УДК 51.77

М. А. Сполохова, аспирант
С. Б. Пономарев, доктор медицинских наук, профессор
Е. Л. Аверьянова, кандидат медицинских наук
ИжГТУ имени М. Т. Калашникова

МАТЕМАТИЧЕСКАЯ МОДЕЛЬ ОЦЕНКИ РАЗВИТИЯ СПИДА У ВИЧ-ИНФИЦИРОВАННЫХ ПАЦИЕНТОВ

Разработан метод оценки динамики развития СПИДа у ВИЧ-инфицированных пациентов с помощью математического моделирования. Построена модель и определена вероятность развития заболевания по стадиям.

Ключевые слова: математическое моделирование, регрессионная линейная модель, пошаговый отбор регрессоров, метод наименьших квадратов, множественная регрессия.

Инфекция, вызываемая вирусом иммунодефицита человека (ВИЧ), признана одним из опаснейших инфекционных заболеваний человека, в финале которой развивается синдром приобретенного иммунного дефицита – СПИД. Главная опасность ВИЧ-инфекции – практически неизбежная гибель инфицируемого после заражения ВИЧ. При этом определить время наступления финальной стадии СПИДа достаточно сложно из-за большого числа факторов, влияющих на течение инфекции [1].

Цель исследования – это разработка метода оценки динамики развития СПИДа у ВИЧ-инфицированных пациентов с помощью математического моделирования.

В качестве входных переменных рассмотрено 49 регрессоров (x_i), лабораторно полученных характеристик здоровья ВИЧ-инфицированных. В качестве результирующего фактора взята переменная (y), отражающая динамику заболевания в течение года после проведенного исследования. Статистическая база данных, на основе которых было проведено моделирование зависимости $y = f(x_i)$, включала 45 наблюдений. Группа проверки эффективности модели включала 20 наблюдений.

Построение математической модели, в которой отобраны факторы x_i, оказывающие весомое влияние на значение y, осуществлялось с помощью процедуры пошагового включения фактора в регрессионную линейную модель $y = f(x_i)$ [2].

1. На первом шаге из исходного набора объясняющих переменных были выбраны переменные, имеющие наибольший по модулю коэффициент корреляции с зависимой переменной y.

2. Второй шаг состоял из двух подшагов:

2.1. На первом из них ищется тот x_s, удаление которого приводит к наименьшему уменьшению коэффициента детерминации. Затем сравнивается значение F-статистики для проверки гипотезы H_0 о незначимости этого регрессора с некоторым заранее заданным пороговым значением $F_{\text{искл}}$. Если $F < F_{\text{искл}}$, то x_s удаляется из списка регрессоров. Заметим, что гипотеза H_0 о равенстве коэффициента при x_s нулю эквивалентна гипотезе о равенстве коэффициентов детерминации до и после удаления регрессора, а также гипотезе о том, что коэффициент частной корреляции x_s и y равен 0.

2.2. Второй подшаг состоит в попытке включения нового регрессора из исходного набора предсказывающих переменных. Ищем переменную x_p с наибольшим по модулю частным коэффициентом корреляции (исключается влияние ранее включенных в уравнение регрессоров) и сравниваем значение F-статистики для проверки гипотезы H_0 о незначимости этого регрессора с некоторым заранее заданным пороговым значением $F_{\text{вкл}}$. Если $F > F_{\text{вкл}}$, то x_p включается в список регрессоров.

Обычно выбирают $F_{\text{искл}} < F_{\text{вкл}}$.

Второй шаг повторяется до тех пор, пока происходит изменение списка регрессоров.

В программном продукте SPSS реализован алгоритм пошагового отбора регрессоров.

Задаем параметры $F_{\text{вкл}} = 0{,}05$ и $F_{\text{искл}} = 0{,}10$. Таким образом, из 49 регрессов отобраны 7, которые вносят наибольший вклад в объяснение вариации зависимой переменной (табл. 1). Это диагноз (D), размеры печени по Курлову (R), кашель (K), ОАК лимфоциты (L), миалгия (M), туберкулез (T), затылочные лимфоузлы (Z) (табл. 1).

