УДК 61:311.1

Миронова П. Н., Владимирова Л. В.

Построение логистической регрессии в медицине

Рекомендовано к публикации доцентом Владимировой Л. В.

В данной работе рассматривается задача о построении логистической регрессии с бинарной зависимой переменной. Логит-преобразование позволяет получать значение величины, принимающей значение в $(0, 1)$ и интерпретируется как функция распределения некоторой случайной величины [1].

1. Логит модель. Пусть координаты вектора $x = (x_1, x_2, \ldots, x_n)$ являются предикторами (характеристиками) пациентов и имеется N наблюдений $(x^{(i)}, y_i)$, $x^{(i)} = (x_{i1}, x_{i2}, \ldots, x_{in})$, y_i — зависимая переменная, которая может принимать значения ноль и единица, $i = \overline{1, N}$. Введем функцию логистического распределения, которую называют логит моделью $F(z) = \frac{e^z}{1+e^z}$, где $z = \beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n$, $\beta = (\beta_0, \beta_1, \ldots, \beta_n)^T$ — вектор неизвестных параметров. Эта функция является функцией от вектора предикторов и от вектора неизвестных параметров: $F = F(x, \beta)$.

2. Функция максимума правдоподобия. По данным наблюдений составим функцию максимального правдоподобия для логистической регрессии

$$L(\beta) = \prod_{i=1}^{N} \left(F(x^{(i)}, \beta)\right)^{y_i} \left(1 - F(x^{(i)}, \beta)\right)^{(1-y_i)} = \\ = \prod_{i=1}^{N} \left(\frac{F(x^{(i)}, \beta)}{1 - F(x^{(i)}, \beta)}\right)^{y_i} \left(1 - F(x^{(i)}, \beta)\right).$$

Заметим, что $1 - F(z) = \frac{1}{1+e^z}$. Тогда функция правдоподобия будет иметь вид

$$L(\beta) = \prod_{i=1}^{N} z_i^{y_i} (1 + e^{z_i})^{-1}, \tag{1}$$

где $z_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_n x_{in}$. Рассмотрим логарифм от функции (1)

Миронова Полина Николаевна – магистр, Санкт-Петербургский государственный университет; e-mail: pollimir@bk.ru, тел. +79052552988

Владимирова Людмила Васильевна – доцент, Санкт-Петербургский государственный университет; e-mail: sergvlad@sp.ru, тел.: +7(921)3486184

$$\ln L(\beta) = \sum_{i=1}^{N} y_i z_i - \sum_{i=1}^{N} (1 + e^{z_i}). \qquad (2)$$

Так как аргументы максимумов функций (1) и (2) равны, то будем искать максимум функции (2) по неизвестному вектору параметров β. Сначала применим простейший случайный поиск [2] для нахождения глобального максимума функции (2). Полученный вектор $\widetilde{\beta} = \arg\max\limits_{\beta \in \Pi} \ln L(\beta)$ используется как начальный для градиентного метода с целью уточнения полученного результата.

3. Оценка значимости коэффициентов модели. Проверка гипотез H_0: β_i=0, i=0, 1, ..., n, имеет большое значение. От результата проверки гипотезы при каждом i зависит оставлять или выбросить переменную фактор x_i из уравнения регрессии. Если гипотеза подтвердилась с большой вероятностью, то x_i удаляют из регрессии, если нет, то оставляют. Для логит моделей проверка гипотез о значимости коэффициентов может проводиться с помощью z статистики [1].

Рассмотрим нулевую гипотезу: H_0 : β_i=0. Сформулируем альтернативную гипотезу: H_1 : $\beta_i \neq 0$. Зададим уровень значимости α.

Вычислим значение статистики z: $z = \frac{\widehat{\beta}_i}{\widehat{\sigma}(\beta_i)}$, где $\widehat{\beta}_i$ — полученная оценка, $\widehat{\sigma}(\beta_i)$ — оценка стандартной ошибки регрессии, которая определяется как диагональный элемент обратной матрицы вторых производных логарифма правдоподобия $\ln L(x, \beta)$.

