

## СТАТИСТИЧЕСКИЕ МЕТОДЫ И ТЕОРИЯ ВЕРОЯТНОСТЕЙ

### ВЫБОР НАИЛУЧШЕЙ РЕГРЕССИИ ДЛЯ ПРОГНОЗА С ИСПОЛЬЗОВАНИЕМ ЭКЗАМЕНУЮЩЕЙ ВЫБОРКИ

Авербух Е. А.

(Калинин)

Предложен математико-статистический критерий выбора наилучшей регрессии с учетом точности прогноза по обучающей и экзаменуемой выборкам при существенном различии в них уровня шума.

Регрессионные модели довольно широко используются в прогнозировании [1, 2]. Условия их эксплуатации и обучения могут значительно различаться, в частности, уровнем внешнего шума из-за неодинаковых режимов наблюдения (измерения) зависимых переменных, присутствия неконтролируемых помех и т. п. Это сильно усложняет выбор наилучшей регрессии, поскольку, с одной стороны, такие распространенные критерии, как  $C_p$ -статистика [3], статистика AEV, предложенная в [4], AIC [5] или MSEF [6], а также обобщающий их критерий [7] не позволяют одновременно учитывать точность прогноза по обучающей и экзаменуемой выборкам, что затрудняет выбор и снижает эффективность ее автоматизации. С другой стороны, выбор наилучшей регрессии с учетом точности прогноза на экзаменуемой выборке можно условно классифицировать следующим образом.

I. Процедуры, не требующие априорной информации о виде закона распределения наблюдений зависимой переменной:

I.1. те, в которых в качестве основного решающего правила при выборе наилучшего вида модели используют невязку (в среднем квадратичном) на всей экзаменуемой выборке [2, 8] и ее производные [9];

I.2. типа jack-knife [1, 10].

II. Базирующиеся на априорной информации о законе распределения наблюдений зависимой переменной [11].

Все они не предназначены для использования в случаях, когда условия обучения и эксплуатации модели различаются между собой по уровню шума. Кроме того, иерархия решающих правил в процедурах I.1 существенно затягивает время поиска оптимального решения. Эффективность же использования процедур второго типа очень чувствительна к отклонению от нормальности зависимых переменных.

Цель настоящей работы — вывод математико-статистического критерия для выбора наилучшей регрессии, позволяющего одновременно учитывать точность прогноза по обучающей и экзаменуемой выборкам даже при существенном различии в них уровня шума, а также анализ работоспособности предложенного критерия на модельных примерах.

#### ПОСТАНОВКА ЗАДАЧИ

Рассмотрим классическую схему линейной регрессии

$$Y_i = X_i B + E_i = W_i B_1 + V_i B_2 + E_i, E_i \sim N_{n_i}(0, \sigma_i^2 I_{n_i}), \quad (1)$$

где  $i$  — индекс независимых выборок экспериментальных данных;  $i=1$  для обучающей,  $i=2$  — для экзаменуемой выборки;  $Y_i$  —  $n_i$ -вектор наблю-

дений зависимой переменной;  $E_i$  —  $n_i$ -вектор шума;  $X_i = (x_1, \dots, x_{n_i}) = [W_i | V_i]$  — блочная  $n_i \times p$ -матрица наблюдений  $p$  независимых переменных (регрессоров), причем  $W_i = (w_{i1}, \dots, w_{in_i})^T$  —  $n_i \times k$ -матрица,  $V_i = (v_{i1}, \dots, v_{in_i})^T$  —  $n_i \times m$ -матрица,  $k+m=p$ ;  $B^T = [B_1^T | B_2^T]$  —  $p$ -вектор неизвестных параметров, причем  $B_1$  —  $k$ - и  $B_2$  —  $m$ -векторы;  $I_{n_i}$  — единичная  $n_i \times n_i$ -матрица. Разбиение матрицы  $X_i$  и вектора  $B$  на блоки произвольное ( $k=1, \dots, p$ ,  $m=1, \dots, p$  при  $k+m=p$ ), допускающее любые сочетания регрессоров в блоках. Следовательно, принятые обозначения наряду с удобством не ограничивают общности рассуждений.

