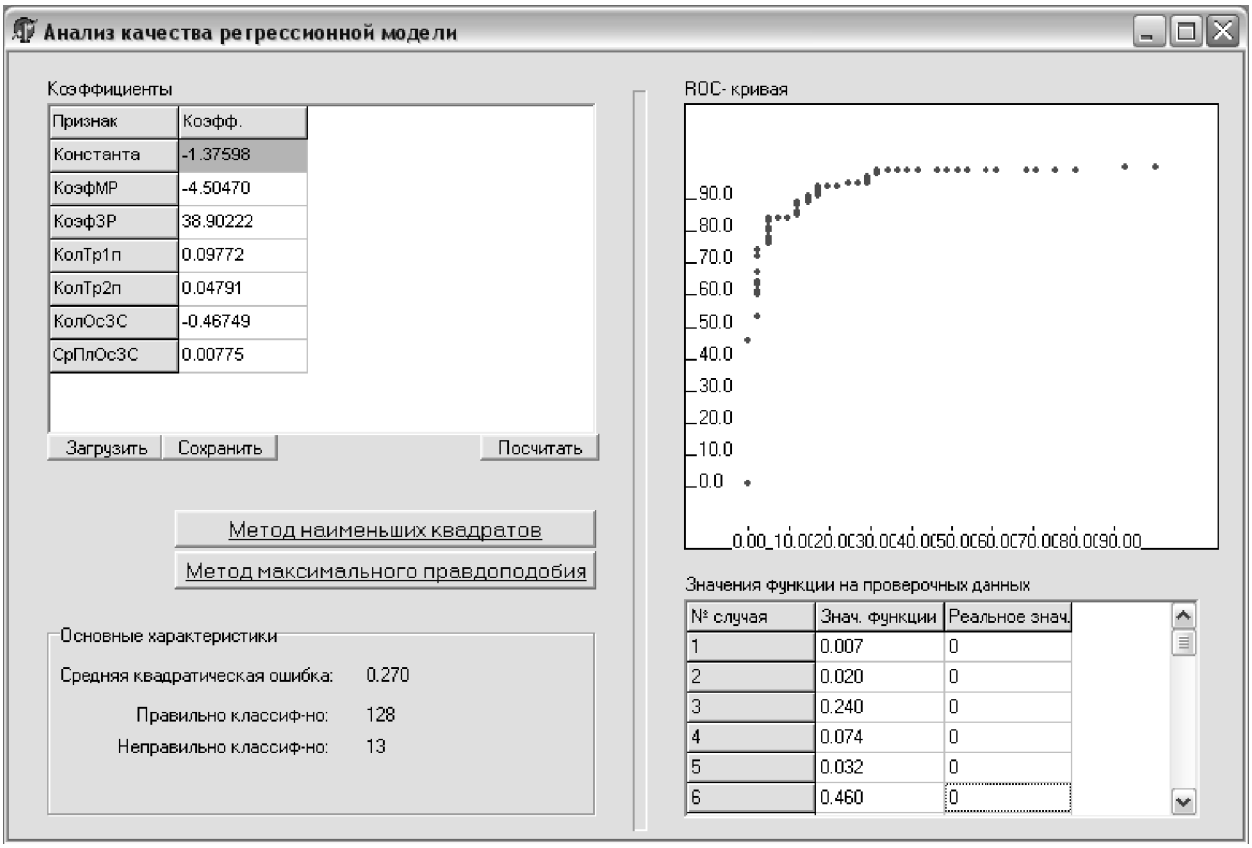

Рис. 1. Окно программы «Анализ качества регрессионной модели»