Таблица 1. **Отобранные регрессоры с использованием метода пошагового отбора факторов в программном продукте SPSS**

Модель	Включенные переменные	Метод
1	D	Включение (критерий: вероятность F-включения >= ,050)
2	R	Включение (критерий: вероятность F-включения >= ,050)
3	K	Включение (критерий: вероятность F-включения >= ,050)
4	L	Включение (критерий: вероятность F-включения >= ,050)
5	M	Включение (критерий: вероятность F-включения >= ,050)
6	T	Включение (критерий: вероятность F-включения >= ,050)
7	Z	Включение (критерий: вероятность F-включения >= ,050)

Обозначим выбранные объясняющие регрессоры следующим образом (табл. 2).

Таблица 2. **Отобранные регрессоры и их трактовка**

Включенные переменные	Обозначение	Описание фактора	Принимаемые дискретные значения
x2	t_1	D	1–4
x22	t_2	R	Различные
x11	t_3	K	0,1
x43	t_4	L	Различные
x13	t_5	M	0,1
x3	t_6	T	0–8
x24	t_7	Z	0,1

Методом наименьших квадратов определяются коэффициенты в линейной регрессии вида [3]:

$$y = \alpha_1 \cdot t_1 + \alpha_2 \cdot t_2 + \alpha_3 \cdot t_3 + \alpha_4 \cdot t_4 + \\ + \alpha_5 \cdot t_5 + \alpha_6 \cdot t_6 + \alpha_7 \cdot t_7 + e, \quad (1)$$

где e – остатки.

Проанализируем результаты регрессионной статистики и дисперсионного анализа (табл. 3–5).

Таблица 3. **Регрессионная статистика в программе SPSS**

Вид фактора	Числовое значение
Множественный *R*	0,991552
R-квадрат	0,983176
Нормированный *R*-квадрат	0,954204
Стандартная ошибка	0,525192
Наблюдения	45

Таблица 4. **Дисперсионный анализ в программе SPSS**

Параметр	Df	SS	MS	*F*	Значимость *F*
Регрессия	7	612,5186	87,50265	317,237	4,97
Остаток	38	10,48143	0,275827	4	E-31
ИТОГО	45	623			

Таблица 5. **Дисперсионный анализ в программе SPSS (Y-пересечения и t-факторы)**

Параметр	Коэффициенты	Стандартная ошибка	*t*-статистика	*P*-значение	Нижние 95 %	Верхние 95 %	Нижние 95 %	Верхние 95 %
Y-пересечение	0	#Н/Д	#Н/Д	#Н/Д	#Н/Д	#Н/Д	#Н/Д	#Н/Д
t_1	–0,0419	0,14738	–0,28	0,347	–0,3403	0,2563	–0,34	0,256
t_2	0,11712	0,02151	5,444	3,29E-0,6	0,07357	0,1606	0,073	0,160
t_3	0,35218	0,20452	1,721	0,093	–0,0618	0,7662	–0,06	0,766
t_4	0,02213	0,00555	3,986	0,000294	0,01089	0,0333	0,010	0,033
t_5	0,69682	0,18363	3,794	0,000517	0,32506	1,0685	0,325	1,068
t_6	0,19642	0,04706	4,173	0,000168	0,10115	0,2916	0,101	0,291
t_7	0,76425	0,34928	2,188	0,034	0,05716	1,4713	0,057	1,471

Таким образом, получили множественную регрессионную модель.

Для оценки адекватности построенной модели используется коэффициент детерминации R^2 (отношение объясненной дисперсии к общей):

$$R^2 = \frac{RSS}{TSS}, \quad (2)$$

где

$$RSS = \sum_{t=1}^{T} (y_t^{\text{mod}} - \overline{y})^2, \; TSS = \sum_{t=1}^{T} (y_t - \overline{y})^2. \quad (3)$$

Коэффициент детерминации R^2 характеризует качество подгонки регрессионной модели к значениям временного ряда y_t. Если $R^2 = 0$, то регрессия не улучшает качество предсказаний y_t^{mod} по сравнению с тривиальным предсказанием $\overline{y}$. Если же $R^2 = 1$, то говорят о точной подгонке модели, т. е. все точки наблюдений удовлетворяют уравнению регрессии.

