# КОМПЬЮТЕРНЫЕ НАУКИ

УДК 681.5.01

# ОЦЕНКА ЭФФЕКТИВНОСТИ АЛГОРИТМОВ РОБАСТНОГО ОЦЕНИВАНИЯ

*АНТОНОВ В.А., ШАМША Б.В.*

Методом имитационного моделирования строятся модели качественных показателей алгоритмов робастного оценивания. В виде зависимых переменных моделей качественных показателей используются среднеквадратическая ошибка и медиана абсолютных отклонений параметров регрессии. Независимыми параметрами в моделях выступают длина выборки, вид засоренности зависимой и независимых переменных регрессии.

## 1. Введение

При обработке технико-экономической информации для построения статических зависимостей часто используется регрессионный анализ, который требует соблюдения определенных предпосылок и, в частности, нормальности закона распределения остатков. На практике данные, полученные с реальных объектов или процессов, как правило, имеют ограниченную длину выборки, имеют мультиколлинеарность и подвержены влиянию выбросов.

В этих условиях применение регрессионного анализа некорректно. Случаи, когда остатки не имеют нормального закона распределения, чаще всего обусловлены влиянием выбросов. В таких условиях необходимо перейти к другим методам оценивания, в частности, к робастным оценкам.

В настоящее время разработан ряд алгоритмов робастного оценивания, которые имеют ограниченную область использования. В этой связи возникает задача выбора области изменения исходной информации для наиболее эффективного использования тех или иных методов.

В данной работе предлагается методология оценки области применимости нескольких алгоритмов робастного оценивания параметров регрессии. Для этого с помощью имитационного моделирования методом Монте-Карло определяются зависимости качественных показателей алгоритмов от статистических свойств исходных данных. Далее предполагается определить модели этих зависимостей в аналитическом виде, что позволит определять значения критериев эффективности каждого алгоритма для конкретных статистических свойств исходных данных, а следовательно, и наиболее предпочтительный алгоритм оценивания.

## 2. Алгоритмы робастного оценивания параметров регрессии

Пусть математическая модель регрессии представлена в линейном виде

$$y_i = x'_i \beta + u_i,\ i = 1,2,....,n, \qquad (1)$$

где $y_1, y_2, ..., y_n$ – значения выхода; $x_1, x_2, ..., x_n$ – p-размерные векторы независимых переменных в уравнении регрессии; $\beta$ – p-размерный вектор неизвестных параметров, которые будут оценены; $u_1, u_2, ..., u_n$ – независимые от X случайные помехи с распределением F.

В статье рассматриваются восемь алгоритмов робастного оценивания параметров регрессии.

1) Метод наименьших квадратов медиан (LMS) [5]. Оценки получаем от минимизации квадратов медиан ошибок, т.е. решая

$$\min_\beta \operatorname{median}_{i \le n} (y_i - \sum_{k=1}^{p} x_{ki}\beta_k)^2, \qquad (2)$$

LMS имеет точку пробоя близко к 50 %, но из-за своей $n^{-1/3}$ скорости сходимости он имеет нулевую эффективность для центральной гауссовой модели.

2) Метод перевзвешенных наименьших квадратов (RLS) [5], вычисляемый

$$\sum_{i=1}^{n} \omega_i r_i^2, \qquad (3)$$

где $\omega_i = \begin{cases} 1 & ,| r_i / \hat{\sigma} | \le 2.5, \\ 0 & \text{иначе.} \end{cases}$ – веса, вычисленные от LMS остатков и их оценки масштаба

$$\hat{\sigma} = \sqrt{\frac{\sum_{i=1}^{n} \omega_i^0 r_i^2}{\sum_{i=1}^{n} \omega_i^0 - (p+1)}}. \qquad (4)$$

3) Метод функциональных наименьших квадратов (FLS) [1], который является M-оценкой с тригонометрической функцией $\psi$. FLS вычисляется решением уравнения

$$\frac{1}{t}\frac{1}{n^2}\sum_{j=1}^{n}\sum_{i=1}^{n} x_{ik} \sin\{t[(y_i - y_j) - (x_i - x_j)\beta]\} = 0, \quad k = 1,..,p, \qquad (5)$$

где $t \in T$, T – окрестность нуля.

