Таблица 3

Результаты оценивания параметров и ошибок модели

		$\theta(1)=\begin{pmatrix}0.98 & 1.96\\ 0 & 0.98\end{pmatrix}$	$\Sigma=\begin{pmatrix}1 & 0\\ 0 & 1\end{pmatrix}$
LDL -разложение	Оценка	$\begin{pmatrix}0.9795 & 1.8633\\ 0.1182 & 0.8826\end{pmatrix}$	$\begin{pmatrix}1.0395 & -0.1849\\ 0.0224 & 1.0278\end{pmatrix}$
	MSE	$\begin{pmatrix}0.0045 & 0.0090\\ 0.0124 & 0.0103\end{pmatrix}$	$\begin{pmatrix}0.0324 & 0.0224\\ 0.0224 & 0.0298\end{pmatrix}$

Согласно данным табл. 3, погрешность оценивания параметров составляет 0,5–3%.

Результаты компьютерного моделирования показывают, что подход, основанный на использовании модифицированной функции максимального правдоподобия и разложении Холецкого, достаточно эффективен и дает погрешности оценивания параметров процесса скользящего среднего порядка 1–3%.

Литература

1. Харин, Ю. С. Математическая и прикладная статистика / Ю.С. Харин, Е.Е. Жук. – Минск: БГУ, 2005. – 279 с.
2. Бокс, Дж. Анализ временных рядов. Прогноз и управление / Дж. Бокс, Г. Дженкинс. – М.: Мир, 1974. – Вып.1.
3. Андерсон, Т. Статистический анализ временных рядов / Т. Андерсон. – М.: Мир, 1976. – 757 с.
4. Hamilton, J. Time Series Analysis / J. Hamilton.-Princeton University Press, 1994. – 800 p.

Статья поступила 01.02.2018 г.

ПРИМЕНЕНИЕ ЛОГИСТИЧЕСКОЙ РЕГРЕССИИ ПРИ ОБРАБОТКЕ ЭКОНОМИЧЕСКИХ ДАННЫХ

Белько И.В.,
доктор физико-математических наук, профессор,
Криштапович Е.А.,
Белорусский государственный аграрный технический университет

Применение логистической регрессии при обработке экономических данных продемонстрировано на примере статистического анализа результатов тестирования пациентов с болезнью органов дыхания. В Республиканском научно-практическом центре пульмонологии и фтизиатрии разработана анкета, результаты которой используются в лечебной практике. Проведен опрос пациентов контрольной группы из 65 чел., обладающих лекарственной чувствительностью (ЛЧ), и 115 пациентов основной группы с лекарственной устойчивостью (ЛУ) возбудителя туберкулеза.

Анкета состоит из следующих клинических, клинико-рентгенологических и социальных показателей, многие из которых являются качественными, в частности бинарными: X_1 – возраст, X_2 – пол, X_3 – злоупотребление алкоголем, X_4 – курение, X_5 – сторона поражения легких, X_6 – распространенность процесса, X_7 – объем поражения, X_8 – скиалогическая картина, X_9 – количество полостей распада, X_{10} – нарушение бронхиальной проходимости, X_{11} – наличие синдрома диссеминации, X_{12} – вовлечение в процесс плевры, X_{13} – наличие пневмоторакса, X_{14} – аномалии развития легких, X_{15} – аномалии развития сосудов средостения. Основной зависимой переменной служит переменная Y, которая принимает значение $Y = 1$ при наличии ЛЧ или