Коэффициент детерминации R^2 = 0,983, это говорит о том, что 7 регрессоров оценивает 98,3 % вариации результата.

Для определения степени значимости коэффициентов регрессионной модели, следовательно, и значимости всей модели используют статистику Фишера F [4]:

$$F = \frac{R^2}{1 - R^2} \cdot \frac{n - k}{k - 1}, \quad (4)$$

где n – количество наблюдений; k – количество оцениваемых параметров.

Выдвигают основную и альтернативную гипотезы:

$$H_0 : \beta_i = 0; \quad (5)$$

$$H_1 : \beta_i \neq 0. \quad (6)$$

При выполнении основной гипотезы статистика F имеет распределение Фишера с $(k-1, n-k)$ – степенями свободы. Величину F сравнивают с табличным значением $F_\alpha(k-1, n-k)$ на уровне значимости α. Если $F < F_{\text{табл}}(k-1, n-k)$, то принимают основную гипотезу о равенстве коэффициентов регрессии нулю. Если $F > F_{\text{табл}}(k-1, n-k)$, то основную гипотезу отвергают в пользу альтернативной о значимости коэффициентов регрессии, следовательно, модель адекватная.

Оценку надежности уравнения линейной регрессии в целом дает *F*-критерий Фишера: для оцененной модели он составил 317,24. Вероятность случайно получить такое значение *F*-критерия не превышает допустимый уровень значимости 1 %. Следовательно, полученное значение критерия Фишера не случайно, оно сформировалось под влиянием существенных факторов, т. е. подтверждается статистиче-

ская значимость всей модели и показателя тесноты связи R^2.

На рис. 1 изображен график остатков в регрессионной модели, он представляет собой стационарный ряд.

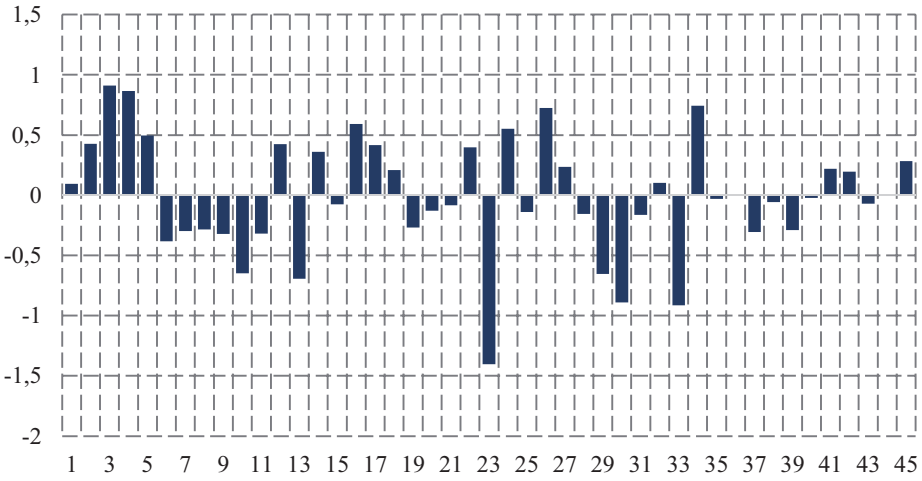

Рис. 1. График остатков в регрессионной модели

Оценим значимость коэффициентов в регрессионной модели с помощью критерия Стьюдента. На уровне значимости 10 % значим коэффициент при факторе t_3, на уровне в 5 % – при факторе t_7, на уровне в 1 % – при факторе t_2, t_4, t_5, t_6.