В качестве начального приближения используется LTS (см. ниже).

4) Метод наименьших усеченных квадратов (LTS) [5]. Оценки определяются как p-вектор:

$$\hat{\beta} = \arg\min_\beta \sum_{i=1}^{J} u_{[i]}^2, \qquad (6)$$

где $u_{[1]}^2 \le u_{[2]}^2 \le ... \le u_{[n]}^2$ – упорядоченный ряд квадратов остатков $u_i^2 = (y_i - \sum_{k=1}^{p} x_{ki}\beta_k)^2$; J – самое

большое целое число, меньше чем или равное n/2+1. Он имеет ту же самую асимптотическую эффективность для гауссовой модели, как M-оценки Хьюбера [9]. Точка пробоя LTS – 50 %, его скорость сходимости – $n^{-1/2}$.

5) Метод наименьших абсолютных отклонений (LAD) позволяет получить параметры, решая задачу минимизации:

$$\min_\beta \rho(\beta) = \sum_{k=1}^{n} | y_i - \sum_{k=1}^{p} x_{ki}\beta_k | . \qquad (7)$$

Его асимптотическое распределение, как известно, является нормальным в случае независимых ошибок [9], однако его точка пробоя является нулевой.

6) S-оценка [4]. Для каждого вектора $\beta$ вычисляем дисперсию $S(u_1(\beta),...,u_n(\beta))$, которую получаем как решение

$$1/n\sum_{i=1}^{n}\rho(u_i/S) = b , \qquad (8)$$

где $\frac{b}{\sup_{\rho(u)}} = 0.5$, чтобы обеспечить точку пробоя на уровне 50%.

S-оценка определяется как $\arg\min_\beta S(u_1(\beta),...,u_n(\beta))$ и конечная оценка масштаба определяется как $\hat{\sigma} = S(u_1(\beta),...,u_n(\beta))$.

Мы использовали $\rho$ -функцию с параметром c=1,548 в виде

$$\rho(x) = \begin{cases} 3(x/c)^2 - 3(x/c)^4 + 3(x/c)^6 & ,|x| \le c \\ 1 & ,|x| > c \end{cases} . \qquad (9)$$

С выбрано, чтобы получить точку пробоя 0,5. Увеличение (уменьшение) значения с повышает (понижает) асимптотическую эффективность в центральной гауссовой модели, но понижает (повышает) точку пробоя.

7)MM-оценка [3]. Начальную оценку $\hat{\beta}_0$ с высокой точкой пробоя вычисляем как S-оценку. Далее вычисляем M-оценку с другой $\rho$ -функцией (c=4,687, $\rho_1(u) \le \rho_0(u)$ и $\rho_1(u) = \sup\rho_0(u) = a$), как решение

$$\sum_{i=1}^{n}\psi_1(u_i(\beta)/S_n) = 0 , \qquad (10)$$

которое удовлетворяет $Q(\hat{\beta}_1) \le Q(\hat{\beta}_0)$, где $\psi_1$ – первая производная $\rho_1$ и $Q(\beta) = \sum_{i=1}^{n}\rho_1(u_i(\beta)/S_n)$. MM-оценка имеет такое же асимптотически нормальное распределение, как S-формула оценки.

8) Одношаговая GM-оценка (S1S) [8]. Рассмотрим класс оценок, определенных как решение уравнения

$$\sum_{i=1}^{n}\omega_i\psi(u_i(\beta)/\sigma\varpi_i)x_i = 0 , \qquad (11)$$

$$\omega(x_i) = \min\{1, b/(x_i - m_x)'C_x^{-1}(x_i - m_x)\} ,$$

где b – 95% процентов от $\chi^2(p)$; $m_x, C_x$ – оценка минимума объема эллипсоида (MVE) для оценки положения и ковариации [6];

$$\psi(t) = \min\{1, c/|t|\}\max t .$$

Используя LTS как начальную оценку, одношаговая GM-оценка может быть вычислена итеративно:

$$\beta = \hat{\beta}_0 + [\sum_{i-1}^{n}\psi'(u_i(\hat{\beta}_0)/\hat{\sigma}\omega_i)x_i x'_i]^{-1} \times \times \sum_{i-1}^{n}\hat{\sigma}\omega_i\psi(u_i(\hat{\beta}_0)/\hat{\sigma}\omega_i)x_i \qquad (12)$$