В качестве прототипа для предлагаемого критерия выберем довольно универсальный критерий Янга  $H_{1i}(\rho)$ ,  $i=1$  [7], состоящий в минимизации нормированной, т. е. отнесенной к дисперсии шума  $\sigma_1^2$  суммы дисперсии прогноза зависимой переменной на обучающей выборке и взвешенного квадрата его смещения по модели оценками  $\hat{B}_1$ , вычисленными методом наименьших квадратов (м. н. к.) и равными

$$\hat{B}_1 = (W_1^T W_1)^{-1} W_1^T Y_1, \quad (2)$$

вида

$$H_{1i}(\rho) = \sum_{j=1}^{n_i} (\omega_{ij} \hat{B}_1 - \omega_{ij} M[\hat{B}_1])^2 / \sigma_i^2 + \rho \sum_{j=1}^{n_i} (\omega_{ij} M[\hat{B}_1] - x_{ij} B)^2 / \sigma_i^2, \quad \rho > 0, \quad i=1, \quad (3)$$

где  $M[\cdot]$  — математическое ожидание (МО).

Для этого пользуются статистикой  $C_{11}(\rho)$ , вычисляемой по формуле

$$C_{11}(\rho) = \rho (S_{11}^2 / S_{011}^2) + (1 + \rho) k - \rho n_1 = \rho C_p + (1 - \rho) k, \quad \rho > 0, \quad (4)$$

частными случаями которой при  $\rho=1$  являются  $C_p$ -статистика, при  $\rho = k/(n_i - k)$  — статистика AEV, при  $\rho > 1$  — статистика AIC, при  $\rho=1$  и  $n_i \gg k$  — MSEF и т. д. В (4) приняты обозначения:  $S_{011}^2$  — остаточная дисперсия для полной модели с оценками параметров  $B$ :  $S_{0ii}^2 = Y_i^T (I_{n_i} - X_i \times (X_i^T X_i)^{-1} X_i^T) Y_i / (n_i - p)$ ,  $i=1$ ;  $S_{1i}^2$  — остаточная сумма квадратов на выборке  $i$  для модели с оценками параметров  $\hat{B}_1$

$$S_{1i}^2 = (W_i \hat{B}_1 - Y_i)^T (W_i \hat{B}_1 - Y_i). \quad (5)$$

Введение весового коэффициента  $\rho$  в (3) позволяет в отличие от других критериев формировать модель с учетом ее предполагаемого использования. Как известно [11], «перебор» регрессоров ведет к увеличению дисперсии прогноза, а «недобор» — к росту его смещения. Поэтому модель с минимальным значением  $C_{11}(\rho)$  при  $\rho$ , близком к нулю, наиболее пригодна для прогноза, а при  $\rho > 1$  — для оценивания. Эффективность же критерия (3) — (4) недостаточно высока не только из-за отсутствия учета точности прогноза модели на экзаменующей выборке, но и ввиду большой чувствительности статистики (4) к значению остаточной суммы квадратов  $S_{011}^2$ , которое может произвольно изменяться в зависимости от того, какую модель исследователь считает полной.

С целью повышения надежности и эффективности процедуры выбора вида регрессионных моделей прогноза, расширим критерий Янга на случай, когда условия обучения и эксплуатации модели различны, обеспечивая одновременный учет точности прогноза как на обучающей, так и экзаменующей выборках.

# РЕШЕНИЕ

Рассмотрим штрафную функцию  $H(\rho)$  вида

$$H(\rho) = \sum_{i=1}^2 H_{1i}(\rho), \quad (6)$$

включающую по сравнению с (3) дополнительно нормированную, т. е. отнесенную к дисперсии шума  $\sigma_2^2$  сумму дисперсии прогноза зависимой переменной на экзаменуемой выборке и взвешенного квадрата его смещения. Покажем, что соответствующая статистика  $C(\rho)$ , являющаяся несмещенной оценкой среднего значения функции  $H(\rho)$ , имеет вид

$$C(\rho) = \sum_{i=1}^2 C_{1i}(\rho), \quad (7)$$

где

$$C_{12}(\rho) = \frac{\rho S_{12}^2}{S_{011}^2 \hat{\gamma}} - \rho n_2 + \frac{(1-\rho)}{\hat{\gamma}} \text{tr} \{ (W_2^T W_2) (W_1^T W_1)^{-1} \}, \quad (8)$$

$$\hat{\gamma} = S_{022}^2 (n_1 - k - 2) / S_{011}^2 (n_2 - k). \quad (9)$$

Равенство  $M[C(\rho)] = M[C_{11}(\rho)] + M[C_{12}(\rho)]$  очевидно, в [7] доказано, что  $M[C_{11}(\rho)] = M[H_{11}(\rho)]$ , а в [12] — что  $\hat{\gamma}$  является несмещенной и состоятельной в среднем квадратичном оценкой параметра  $\gamma$