4. ROC-анализ. ROC-анализ представляет собой графическую методику оценивания эффективности модели с помощью двух показателей — специфичности и чувствительности. В бинарной классификации каждое предсказание может иметь четыре исхода:
TP — количество истинно положительных исходов,
TN — количество истинно отрицательных исходов,
FP — количество ложноположительных исходов (ошибки II рода),
FN — количество ложноотрицательных исходов (ошибки I рода).
Чувствительность (sensitivity) является отношением числа истинно положительных наблюдений к числу фактически положительных наблюдений: $Sensitivity = \frac{TP}{FN+TP}$. Специфичность (specificity) определяется как отношение числа истинно отрицательных наблюдений к числу фактически отрицательных наблюдений:
$Specificity = \frac{TN}{FP+TN}$.

5. Численные результаты. Имеется набор данных о пациентах с подозрением на сахарный диабет. Объем выборки: $N = 392$. Из них количество случаев с наличием заболевания: $TP = 130$, количество случаев с отсутствием заболевания: $TN = 262$. Ресурс данных: Национальный институт диабета (США), болезней органов пищеварения и почек. Дата публикации данных: 09-05-1990 [3].

Рассмотрим двумерную задачу $n = 2$ с двумя характеристиками: x_1 = возраст пациента и x_2 = уровень глюкозы в крови. Требуется определить с помощью логистической регрессии вероятность заболевания пациента сахарным диабетом. В таблице 1 представлены некоторые значения данной выборки:

Таблица 1. Значения уровня глюкозы в крови, возраста и наличия заболевания

Возраст x_{1i} (лет)	**Уровень глюкозы в крови (ммоль/л)**	**Наличие заболевания** y_i
21	4.9	0
33	5.7	0
51	9.2	1
53	10.9	1
59	10.5	1

Для рассматриваемой задачи функция максимума правдоподобия имеет вид

$$\ln L(X, \beta) = \sum_{i=1}^{N} y_i(\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i}) - \sum_{i=1}^{N}(1 + e^{(\beta_0+\beta_1 x_{1i}+\beta_2 x_{2i})}). \quad (3)$$

Используя простейший случайный поиск и градиентный метод для нахождения экстремума (3), получаем коэффициенты β_0=7.164, β_1= − 0.051, β_2= − 0.688.

В итоге, функция логистической регрессии для данной задачи имеет вид (рис. 1):

$$F(y = 1|x) = \frac{1}{1 + e^{7.164-0.051x_1-0.688x_2}}.$$

Ниже, в таблице 2, в колонке выходных данных модели представлены значения функции логистической регрессии. Как видно, «наличия заболеваний» и «вероятности наличия заболеваний» в обучающей группе пациентов близки.

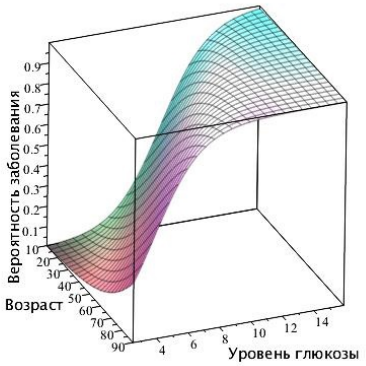

Рис. 1. График функции логистической регрессии

Таблица 2. Сравнение результатов модели с исходными данными

Данные о пациентах					Выходные данные модели
Тип группы пациентов	№	Возраст	Уровень глюкозы в крови	Наличие заболевания	Вероятность наличия заболевания
Обучающая группа пациентов	1	21	4.9	0	0.063
	2	33	5.7	0	0.174
	3	51	9.2	1	0.853
	4	53	10.9	1	0.955
	5	59	10.5	1	0.954
Группа пациентов, для которых необходимо сделать прогноз	12	12	12.6		0.892
	13	12	5.6		0.063
	14	24	5.8		0.123
	15	13	5.8		0.075
	16	36	6		0.229

Пользуясь построенной моделью, можно получить прогноз для новой группы пациентов, о наличии/отсутствии заболевания, которых нам неизвестно.

Согласно данным таблицы 2, доля ошибочно классифицированных случаев составила 22% (с порогом отсечения 0.5). Согласно классификатору моделей, данная модель может быть идентифицирована как «хорошая» [4].

Для оценки значимости коэффициентов модели зададим уровень значимости $\alpha = 0.05$ и рассчитаем оценки стандартных ошибок полученных коэффициентов, значение статистик z. Статистики z для каждого коэффициента регрессии попали в критическую область $(-\infty, -1.96) \cup (1.96, \infty)$, следовательно, все коэффициенты регрессии являются значимыми.