Оценка масштаба остатков получена следующим образом:

$$\hat{\sigma} = 1.4826(1 + 5/(n-p)) \times \text{median}_{i\le n} | u_i(\hat{\beta}_0) | . \qquad (13)$$

Оценка масштаба может быть получена также с помощью Qn алгоритма из [7].

GM-формула оценки имеет асимптотически нормальное распределение и ограниченную функцию влияния.

### 3. Описание экспериментальных исследований

В связи с регресс-эквивариантностью всех оценок все элементы вектора параметров регрессии приняты равными 1. В литературе показано, что несмотря на асимптотическую зависимость некоторых методов от количества независимых переменных регрессии, при выполнении условия n/p>5 и p<10 качество работы методов изменяется незначительно. Поэтому при моделировании будем использовать простую линейную модель без потери общности.

Независимые переменные X в модельном эксперименте генерировались двумя способами: как независимые стандартные нормальные переменные и как аналогичные, содержащие 10% выбросов, образованных перемещением 10% крайних точек правого хвоста вправо на 50. Сгенерированные независимые переменные длиной n использовались без изменений в течение всего моделирования.

Ошибки $u_i$ генерировались как случайная величина с распределением, являющимся загрязненным нормальным распределением вида

$$F(e) = (1-\lambda)N(0,1) + \lambda N(0,\sigma) , \qquad (14)$$

где $\lambda$ – принимает значения 0,05, 0,1, 0,2, 0,3; $\sigma$ – принимает значения 2, 3, 4, 5, 10, 15.

В данной модели степень загрязнения имеет прямую зависимость от значений $\lambda$ и $\sigma$.

Рассматривалось три значения длины выборок – n=50, 100, 200.

Для сравнения робастных методов будем использовать несколько критериев. Первый из них – среднеквадратичная ошибка (MSE). Но критерий MSE не является робастным, так как он может быть подвержен высокому влиянию маловероятных со-

бытий. Поэтому необходимо использовать более робастные критерии. В [2] была рассмотрена робастная альтернатива MSE – медиана абсолютных отклонений (MAD), определяемая как

$$MAD = \frac{1}{p}\sum_{i=1}^{p} \text{med}(|\hat{\beta}_i - \beta_i|). \qquad (15)$$

Кроме этих критериев, в [2] рассматривались еще два критерия, отражающие смещение. Это среднее смещение параметров регрессии относительно среднего арифметического и его робастный аналог – среднее смещение относительно медианы. Но результаты, приведенные в [2], показали малую ценность этих критериев для определения наиболее предпочтительного робастного алгоритма. Поэтому в данной работе ограничимся MSE и MAD.

Для корректной интерпретации результатов и возможности их дальнейшего использования в целях определения наиболее предпочтительного метода необходимо применять относительные критерии для сравнения. Поэтому в качестве критериев будем использовать отношения MSE и MAD, полученные для робастного метода, к MSE и MAD, полученные для метода наименьших квадратов:

$$MSE = \frac{MSE_{\text{робастный метод}}}{MSE_{LS}},$$
$$MAD = \frac{MAD_{\text{робастный метод}}}{MAD_{LS}}. \qquad (16)$$

Моделирования выполняли методом Монте-Карло с использованием 200 дублирований.

В нашем модельном эксперименте для генерации нормальных случайных величин использовался программный продукт MAPLE/R5. Генерация смешанного нормального распределения (14) выполнялась следующим образом. Генерировались две нормальных случайных величины $N(0,1)$, $N(0,\sigma)$ и случайная величина дискретного распределения с параметром $\lambda$. Далее в соответствии с дискретно-распределенной выборкой в нормальной выборке $N(0,1)$ заменялись $(n*\lambda)$-элементов на соответствующие элементы выборки $N(0,\sigma)$.