$$\gamma = \sigma_2^2 / \sigma_1^2 > 0, \quad (10)$$

причем все суммы квадратов  $S_{0ii}^2$  и  $S_{1i}^2 V$   $i=1, 2$  взаимонезависимы. Поэтому достаточно убедиться в справедливости

$$M[C_{12}(\rho)] = M[H_{12}(\rho)]. \quad (11)$$

Подставляя в (5) соотношение (1) для  $i=2$ , на основании свойств МО квадратичных форм [3, с. 23] с учетом независимости обучающей и экзаменуемой выборок и вида модели (1) ( $M[E_2]=0$ ) можно записать

$$\begin{aligned} M[S_{12}^2] &= [W_2 M[\hat{B}_1] - X_2 B]^T [W_2 M[\hat{B}_1] - X_2 B] + \text{tr} \{ D[W_2 \hat{B}_1 - X_2 B - E_2] \} = \\ &= [W_2 M[\hat{B}_1] - X_2 B]^T [W_2 M[\hat{B}_1] - X_2 B] + \\ &\quad + \sigma_2^2 n_2 + \sigma_2^2 \text{tr} \{ W_2^T W_2 (W_1^T W_1)^{-1} \}. \end{aligned} \quad (12)$$

В то же время поскольку для оценок  $\hat{B}_1$  (2) справедливо [13, с. 95]

$$M[\hat{B}_1] = (W_1^T W_1)^{-1} W_1^T V_1 B_2 + B_1,$$

то с учетом (1) запишем

$$\sum_{j=1}^{n_2} (\omega_{2j} \hat{B}_1 - \omega_{2j} M[\hat{B}_1])^2 = \Delta \hat{B}_1^T W_2^T W_2 \Delta \hat{B}_1,$$

где  $\Delta \hat{B}_1 = (W_1^T W_1)^{-1} W_1^T E_1$  и, следовательно,  $M[\Delta \hat{B}_1] = 0$ ,  $D[\Delta \hat{B}_1] = \sigma_1^2 (W_1^T W_1)^{-1}$ , а по свойству МО квадратичных форм [3, с. 22] имеет место равенство

$$M \left[ \sum_{j=1}^{n_2} (\omega_{2j} \hat{B}_1 - \omega_{2j} M[\hat{B}_1])^2 \right] = \sigma_1^2 \text{tr} \{ W_2 (W_1^T W_1)^{-1} W_2^T \}. \quad (13)$$

Принимая во внимание (12), (13), а также то, что  $S_{011}^2$  — несмещенная оценка  $\sigma_1^2$  и не зависит ни от  $S_{11}^2$ , ни от  $S_{12}^2$  [3, с. 56], убеждаемся в справедливости (11), что и требовалось доказать.

Таким образом, нами предложен расширенный критерий Янга вида (6) — (7), менее чувствительный к выбору полной модели и позволяющий

одновременно учитывать ее реальную прогнозирующую способность. Процедура выбора вида модели на базе этого критерия сводится к:

формированию двух независимых выборок объемами, желательными,  $n_i \geq 25 + 2p$ ,  $i = 1, 2$ , с возможно близкими структурами, т. е. размахом по независимым переменным и  $n_1 \cong n_2$  [8];

оцениванию параметра  $\hat{\gamma}$  по (9), если он априорно не известен;

присваиванию множителю  $\rho$  конкретного значения в зависимости от предполагаемого использования модели ( $\rho = 0,2 - 0,5$  — для моделей прогноза [7]);

оцениванию параметров  $\hat{B}$ , для всевозможных наборов регрессоров по (2);

выбору модели с минимальным значением  $C(\rho)$ , вычисленным по (7).

### ИМИТАЦИОННОЕ МОДЕЛИРОВАНИЕ

Для проверки работоспособности предложенного критерия и установления области его эффективного применения проведены исследования на микроЭВМ «Электроника» МС 1201,02 на модели вида

$$y_{ji} = 2,5x_{01ji} + 7x_{02ji} + 0x_{03ji} + e_{ji}, \quad j = 1, \dots, n_i, \quad i = 1, 2. \quad (14)$$

Исследования проводились в два этапа. На первом изучена работоспособность критерия (7) для случая, когда условия обучения и эксплуатации модели в целях прогноза идентичны ( $\gamma = 1$ ). На втором проанализирована ситуация  $\gamma \neq 1$ .