Коэффициенты в уравнении регрессии представляют собой частную корреляцию с результирующим фактором. Так, факторы t_3, t_5, t_7 в статистической базе могут принимать значения 0 и 1, следовательно, при интерпретации результатов мы говорим, что [5]:

- наличие кашля у пациента t_3 усугубляет показатель развития СПИДа и вносит вклад в рост результирующего показателя y на 0,352 доли;
- наличие миалгии (боли в мышцах) у пациента t_5 усугубляет показатель развития СПИДа и вносит вклад в рост результирующего показателя y на 69,7 %;
- наличие увеличенных затылочных лимфоузлов t_7 усугубляет показатель развития СПИДа и вносит вклад в рост y на 0,764 доли.

Можно сказать по коэффициентам модели, что наличие сразу всех этих трех факторов у больного уже может свидетельствовать как минимум о первой либо второй стадии развития СПИДа (0,352+0,697+0,764), при наличии других факторов. Эти коэффициенты регрессии являются значимыми [6].

Коэффициент при t_1, равный –0,042, означает, что при переходе диагноза на единицу в сторону роста уменьшается величина, характеризующая стадию развития СПИДа на 0,042 %.

Коэффициент при t_2, равный 0,117, означает, что при увеличении размеров печени у пациента увеличивается величина, характеризующая стадию развития СПИДа на 0,117 доли от значения t_2.

Значение t_4 характеризует показатель лимфоцитов в общем анализе крови больного. Коэффициент в множественной регрессии при t_4 равен 0,022, это свидетельствует о том, что при росте на 1 % показателей лимфоцитов в крови увеличивает величину, характеризующую стадию развития СПИДа на 0,022 %.

Стадии развития туберкулеза также усугубляют риск негативного исхода заболевания. При росте показателя, характеризующего стадии туберкулеза на единицу, увеличивается вероятность перехода на новую стадию развития СПИДа с долей 0,196 в результирующий фактор y.

Рассчитаем среднюю относительную ошибку аппроксимации по формуле:

$$\overline{\delta} = \frac{1}{N}\sum_{t=1}^{N}\frac{\left|y_t - y_t^{\text{mod}}\right|}{y_t}100\,\%\,. \qquad (7)$$

Здесь N – количество наблюдений; y_t^{mod} – значение уровня ряда в момент времени t, рассчитанное по модели.

Средняя относительная ошибка аппроксимации на данных базы *обучения* составила 11,2 %.

График ошибки аппроксимации для базы обучения приведен на рис. 2.

Средняя относительная ошибка аппроксимации на данных базы *проверки* составила 19,5 %.

На рис. 3 представлен график, где черными точками обозначены статистические данные по стадиям развития СПИДа у пациентов, а белыми – определение стадии СПИДа по множественной регрессионной модели.

Таким образом, благодаря построенной математической модели, представляющей собой множественную регрессию, которая зависит от семи факторов, мы можем определить стадию развития СПИДа с вероятностью 86 %.