Для вычисления параметров регрессии методами LMS, LTS, RLS использовалась программа PROGRESS, описанная в [5]. Для вычисления оценок методом LAD применялалась подпрограмма вычисления симплекс-методом линейного программирования из MAPLE /R5. Для вычисления S-, MM-, S1S- оценок и FLS использовались подпрограммы библиотеки статистических программ SatLib. Для вычисления параметра масштаба остатков применялась программа Qn [7].

**4. Результаты модельного эксперимента**

Моделирование показало, что результаты, основанные на MSE, грубо совместимы с теми, которые основаны на критерии MAD. Поэтому ограничимся анализом результатов только для MSE.

Относительно увеличения объема выборки все методы, кроме LMS, более точно оценивают параметры модели либо стабилизируются в окрестности его значения для n=100. Для LMS с увеличением выборки качество работы ухудшается. На рис.1 изображены зависимости эффективности RLS и MM-оценок от длинны выборки n.

Для случая нормально распределенных независимых переменных MM- и S1S- оценки работают наилучшим образом, хотя MM-оценки немного лучше. При малой степени засоренности нормального распределения ошибки MM-, S1S-, FLS-оценки показали высокую эффективность и работают так же хорошо, как и метод наименьших квадратов. Оценки с низким уровнем эффективности (LMS, S, LAD) работают намного хуже, чем метод наименьших квадратов. Оставшиеся оценки показывают средние результаты (рис. 2). При увеличении загрязнения качество работы LMS, LAD, RLS, S-оценок улучшается относительно остальных оценок (рис. 3) При максимальном загрязнении ($\lambda$=0,3, $\sigma$=10,15) все методы работают

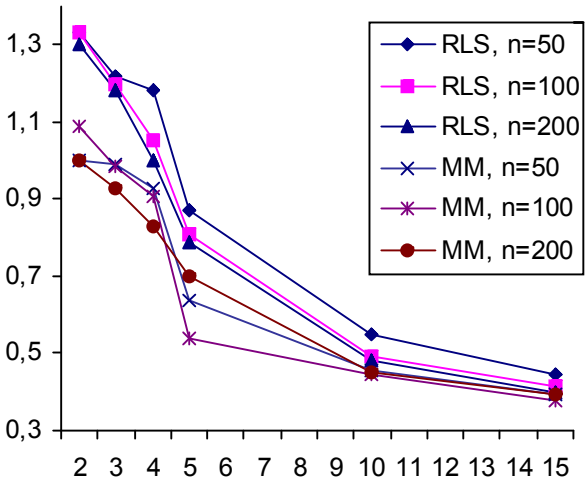


Рис. 1. Зависимость MSE от значения масштаба загрязняющей выборки и длины выборки при $\lambda$=0,05 при нормально распределенной независимой переменной

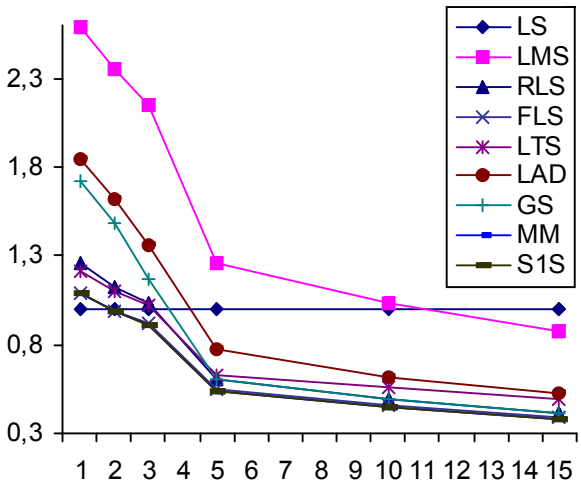


Рис. 2. Зависимость MSE от значения масштаба загрязняющей выборки при нормально распределенной независимой переменной, $\lambda$=0,05, n=100