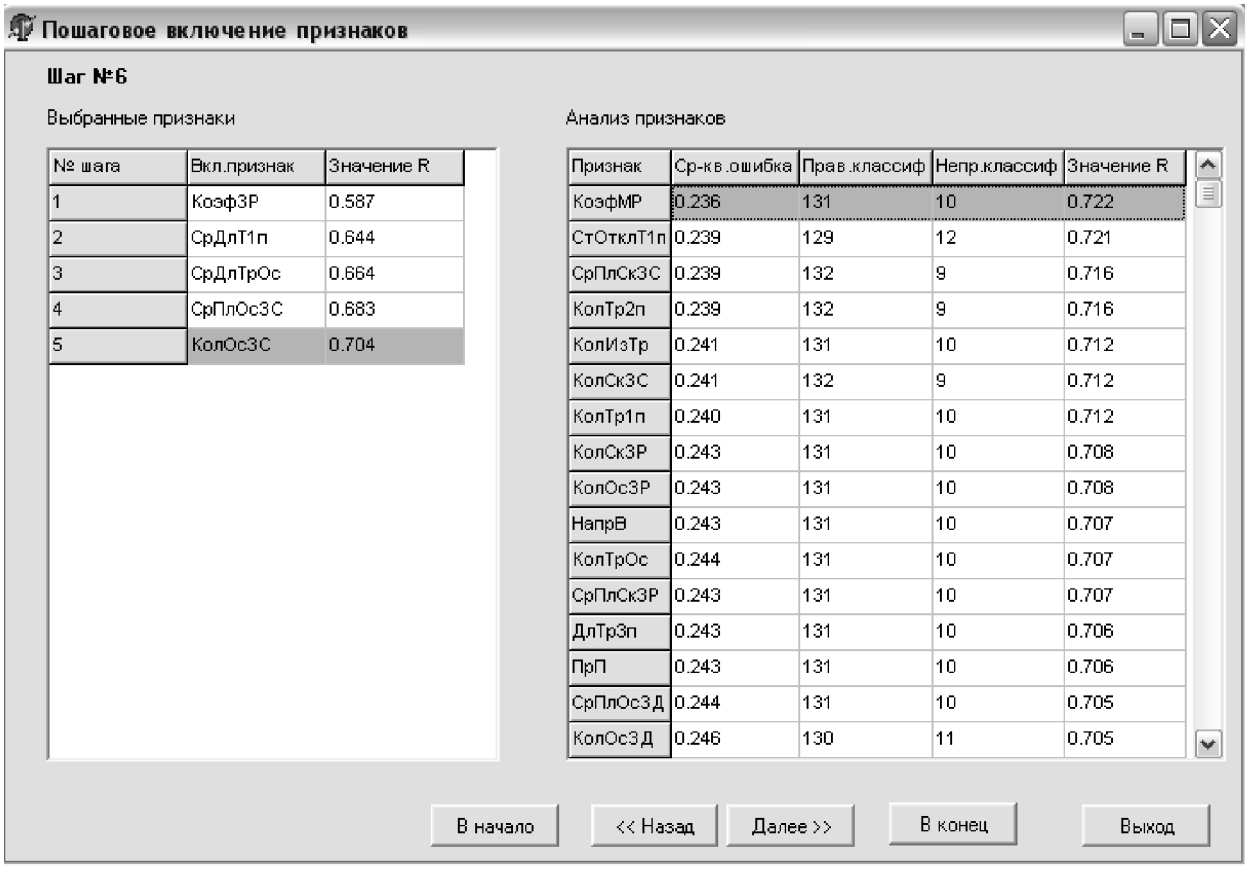

Рис. 2. Окно программы «Пошаговое включение признаков»

Рассмотрим два примера использования разработанной программы.

Пример 1. Расчет коэффициентов $\mathbf{a} = (a_0, a_1, .a_2, ., a_m)$ регрессионной модели на основании выбранного пользователем метода (в данном примере — метод максимального правдоподобия), выбранных пользователем признаков разрушения костей, которые используются для определения вида травмирующего воздействия, а также оценка качества построенной модели.

Коэффициенты регрессионной модели $\mathbf{a} = (a_0, a_1, .a_2, ., a_m)$ приведены в табл. 1, а показатели, характеризующие качество модели — ниже.