значение $Y = 0$ при отсутствии ЛЧ. Исследование связей между показателями проводится эконометрическими методами регрессионного анализа [1]. При этом качественные показатели с уровнями больше двух вызывают наличие их мультиколлинеарности и не позволяют эффективно применять классический регрессионный анализ. Для устранения мультиколлинеарности такого типа, согласно теории, вместо каждой качественной переменной вводятся бинарные переменные в числе на единицу меньше числа уровней самой переменной. Например, переменная X_5, задающая сторону поражения легких, имеет три уровня: 1 – правое, 2 – левое, 3 – оба легких, переменная X_9, задающая количество полостей распада, имеет 4 уровня: 1 – одна полость, 2 – две полости. 3 – больше двух полостей, 4 – полости распада отсутствуют. Вместо переменной X_5 вводятся две бинарные переменные: $X_{5,1}$ принимает значение 1 для первого уровня и значение 0 для остальных уровней; переменная $X_{5,2}$ принимает значение 1 для второго уровня и значение 0 для остальных уровней. Таким же образом вместо переменной X_9 вводятся три бинарные переменные: $X_{9,1}$ принимает значение 1 для первого уровня и значение 0 – для остальных уровней; переменная $X_{9,2}$ принимает значение 1 для второго уровня и значение 0 – для остальных уровней; переменная $X_{9,3}$ принимает значение 1 для третьего уровня и значение 0 – для остальных уровней. В применяемом нами пакете прикладных программ SPSS [2] такое разбиение переменных производится автоматически. Этот пакет позволяет также проводить линейную множественную и логистическую регрессию, которая необходима при наличии бинарной зависимой переменной. Условие нормальности распределения случайных остатков, необходимое для качественных свойств оценок при использовании метода наименьших квадратов, обеспечивается относительно большим объемом выборки наблюдений (в нашем случае – 180). В результате исследования из всех показателей анкеты нами отобрано 7, значения которых позволяют с высокой вероятностью судить о наличии или отсутствии ЛЧ. На первом этапе проводим логистическую регрессию по всем факторам и всем наблюдениям. Однако ее итог имеет низкое качество, он дает лишь 64% верно предсказанных наблюдений. При первоначальном отборе факторов используются коэффициенты корреляции и значимость коэффициентов уравнения регрессии. При последующем пошаговом отборе факторов незначимыми оказываются факторы X_4, X_{11}, X_{13}, X_{14}, X_{15}. Неожиданное исключение фактора X_4 (курение), видимо, можно объяснить недостоверностью информации в анкетах. На шестом шаге исключения для дальнейшего анализа мы оставляем 151 наблюдение по следующим семи факторам: X_1 – возраст, X_3 – злоупотребление алкоголем, X_5 – сторона поражения легких, X_7 – объем поражения, X_9 – количество полостей распада, X_{10} – нарушение бронхиальной проходимости, X_{12} – вовлечение в процесс плевры. Отметим, что большинство факторов будут качественными, при этом факторы X_5 и X_9 имеют более двух уровней. Поэтому, как отмечалось выше, они разбиваются на бинарные: X_5 – на $X_{5,1}$ и $X_{5,2}$; X_9 – на $X_{9,1}$, $X_{9,2}$ и $X_{9,3}$.

Для сравнения приведем уровни значимости коэффициентов при отобранных факторах: X_1 – 0,000; X_3 – 0,060; X_5 – 0,039; $X_{5,1}$ – 0,855; $X_{5,2}$ – 0,035; X_7 – 0,000; X_9 – 0,014; $X_{9,1}$ – 0,091; $X_{9,2}$ – 0,087; $X_{9,3}$ – 0,017; X_{10} – 0,021; X_{12} – 0,000; константа – 0,160. Как видим, все коэффициенты имеют высокую значимость, кроме коэффициента при первой бинарной переменной $X_{5,1}$ (правое легкое) для стороны поражения легких. Можно отметить важную особенность фактора X_5: несмотря на то что в целом коэффициент при X_5 значим, его первый уровень $X_{5,1}$ оказывается незначимым. В отличие от X_5, сам фактор X_9 (количество полостей распада) значим вместе со всеми его уровнями.