**Этап 1.** Выборочные значения наблюдений независимых переменных задавали генератором независимых равномерно распределенных чисел [14] в диапазонах, одинаковых для всех трех переменных в рамках одной выборки, равных:  $X_1 \in [0; 2]$  в сочетании с  $X_2 \in [0; 1]$ ;  $X_2 \in [0; 2]$ ;  $X_2 \in [0; 3,5]$ . Объемы выборок варьировали следующим образом:  $n_1 = 20$ ,  $n_2 = 60$ ,  $n_1 = 30$ ,  $n_2 = 30$ ,  $n_1 = 60$ ,  $n_2 = 20$ . Аддитивный шум  $e_{ji}$  задавали аналогичным генератором нормально распределенных чисел с нулевым математическим ожиданием и относительной дисперсии (в % от средней выборочной величины истинных значений зависимой переменной), равной  $\sigma_{\text{от}}^2 = \sigma_{\text{от}}^2 = 5$ ; 15%. Для множителя  $\rho = 0,2$  и  $\rho = 0,8$  проводили по 100 независимых розыгрышей и вычисляли среднее значение (МОЖ) и среднеквадратичное отклонение (СКО) следующих статистик: евклидовой нормы вектора  $\|\hat{B} - B\|$ ,  $C$ , MSEF, AEF,  $C_{11}(\rho)$ ,  $C(\rho)$  и коэффициента детерминации KDET для всевозможных наборов регрессоров.

В результате коэффициент детерминации оказался слабо чувствительным к изменению вида модели во всех рассмотренных случаях.

Для иллюстрации отдельные результаты исследования приведены в табл. 1. При этом набор регрессоров обозначен трехмерным вектором из нулей и единиц таким, что включение  $j$ -го регрессора в модель соответствует единице в  $j$ -й позиции этого вектора. Так, в табл. 1 с целью анализа влияния объемов выборок и уровня аддитивного шума  $\sigma_{\text{от}}^2$  приведены результаты вычислений для  $X_1 \in [0; 2]$ ,  $X_2 \in [0; 2]$ ,  $\rho = 0,2$ . Полученные данные свидетельствуют о том, что во всех рассмотренных ситуациях наименее предпочтительно использование статистик  $C$  и MSEF, на основании которых наилучшей считается избыточная модель (1,1,1), в которой присутствуют все три независимые переменные (в табл. 1 наименьшие значения МОЖ статистик подчеркнуты). Более надежный выбор наилучшего вида регрессии (1,1,0) дают критерии AEF и  $C(\rho)$ ,  $\rho = 0,2$ . Причем достоверность заключения возрастает с увеличением объема выборки  $n_1$  и уровня внешнего шума  $E$ , а для  $C(\rho)$  — при  $n_1 = n_2$ . Критерий  $C_{11}(\rho)$  в рассмотренных в табл. 1 ситуациях несколько уступает критерию  $C(\rho)$  при наибольшем объеме выборки  $n_1 = 60$ . Критерий AEF также имеет больший разброс по сравнению с  $C(\rho)$ , зачастую соизмеряемый с различием между значениями AEF для наборов регрессоров (1,1,1) и (1,1,0).

Для иллюстрации влияния множителя  $\rho$ , а также структуры обучающей и экзаменующей выборок, отдельные результаты исследования кри-