Показатели качества модели:

- правильно классифицировано: 129;
- неправильно классифицировано: 12;
- среднеквадратическая ошибка: 0.268;
- множественный коэффициент корреляции: 0.701.

Пример 2. Расчет коэффициентов $\mathbf{a} = (a_0, a_1, .a_2, ., a_m)$ регрессионной модели на основании выбранного пользователем метода (в данном примере — метод максимального правдоподобия), путем пошагового включения признаков разрушения кости, которые используются для определения вида травмирующего воздействия, а также оценка качества построенной модели.

Таблица 1

Коэффициенты уравнения регрессии, рассчитанные по заданному пользователем набору признаков методом максимального правдоподобия

Признак	Коэффициент
Константа	–1.225
КоэфМР	–4.603
КоэфЗР	39.279
КолТр1п	0.108
КолОс3с	–0.495
СрПлОс3с	0.007
КолОс3Р	–0.592

Коэффициенты регрессионной модели $\mathbf{a} = (a_0, a_1, .a_2, ., a_m)$ приведены в табл. 2, а показатели, характеризующие качество модели — ниже. Данные по включенным в модель признакам и значениям коэффициента множественной корреляции, полученным после вклю-

чения соответствующих признаков, приведены для первых пяти шагов последовательного включения признаков.

Таблица 2

Коэффициенты уравнения регрессии, рассчитанные при пошаговом включении признаков

№ шага	Включенный признак	Коэффициент множ. корреляции
1	КоэфЗР	0.587
2	СрДлТ1п	0.643
3	СрДлТрОс	0.664
4	СрПлОсЗс	0.683
5	КолОсЗс	0.704

Показатели качества модели:

- правильно классифицировано: 131;
- неправильно классифицировано: 10;
- среднеквадратическая ошибка: 0.244;
- множественный коэффициент корреляции: 0.704.

6. ЗАКЛЮЧЕНИЕ

Разработанная программа, предназначенная для исследования возможности определения вида травмирующего воздействия с помощью регрессионного метода, позволяет рассчитывать коэффициенты регрессионной модели и оценивать ее качество. На этой основе становится возможным выбор наилучшего варианта модели для решения задачи определения вида травмирующего воздействия по признакам разрушения длинных трубчатых костей в судебно-медицинской практике, а также определение наиболее значимых признаков.

ЛИТЕРАТУРА

1. Диагностикум механизмов и морфологии переломов при тупой травме скелета. Т. 1. Механизмы и морфология переломов длинных трубчатых костей / В. И. Бахметьев, В. Н. Крюков, В. П. Новоселов и др. — 2-е изд. — Новосибирск: Наука. Сибирская издательская фирма РАН, 2002. — 166 с.

2. *Бахметьев В.И.* Алгоритм распознавания способа травмирующего воздействия по параметрам разрушений длинных трубчатых костей / В. И. Бахметьев, В. А. Голуб, Г. А. Князев, Н. В. Огаркова // Системные проблемы надежности, качества, информационных и электронных технологий в инновационных проектах (Инноватика — 2006) / Материалы Международной конференции и Российской научной школы. Часть 5, Т. 1. — М.: Радио и связь, 2006. — С. 101-104.

3. Логистическая регрессия и ROC-анализ — математический аппарат (http://www.basegroup.ru/regression/logistic.htm).

4. *Дрейпер Н., Смит Г.* Прикладной регрессионный анализ. Часть 2. Эконометрия — I: Регрессионный анализ.

5. *Цыплаков А.А.* Метод максимального правдоподобия в эконометрии. Новосибирск: ЭФ НГУ, 1997, 129 с.

6. *Зандер Е.В.* Эконометрика: Учебно-методический комплекс. Красноярск: РИО КрасГУ, 2003, 36 с.

Статья принята к опубликованию 25 декабря 2006 г.