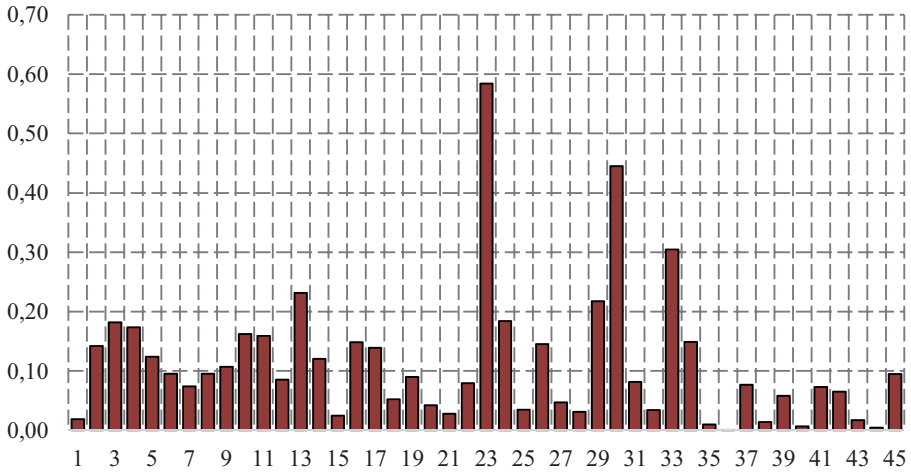

Рис. 2. Ошибка аппроксимации для множественной регрессии

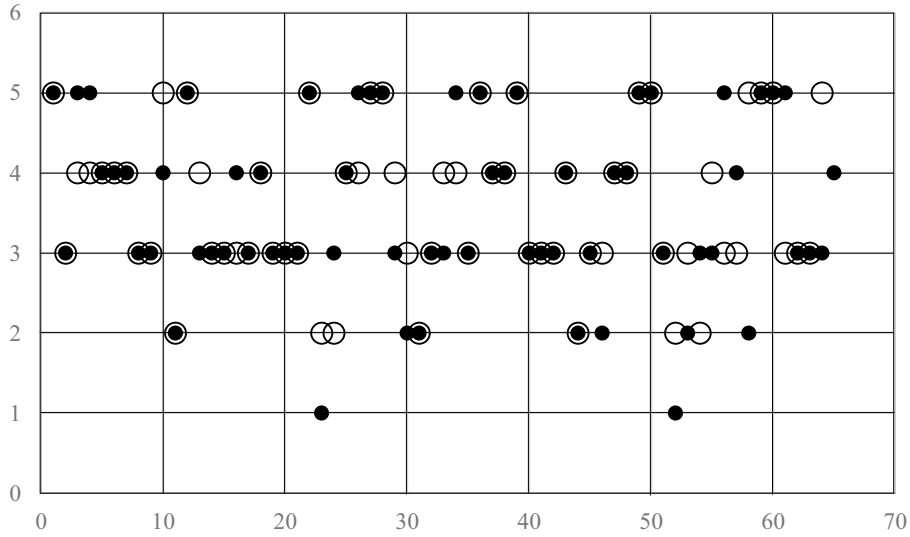

Рис. 3. Статистические данные по стадиям развития СПИДа по множественной регрессионной модели

Библиографические ссылки

1. Прикладная статистика: Основы моделирования и первичная обработка данных. Справочное изд. / С. А. Айвазян и др. – М. : Финансы и статистика, 1983. – 471 с.

2. *Бородич С. А*. Вводный курс эконометрики. – Минск : Новое знание, 2004. – 368 с.

3. *Магнус Я. Р., Катышев П. К., Пересецкий А. А*. Эконометрика. Начальный курс – М. : Дело, 2004. – 576 с.

4. *Носко В. П.* Эконометрика. Введение в регрессионный анализ временных рядов. – М., 2002. – 274 с.

5. *Пономарев С. Б., Горохов М. М., Серебренников А. В., Логинова С. Г.* К вопросу о применении информационных систем для оптимизации тактики ведения больных в местах лишения свободы // Интеллектуальные системы в производстве. – 2007. – № 2. – С. 100–103.

6. *Кудашева Л. Т., Пономарев С. Б., Тененев В. А., Сергиенко А. С.* Использование информационных технологий при ведении санитарно-эпидемиологического мониторинга учреждений уголовно-исполнительной системы // Интеллектуальные системы в производстве. – 2007. – № 2. – С. 132–142.

* * *

M. A. Spolokhova, Post-graduate, Kalashnikov ISTU
S. B. Ponomarev, Doctor of Medicine, Professor, Kalashnikov ISTU
E. L. Averyanova, PhD in Medicine, Kalashnikov ISTU

Mathematical model for assessing the development of AIDS in HIV infected patients

The method for assessing the dynamics of AIDS in HIV infected patients with the help of mathematical modeling has been developed. A mathematical model has been constructed and the probability of developing the disease in stages has been determined.

Keywords: mathematical modeling, linear regression model, stepwise selection of regressors, least squares method, multiple regression.

Получено: 06.05.16