## Результаты модельных исследований

| Наименование статистик | Набор информативных переменных | $n_1=20, n_2=60$    |      |                      |      |
|------------------------|--------------------------------|---------------------|------|----------------------|------|
|                        |                                | $\sigma_{от}^2=5\%$ |      | $\sigma_{от}^2=15\%$ |      |
|                        |                                | МОЖ                 | СКО  | МОЖ                  | СКО  |
| $\ \hat{B}-B\ $        | 1, 1, 1                        | 0,28                | 0,2  | 0,83                 | 0,4  |
|                        | 0, 1, 1                        | 2,85                | 0,0  | 2,92                 | 0,1  |
|                        | 0, 1, 0                        | 3,05                | 0,1  | 3,06                 | 0,2  |
|                        | 1, 1, 0                        | 0,18                | 0,1  | 0,54                 | 0,3  |
| $C_p$                  | 1, 1, 1                        | 4,59                | 6,1  | 4,59                 | 6,1  |
|                        | 0, 1, 1                        | 240,3               | 35,2 | 30,79                | 12,3 |
|                        | 0, 1, 0                        | 294,2               | 49,3 | 34,81                | 17,0 |
|                        | 1, 1, 0                        | 6,03                | 6,6  | 6,06                 | 6,6  |
| MSEP                   | 1, 1, 1                        | 2,73                | 0,9  | 24,56                | 8,1  |
|                        | 0, 1, 1                        | 37,68               | 5,2  | 61,80                | 16,3 |
|                        | 0, 1, 0                        | 45,9                | 7,2  | 69,80                | 22,5 |
|                        | 1, 1, 0                        | 3,24                | 1,0  | 29,14                | 8,7  |
| AEV                    | 1, 1, 1                        | 6,94                | 2,3  | 62,63                | 2,7  |
|                        | 0, 1, 1                        | 67,8                | 9,3  | 111,26               | 29,3 |
|                        | 0, 1, 0                        | 43,6                | 6,9  | 66,27                | 21,4 |
|                        | 1, 1, 0                        | 5,83                | 1,7  | 52,5                 | 15,7 |
| $C_{11}(\rho)$         | 1, 1, 1                        | 3,32                | 1,3  | 3,31                 | 1,2  |
|                        | 0, 1, 1                        | 49,7                | 7,0  | 7,76                 | 2,5  |
|                        | 0, 1, 0                        | 59,6                | 9,9  | 7,76                 | 3,4  |
|                        | 1, 1, 0                        | 2,8                 | 1,3  | 2,85                 | 1,5  |
| $C(\rho)$              | 1, 1, 1                        | 10,47               | 9,1  | 5,74                 | 3,1  |
|                        | 0, 1, 1                        | 52,03               | 9,0  | 5,9                  | 2,9  |
|                        | 0, 1, 0                        | 52,8                | 12,1 | 6,0                  | 4,1  |
|                        | 1, 1, 0                        | 8,02                | 6,4  | 4,8                  | 3,4  |

териев  $C_{11}(\rho)$  и  $C(\rho)$  сведены в табл. 2. Как видно из табл. 1, оба критерия очень чувствительны к «недобору» переменных (0,1,0), а также к одновременному «недобору» и «перебору» (0,1,1), что вполне соответствует соотношению евклидовых расстояний между векторами оценок и истинных значений параметров  $\|\hat{B}-B\|$ . Поэтому в табл. 2 сравнивается взаимное расположение функций  $C_{11}(\rho)$  и  $C(\rho)$  для самого сложного в рассматриваемой модели случая — предпочтения истинной модели (1,1,0) по сравнению с полной моделью (1,1,1). При этом в качестве меры для сравнения критериев, характеризующих их мощность при построении моделей прогноза, представлены средние значения и СКО точек пересечения  $\rho^*$  функций одного и того же критерия для моделей (1,1,1) и (1,1,0), ниже (отмечены знаком  $<$ ) или выше ( $>$ ) которых по этому критерию можно отдать предпочтение истинной модели (1,1,0). Заметим, что по обоим критериям на всем допустимом интервале  $\rho > 0$  отдается предпочтение модели (1,1,0) по сравнению с (0,1,1) и (0,1,0).

Анализ данных, приведенных в табл. 2, позволяет сделать следующие выводы.

1. С ростом объема  $n_1$ , а также с приближением структуры экзаменуемой выборки к структуре обучающей и с уменьшением уровня внеш-