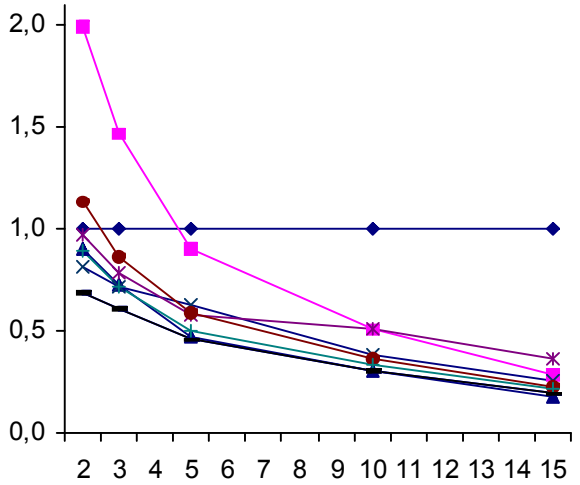


Рис. 3. Зависимость MSE от значения масштаба загрязняющей выборки при нормально распределенной независимой переменной, $\lambda$ =0,3, n=100

практически одинаково, при этом RLS становится наилучшим методом.

Для случая независимых переменных с выбросами все оценки (абсолютный критерий), кроме S1S, показали результаты намного хуже, чем в предыдущем случае. Различие между оценками резко сократилось. LAD и FLS стали хуже, чем остальные. MM- и S-оценки занимают второе и третье место соответственно. RLS, LMS, LTS выполняются практически одинаково и показали средние результаты. S1S-оценки оказались лишь немного хуже, чем для случая нормально распределенных независимых переменных. При любом уровне загрязнения остатков все методы работают намного лучше, чем LS. Значения относительного критерия для независимой переменной с выбросами оказались меньше (рис. 4,5), чем для случая с нормальными независимыми переменными, в связи с тем, что абсолютные значения критерия для LS были намного хуже, чем для робастных методов, в первом случае.

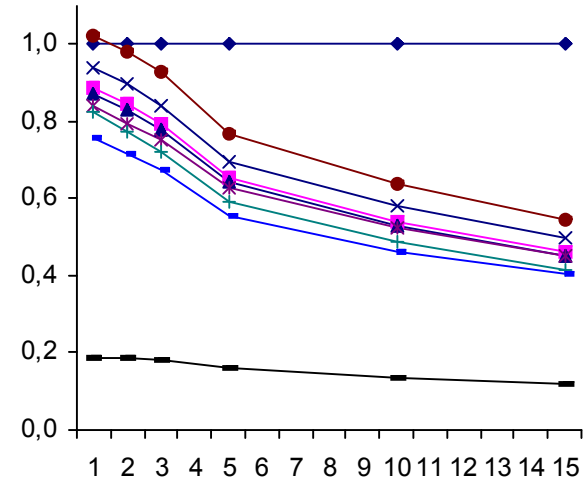


Рис. 4. Зависимость MSE от значения масштаба загрязняющей выборки при независимой переменной с выбросами, $\lambda$ =0,05, n=100

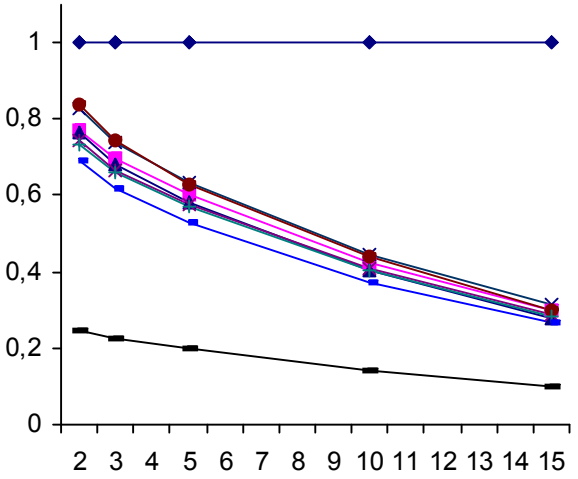


Рис. 5. Зависимость MSE от значения масштаба загрязняющей выборки при независимой переменной с выбросами, $\lambda$ =0,3, n=100

Таким образом, в случае нормально распределенных независимых переменных лучше использовать MM-оценки, так как они наиболее предпочтительны и наименее вычислительно емкие, кроме случая загрязнения ($\lambda$=0,3, $\sigma$=15), для которого наилучшим оказался RLS. Для случая выбросов в независимых переменных наилучший выбор – S1S, так как остальные методы плохо справляются с выбросами двух типов – в независимых переменных и вертикальными выбросами.