При построении логистической регрессии сначала проводится линейная множественная регрессия для логарифма отношения шансов. Для набора значений факторов $\{X_1, X_3, ..., X_{12}\}$ вычисляется относительная частота $\hat{P}$ появления соответствующего значения $Y = 1$. Частота $\hat{P}$ может служить оценкой вероятности P появления значения $Y=1$ для заданного набора факторов. Отношением шансов называется отношение $\frac{P}{1-P}$. Согласно логистической регрессии, проводится оценка

зависимости переменной $L = \ln\left(\frac{P}{1-P}\right)$ от заданных факторов [1]. В нашем случае оценка $\hat{L}$ логарифма отношения шансов задается следующим уравнением:

$$\hat{L} = -1,865 + 0,087X_1 + 1,232X_3 - 0,118X_{5,1} - 1,733X_{5,2} - 1,202X_7 - 1,101X_{9,1} - \\ -2,147X_{9,2} + 3,383X_{9,3} - 1,704X_{10} + 2,440X_{12}. \quad (1)$$

Отметим, что методические аспекты построения уравнения (1) впервые были доложены 20 октября 2017 г. на XVIII Международной научной конференции «Проблемы прогнозирования и государственного регулирования социально-экономического развития» [3].

В пакете SPSS оценка логарифма отношения шансов и значимость коэффициентов по статистике Вальда входят в поэтапный отчет пакета. Переход от оценки $\hat{L}$ к вероятности прогнозного значения PREY совершается также в пакете по формуле $PREY = \frac{\exp\hat{L}}{1+\exp\hat{L}}$.

Полученные значения PREY условной прогнозной вероятности положительного результата ($Y = 1$) принимают значения на отрезке [0,1]. Для перехода к бинарным значениям 1 и 0 прогнозной вероятности выбирается порог отсечения (разделяющее значение). Если предсказанное значение PREY больше порога, то данное значение заменяется единицей, если меньше или равно порогу, то – нулем. Пошаговый перебор значений порога между нулем и единицей проводится с помощью ROC-анализа. Для каждого значения порога вводятся понятие чувствительности *Se* и специфичности *Sp*. Чувствительность – это отношение верно предсказанных положительных исходов к сумме таких исходов и ложно предсказанных отрицательных исходов, специфичность – это отношение верно предсказанных отрицательных исходов к сумме таких исходов и ложно предсказанных положительных исходов. Модель с высокой чувствительностью точнее обнаруживает положительные случаи, модель с высокой специфичностью – отрицательные случаи.

ROC-кривая – это ломаная, концы отрезков которой соответствуют значениям порога и имеют координаты (1-*Sp*, *Se*). Качество модели и сравнение моделей определяется по величине площади под ROC-кривой. Выбор оптимального значения порога зависит от выбранной стратегии. В пакете SPSS в качестве оптимального выбирается порог с максимальной суммой чувствительности и специфичности. Для нашей модели с семью переменными выделятся два порога отсечения – порог 0,3024 с чувствительностью 85,42% и специфичностью 74,76% (сумма – 160,17) и порог 0,4993 с чувствительностью 64,58% и специфичностью 93,2% (сумма – 157,78).

Второй порог ($\approx 0,5$) представляется предпочтительным, поскольку ЛУ более важна для обнаружения. Выбранная модель очень хорошего качества [2. С. 77], для нее площадь под ROC-кривой (AUC) равна 0,884 (рис. 1).

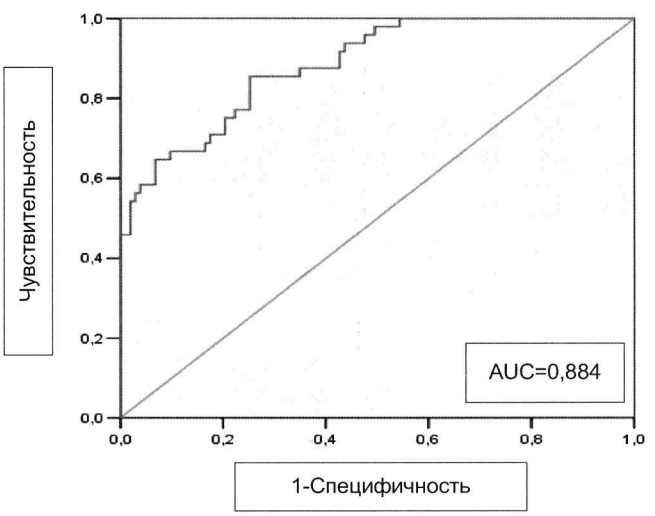

Рис. 1 ROC-кривая для семи переменных (возраст, злоупотребление алкоголем, сторона поражения, объем поражения, количество полостей распада, нарушение бронхиальной проходимости и вовлечение в процесс плевры)