Таблица 1

 $(q=0,2, X_1 \subset [0; 2], X_2 \subset [0; 2])$ 

| $n_1=n_2=30$                     |                            |                                 |                              | $n_1=60, n_2=20$                |                             |                                  |                              |
|----------------------------------|----------------------------|---------------------------------|------------------------------|---------------------------------|-----------------------------|----------------------------------|------------------------------|
| $\sigma_{от}^2=5\%$              |                            | $\sigma_{от}^2=15\%$            |                              | $\sigma_{от}^2=5\%$             |                             | $\sigma_{от}^2=15\%$             |                              |
| МОЖ                              | СКО                        | МОЖ                             | СКО                          | МОЖ                             | СКО                         | МОЖ                              | СКО                          |
| 0,20<br>2,76<br>2,95<br>0,15     | 0,1<br>0,0<br>0,0<br>0,1   | 0,59<br>2,82<br>2,92<br>0,39    | 0,3<br>0,1<br>0,1<br>0,4     | 0,31<br>2,93<br>3,02<br>0,16    | 0,2<br>0,1<br>0,0<br>0,1    | 0,92<br>3,02<br>3,05<br>0,46     | 0,5<br>0,2<br>0,2<br>0,3     |
| 8,43<br>503,9<br>542,6<br>7,88   | 6,8<br>54,6<br>55,5<br>8,2 | 8,43<br>65,2<br>58,8<br>8,09    | 6,8<br>21,1<br>16,9<br>8,3   | 4,48<br>616,9<br>814,5<br>9,75  | 8,8<br>90,6<br>96,1<br>16,1 | 4,49<br>70,06<br>88,86<br>5,94   | 8,8<br>21,1<br>31,5<br>10,2  |
| 4,67<br>76,31<br>82,16<br>4,88   | 1,0<br>7,9<br>8,0<br>1,2   | 42,06<br>118,6<br>112,8<br>43,0 | 8,9<br>27,4<br>22,0<br>10,8  | 2,85<br>45,70<br>59,68<br>3,2   | 0,7<br>6,4<br>6,7<br>1,1    | 25,61<br>71,02<br>84,93<br>28,10 | 5,9<br>18,8<br>21,1<br>6,8   |
| 12,62<br>142,44<br>79,42<br>9,11 | 2,7<br>14,7<br>7,7<br>2,2  | 113,6<br>221,3<br>109,0<br>80,3 | 23,9<br>51,1<br>21,3<br>20,1 | 7,90<br>86,83<br>58,18<br>6,09  | 1,8<br>12,1<br>6,6<br>2,1   | 71,08<br>134,95<br>82,8<br>53,4  | 16,2<br>35,7<br>20,6<br>12,9 |
| 4,09<br>102,4<br>109,3<br>3,18   | 1,4<br>10,9<br>11,1<br>1,4 | 4,08<br>14,64<br>12,60<br>3,02  | 1,5<br>4,2<br>3,4<br>1,7     | 3,30<br>124,98<br>163,7<br>3,6  | 1,8<br>18,1<br>19,3<br>3,2  | 3,30<br>15,61<br>18,55<br>2,8    | 2,8<br>5,6<br>6,3<br>3,03    |
| 9,49<br>248,2<br>277,3<br>7,13   | 0,8<br>14,5<br>15,0<br>0,5 | 9,49<br>34,48<br>32,68<br>7,18  | 1,3<br>5,1<br>5,4<br>1,2     | 10,36<br>205,7<br>290,5<br>9,18 | 0,4<br>22,2<br>24,5<br>0,5  | 1,75<br>133,2<br>175,80<br>0,84  | 0,4<br>15,4<br>16,7<br>0,6   |

него шума чувствительность критерия  $C(\rho)$  к изменению  $\rho$  резко снижается.

2. С возрастанием объема  $n_i$  вероятность принятия ложной гипотезы о том, что наилучший вид модели (1,1,1) по критерию  $C_{11}(\rho)$  увеличивается почти до единицы за счет сужения интервала значений  $\rho$ , на котором возможен правильный выбор наилучшей регрессии.

Таким образом, исследования подтверждают работоспособность предложенного критерия и процедуры выбора вида регрессионной модели с его использованием при  $\gamma=1$  (в особенности, когда имеется возможность формирования экзаменуемой выборки, близкой по структуре к обучающей).

**Этап 2.** На основании выводов, полученных в результате моделирования идентичных условий обучения модели и эксплуатации ее в целях прогноза ( $\gamma=1$ ), изучение работоспособности предложенного критерия  $C(\rho)$  в условиях  $\gamma \neq 1$  проводилось для этой же модели (14), по следующей схеме.

Объемы выборок устанавливали равные между собой:  $n_1=n_2=20$  и  $n_1=n_2=60$ . По аналогии с первым имитационным экспериментом выборочные значения наблюдений независимых переменных задавали генераторы

Значения точек пересечения функций  $C_{11}(q)$  и  $C(q)$ 

| Наименование статистик | Уровень шума, $\sigma_{от}^2$ , % | $n_1=20, n_2=60$     |      |                      |      |
|------------------------|-----------------------------------|----------------------|------|----------------------|------|
|                        |                                   | $X_2 \subset [0; 2]$ |      | $X_2 \subset [0; 1]$ |      |
|                        |                                   | МОЖ                  | СКО  | МОЖ                  | СКО  |
| $C_{11}(\rho)$         | 5                                 | $<0,69$              | 0,13 | $<0,69$              | 0,13 |
|                        | 15                                | $<0,38$              | 0,11 | $<0,38$              | 0,11 |
| $C(\rho)$              | 5                                 | $<1,0$               | 0,08 | $<0,86$              | 0,10 |
|                        | 15                                | $<0,8$               | 0,14 | $<0,59$              | 0,12 |