**5. Разработка моделей качественных показателей алгоритмов робастного оценивания**

Результатом указанного выше модельного эксперимента являются зависимости качественных показателей алгоритмов робастного оценивания от статистических свойств исходных данных. Зависимости представляют собой наборы данных, по которым можно построить функциональные зависимости в аналитическом виде.

В модельном эксперименте использовались следующие статистические характеристики исходных данных:

– длина выборки – n;

– наличие выбросов в независимых переменных исходных данных – $\alpha$;

– степень засоренности распределения ошибок (14) регрессионной модели – $\lambda$;

– соотношение масштабов основной и загрязняющей составных частей распределения ошибок (14) регрессионной модели – $\sigma$.

В качестве критериев эффективности использовались MSE и MAD из (16). В связи со сказанным выше модели качественных показателей в общем виде можно представить следующими выражениями:

$$\begin{aligned} MSE_i &= F_i(n,\alpha,\lambda,\sigma), \\ MAD_i &= F'_i(n,\alpha,\lambda,\sigma), \\ i &= 1,2,...,8 \end{aligned} \qquad (17)$$

В моделях (17) $\alpha$ и n имеют качественный характер. Так, $\alpha$ показывает, есть выбросы в независимых переменных или нет, а n – определяет малую, среднюю или большую длину выборки. Поэтому целесообразно исключить эти параметры из представления моделей в виде (17) и построить ряд моделей для конкретных значений $\alpha$ и n в следующем виде:

$$\begin{aligned} &MSE^{\alpha,n}{}_i = F_i(\lambda,\sigma), \\ &MAD^{\alpha,n}{}_i = F'_i(\lambda,\sigma), i = 1,2,...,8. \end{aligned} \qquad (18)$$

Модели в виде (18) представляют собой непрерывные функции, которые позволят определять наиболее предпочтительный алгоритм робастного оценивания посредством решения задач минимизации в виде

$$\begin{aligned} &\min_i MSE^{\alpha,n}{}_i, i = 1,2,...,8, \\ &\min_i MAD^{\alpha,n}{}_i, i = 1,2,...,8. \end{aligned} \qquad (19)$$

В связи с нелинейным характером зависимостей (18) для определения моделей в явном виде можно использовать регрессионный анализ с нелинейным видом уравнения регрессии или нелинейным относительно независимых переменных регрессии. Для параметрической идентификации существуют формальные методы, чего нельзя сказать о структурной идентификации. Чтобы преодолеть проблему структурной идентификации при построении моделей, можно использовать аппарат искусственных нейронных сетей (ИНС). ИНС целесообразно использовать потому, что он является универсальным аппроксиматором, который не требует структурной идентификации. Будем использовать ИНС в режиме обучения, что позволит определить ее синаптические веса. Далее, в соответствии с выбранной топологией ИНС, функциями активации и определенными синаптическими весами, формируются модели качественных показателей алгоритмов робастного оценивания в аналитическом виде.

В данной работе представлена методика построения моделей качественных показателей алгоритмов робастного оценивания. Также определены посредством модельного эксперимента области применимости алгоритмов робастного оценивания и основные статистические характеристики, определяющие их применимость.