Для сравнения проведена логистическая регрессия для модели с тремя переменными (возраст, объем поражения и вовлечение в процесс плевры). Уравнение логистической регрессии имеет вид:

$$\hat{L} = -2,729 + 0,067X_1 - 0,636X_7 + 1,58X_{12}. \quad (2)$$

По статистике Вальда все коэффициенты уравнения (2) значимы. Отметим, что для данной модели площадь под ROC-кривой равна 0,827 (рис. 2).

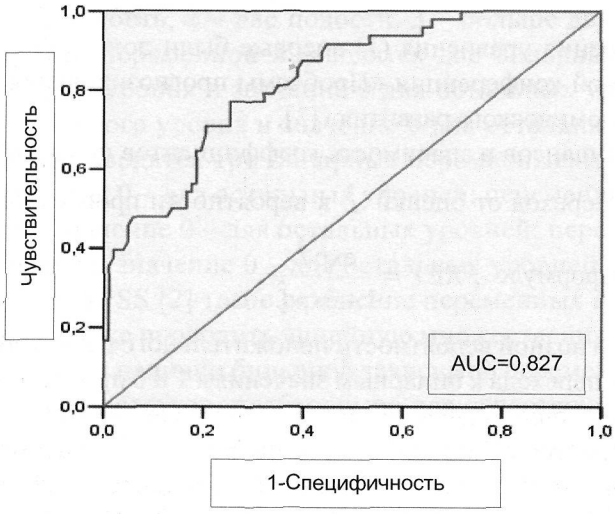

Рис. 2. ROC-кривая для трех переменных (возраст, объем поражения и вовлечение в процесс плевры)

Полученные значения оценки прогнозной вероятности лекарственной чувствительности возбудителя туберкулеза органов дыхания при заданных значениях показателей позволяют проводить диагностику ЛЧ. Эту модель можно использовать для оценки прогнозной вероятности положительного исхода [4]. Дополнительно отметим, что согласно семифакторной модели из 29 пропущенных в начале наблюдений верно классифицируются 28 (96,5%), и это свидетельствует о качестве модели. Адекватность модели подтверждается и высокой долей (84,1%) верно предсказанной лекарственной чувствительности.

Применяемая нами поэтапная методика исследования выборки наблюдений может быть использована для экономических, финансовых и других показателей.

Литература

1. D. Gujarti, Basic Econometrics, Mc Graw Hill, Inc., 1995. – 838 p.

2. Многомерный статистический анализ в экономических задачах: компьютерное моделирование в SPSS: учеб. пособие / Под ред. И.В. Орловой. – М.: Вузовский учебник, 2013. – 310 с.

3. Белько, И.В. Регрессионный анализ и оценка факторов, влияющих на заболевание органов дыхания / И.В. Белько, Е.А. Криштапович, О.М. Калечиц // Проблемы прогнозирования и государственного регулирования социально-экономического развития: материалы XVIII Международной научной конференции (Минск, 19–20 октября 2017 г.): в 3 т. / редкол. В.В. Пинигин [и др.]. – Минск: НИЭИ М-ва экономики Респ. Беларусь, 2017. – Т. 3. – С. 154–155.

4. Эконометрика и экономико-математические методы и модели: учеб. пособие / Г.О. Читая [и др.]; Под ред. Г.О. Читая, С.Ф. Миксюк. – Минск: БГЭУ, 2018. – 511 с.

Статья поступила 27.01.2018 г.