Результаты модельных исследований

| Наименование статистик | Набор информативных перемен | Тип результата | $n_1=n_2=20$         |            |            |
|------------------------|-----------------------------|----------------|----------------------|------------|------------|
|                        |                             |                | $\sigma_{1от}^2=5\%$ |            |            |
|                        |                             |                | $\gamma=0,6$         | $\gamma=2$ | $\gamma=3$ |
| $C(\rho)$              | 1, 1, 1                     | МОЖ            | 15,40                | 4,69       | 4,08       |
|                        |                             | СКО            | 2,6                  | 1,6        | 1,5        |
|                        | 0, 1, 1                     | МОЖ            | 87,04                | 67,67      | 61,73      |
|                        |                             | СКО            | 13,4                 | 8,8        | 7,1        |
|                        | 0, 1, 0                     | МОЖ            | 183,90               | 90,66      | 79,83      |
|                        |                             | СКО            | 14,5                 | 7,7        | 7,4        |
|                        | 1, 1, 0                     | МОЖ            | 9,95                 | 3,35       | 2,99       |
|                        |                             | СКО            | 1,7                  | 1,9        | 1,8        |
| $\hat{C}(\rho)$        | 1, 1, 1                     | МОЖ            | 17,60                | 5,25       | 4,55       |
|                        |                             | СКО            | 6,3                  | 2,1        | 1,9        |
|                        | 0, 1, 1                     | МОЖ            | 130,71               | 127,43     | 127,67     |
|                        |                             | СКО            | 17,6                 | 12,9       | 18,2       |
|                        | 0, 1, 0                     | МОЖ            | 151,20               | 152,07     | 152,59     |
|                        |                             | СКО            | 16,4                 | 17,4       | 17,6       |
|                        | 1, 1, 0                     | МОЖ            | 11,82                | 3,48       | 3,08       |
|                        |                             | СКО            | 5,5                  | 2,7        | 2,6        |

ром независимо равномерно распределенных чисел в диапазонах, одинаковых для всех трех переменных в рамках одной выборки, равных соответственно  $X_1 \subset [0; 2]$  в сочетании с  $X_2 \subset [0; 1]$  и  $X_2 \subset [0; 2]$ . Относительные дисперсии внешнего шума задавали равными  $\sigma_{1от}^2 = 5$  и  $\sigma_{2от}^2 = 3; 10; 15\%$  (т. е.  $\gamma = 0,6; 2; 3$  для  $\sigma_{1от}^2 = 5\%$ ), а также  $\sigma_{1от}^2 = 15$  и  $\sigma_{2от}^2 = 9; 30\%$  (т. е.  $\gamma = 0,8; 2$  для  $\sigma_{1от}^2 = 15\%$ ). Аналогично для  $\rho = 0,2$  и  $\rho = 0,8$  проводились по 100 независимых розыгрышей и определялись МОЖ и СКО статистик  $C_{11}(\rho)$ ,  $C(\rho)$  для точного значения  $\gamma$  и  $\hat{C}(\rho)$  для оценки параметра  $\gamma$ , вычисленной по (9).

Для иллюстрации чувствительности критерия на базе статистики  $\hat{C}(\rho)$  к изменению объема выборки, уровня внешнего шума, величине  $\gamma$ , а также к погрешности оценивания  $\gamma$  отдельные результаты исследования приведены в табл. 3. Анализ этих результатов позволяет сделать следующие выводы.

1. Почти во всех рассмотренных ситуациях критерии на базе  $C(\rho)$  и  $\hat{C}(\rho)$  отдают предпочтение истинной модели (1,1,0) и очень чувствитель-

для моделей (1, 1, 1) и (1, 1, 0)

| $n_1=n_2=30$         |      |                      |      | $n_1=60, n_2=20$     |      |                      |      |
|----------------------|------|----------------------|------|----------------------|------|----------------------|------|
| $X_2 \subset [0; 2]$ |      | $X_2 \subset [0; 1]$ |      | $X_2 \subset [0; 2]$ |      | $X_2 \subset [0; 1]$ |      |
| МОЖ                  | СКО  | МОЖ                  | СКО  | МОЖ                  | СКО  | МОЖ                  | СКО  |
| $<0,23$              | 0,09 | $<0,23$              | 0,09 | $<0,08$              | 0,12 | $<0,08$              | 0,12 |
| $<0,25$              | 0,10 | $<0,25$              | 0,10 | $<0,13$              | 0,15 | $<0,13$              | 0,15 |
| $<1$                 | 0,12 | $>0,14$              | 0,07 | $<0,92$              | 0,05 | $<1$                 | 0,07 |
| $<1$                 | 0,15 | $>0,21$              | 0,09 | $<1$                 | 0,08 | $<1$                 | 0,08 |