**Литература:** 1. *Meintanis, S.G. and G.S. Donatos.* A comparative study of some robust methods for coefficient-estimation in linear regression, Computational Statistics & Data Analysis, 23, (1997). P. 525-540. 2. *You Jiazhong,* A Monte Carlo comparison of several high breakdown and efficient estimators, Computational Statistics & Data Analysis. Vol: 30, Issue: 2, (1998). P. 25-55. 3. *Hennig C.,* Efficient high-breackdown-point estimator in robust regression: which function choose?, Statistics & Decision 13, (1995). P. 221-241. 4. *Rousseeuw, P.J. and Yohai V.J.*, Robust regression by means of S-estimators, in: J.Franke, W.Hardle, R.D.Martin (Eds.), Robust and Nonlinear Time Series Analysis (Lecture Notes in Statistics 26) (New York: Springer-Verlag, 1984). P. 256-272. 5. *Rousseeuw P.J., Hubert M.* Recent development in PROGRESS, Computational Statistics & Data Analysis, 20, (1997). P. 321-340. 6. *Rousseeuw, P.J.* Least median of squares regression, J. Am. Statist. Assoc., 79, (1984). P. 871-880. 7. *Rousseeuw P.J., Croux C.* Explicit Scale Estimator with High Breakdown Point, L1-Statistical Analisis and Reated Methods, (1992). P. 77-92. 8. *Coakley C.W. and Hettmansperger T.P.* A bounded influence, high breakdown, efficient regression estimator, J. Am. Statist. Assoc., 88, (1993). P. 872-880. 9. *Хьюбер П.* Робастность в статистике. М.: Мир, 1984. 302с

Поступила в редколлегию 12.03.2000

**Рецензент:** д-р техн. наук, проф. Путятин В.П.

**Антонов Владислав Александрович,** аспирант кафедры ИУС ХТУРЭ. Научные интересы: робастная статистика. Адрес: Украина, 61172, Харьков, ул. С.Грицевца, 24, кв. 43, тел. 40-94-51.

**Шамша Борис Владимирович,** канд. техн. наук, доцент, профессор кафедры ИУС ХТУРЭ. Научные интересы: обработка данных и управление. Адрес: Украина, 61166, Харьков, ул. Космонавтов, 5, кв. 32, тел. 33-27-78.

УДК 519.7

# ВОЗМОЖНЫЕ ИНТЕРПРЕТАЦИИ ЛОГИЧЕСКОЙ АЛГЕБРЫ

*ЯКИМОВА Н.А.*

Доказывается принципиальная возможность рассматривать как логические алгебры некоторые частные виды алгебр. Устанавливается соответствие между элементами этих алгебр и элементами векторного логического пространства, а также между операциями над элементами этих алгебр и операциями над элементами векторного логического пространства.

Рассмотрим несколько частных алгебр, которые можно представлять как алгебры логического типа. Одной из них является алгебра двоичных кодов. При такой интерпретации логической алгебры берем $p$-компонентные наборы $(a_1, ..., a_p)$ цифр из двухэлементного множества $G=\{0, 1\}$. Им соответствуют векторы булева пространства размерности $p$. При этом нулевому вектору соответствует нулевой набор (0, ..., 0), а единичному - единичный набор (1, ..., 1). В роли базисных векторов $a_1$, ..., $a_p$ используются всевозможные двоичные наборы, в состав которых входит по одной единице: (1, 0, ..., 0), (0, 1, ..., 0), ..., (0, ..., 0, 1) [1]. Под дизъюнкцией базисных векторов $a_{i_1} \vee a_{i_2} \vee ... \vee a_{i_t}$ в алгебре двоичных кодов понимается набор, у которого на $i_1$, $i_2$, ..., $i_t$-х местах стоят единицы, а на остальных местах – нули. Таким образом, каждый двоичный код можно единственным образом представить в виде линейной комбинации базисных кодов.

Дизъюнкция векторов булева пространства отвечает дизъюнкции соответствующих двоичных кодов: $(a_1, ..., a_p) \vee (b_1, ..., b_p) = (a_1 \vee b_1, ..., a_p \vee b_p)$. Конъюнкции векторов при двоично-кодовой интерпретации логической алгебры соответствует конъюнкция двоичных наборов: $(a_1, ..., a_p) \wedge (b_1, ..., b_p) = (a_1 \wedge b_1,$