Таблица 3. Оценки значимости коэффициентов модели

Наименование коэффициента	Оценка коэффициента	Стандартная ошибка	Статистика Z
β_0	-7.17	0.72	-9.9
β_1	0.05	0.01	3.9
β_2	0.69	0.09	7.9

Для оценки модели построим матрицу классов (таблица 4).

Таблица 4. Матрица классов, соответствующая порогу 0.5

	$\widehat{y_i} = 1$	$\widehat{y_i} = 0$	Количество
$y_i = 1$	74(TP)	56 (FN)	130
$y_i = 0$	29(FP)	233(TN)	262
Количество	103	289	392

Общий показатель ошибки (Overall error rate) определяется отношением числа неправильно классифицированных наблюдений к общему числу наблюдений:
$Overall\,error\,rate = \frac{FP+FN}{TP+TN+FP+FN} = 0.22$. Следовательно, точность классификатора составит 78%. Найдем чувствительность и специфичность:
$Sensitivity = \frac{TP}{FN+TP} = 0.57$, $Specificity = \frac{TN}{FP+TN} = 0.89$.

На рис. 2 построены графики специфичности (возрастающая кривая) и чувствительности (убывающая кривая) модели, которые показывают насколько хорошо модель предсказывает отрицательные события. Горизонтальная ось (Cutoff) соответствует порогу отсечения. Точка пересечения двух графиков является оптимальным порогом отсечения. Оптимальный параметр (порог отсечения) соответствует 0.3. Построим матрицу классов, соответствующую данному порогу отсечения 0.3.

Таблица 5. Матрица классов, соответствующая порогу 0.3

	$\widehat{y_i} = 1$	$\widehat{y_i} = 0$	Количество
$y_i = 1$	96(TP)	34 (FN)	130
$y_i = 0$	64(FP)	198(TN)	262
Количество	160	232	392

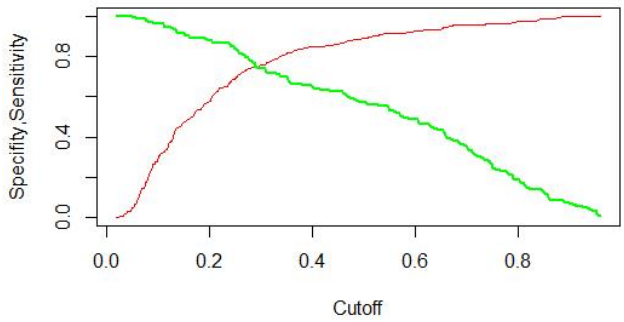

Рис. 2. Графики специфичности и чувствительности модели

Рассчитаем основные показатели модели:

$$Overall\ success\ rate = \frac{TP + TN}{TP + TN + FP + FN} = 0.75,$$

$$Sensitivity = \frac{TP}{FN + TP} = 0.74, \quad Specificity = \frac{TN}{FP + TN} = 0.76.$$

6. Выводы.

1. Анализ данных подтвердил зависимость исхода (наличия/отсутствия заболевания) и двух характеристик пациентов (уровень глюкозы в крови и возраст).
2. Логистическая модель с порогом отсечения 0.5 достаточно точно классифицирует пациентов, однако, в связи со спецификой задачи и полученным результатом ROC-анализа разумнее выбрать порог отсечения равный 0.3. Общая точность модели в таком случае понизится, зато повысится чувствительность модели, что позволит минимизировать ошибки классификации больных пациентов в группу здоровых (ошибки I рода).

Литература

1. Буре В. М., Парилина Е. М., Седаков А. А. Методы прикладной статистики в R и Excel. Учебное пособие. СПб.: Лань, 2016. 152 с. ISBN 978-5-8114-2229-6.

2. Владимирова Л. В. Использование случайного поиска с «памятью» в оценке параметров нелинейной регрессии // Вестник СПГУТД. Серия 1. 2013. № 4. С. 30–34. ISSN 2079-8199.

3. Smith J. W., Everhart J. E., Dickson W. C., Knowler W. C. and Johannes R. Using the ADAP learning algorithm to forecast the onset of diabetes mellitus. 1988.

4. Паклин Н. Б., Орешков В. И. Бизнес-аналитика: от данных к знаниям. Учебное пособие. 2-е изд., испр. СПб.: Питер, 2013. 704 с.