Таблица 3

 $(\rho = 0,2, X_1 \subset [0; 2], X_2 \subset [0; 2])$ 

|                              |            | $n_1=n_2=60$                |            |            |                              |            |  |
|------------------------------|------------|-----------------------------|------------|------------|------------------------------|------------|--|
| $\sigma_{\text{tot}}^2=15\%$ |            | $\sigma_{\text{tot}}^2=5\%$ |            |            | $\sigma_{\text{tot}}^2=15\%$ |            |  |
| $\gamma=0,6$                 | $\gamma=2$ | $\gamma=0,6$                | $\gamma=2$ | $\gamma=3$ | $\gamma=0,6$                 | $\gamma=2$ |  |
| 30,00                        | 4,69       | 10,28                       | 2,17       | 1,72       | 61,58                        | 4,57       |  |
| 7,0                          | 1,6        | 2,5                         | 0,8        | 0,6        | 10,2                         | 0,7        |  |
| 58,25                        | 10,71      | 424,32                      | 143,36     | 107,93     | 375,34                       | 27,76      |  |
| 8,4                          | 2,7        | 20,2                        | 9,7        | 8,87       | 27,7                         | 4,7        |  |
| 67,88                        | 11,45      | 490,87                      | 169,83     | 152,58     | 384,75                       | 28,91      |  |
| 7,2                          | 3,2        | 21,9                        | 13,5       | 12,6       | 18,0                         | 6,1        |  |
| 18,29                        | 3,32       | 4,84                        | 0,79       | 0,20       | 38,52                        | 1,72       |  |
| 4,0                          | 1,8        | 2,7                         | 0,1        | 0,07       | 8,48                         | 0,5        |  |
| 34,42                        | 5,25       | 11,11                       | 2,40       | 1,92       | 67,95                        | 4,98       |  |
| 12,5                         | 2,1        | 5,3                         | 1,1        | 0,7        | 16,41                        | 0,8        |  |
| 43,65                        | 18,17      | 243,08                      | 244,5      | 245,5      | 31,45                        | 30,61      |  |
| 8,2                          | 6,3        | 17,4                        | 17,2       | 17,1       | 6,7                          | 6,2        |  |
| 40,43                        | 19,43      | 290,11                      | 290,6      | 290,9      | 33,83                        | 33,64      |  |
| 6,8                          | 6,6        | 22,6                        | 22,8       | 22,9       | 8,4                          | 8,4        |  |
| 23,28                        | 3,94       | 6,69                        | 0,87       | 0,65       | 43,22                        | 1,92       |  |
| 10,0                         | 3,6        | 5,4                         | 0,2        | 0,17       | 10,2                         | 0,7        |  |

ны к «недобору» (0,1,0) и одновременному «недобору» и «перебору» (0,1,1) информативных регрессоров в модели.

2. Чувствительность статистик  $C(\rho)$  и  $\hat{C}(\rho)$  к «недобору» и одновременному «недобору» и «перебору» регрессоров в модели возрастает с уменьшением  $\gamma$  и  $\sigma_{\text{tot}}^2$  и с увеличением объемов выборок  $n_1=n_2$ .

3. Разброс и смещение статистик  $\hat{C}(\rho)$  из-за неточного определения  $\hat{\gamma}$  снижается с ростом  $\gamma$  и объемов выборок и с уменьшением уровня шума  $\sigma_{\text{tot}}^2$ .

Для иллюстрации влияния множителя  $\rho$ , а также структуры выборок отдельные результаты исследования (по аналогии с данными, представленными в табл. 2) приведены в табл. 4. Их анализ позволяет сделать следующие выводы.

1. Вероятность правильного заключения о лучшем виде регрессии по критерию  $C(\rho)$  при  $\gamma \neq 1$  очень слабо чувствительна к изменениям  $\rho$  и структуры экзаменуемой выборки по отношению к обучающей при сохранении равенства их объемов.

Значения точек пересечения функций  $C(\rho)$  и  $\hat{C}(\rho)$  для моделей (1, 1, 1) и (1, 1, 0)

| Наименование статистик | Уровень шума $\sigma_{\text{шум}}^2$ , % | Параметр $\gamma$ | $n_1=n_2=20$         |      |                      |      | $n_1=n_2=60$         |      |                      |      |
|------------------------|------------------------------------------|-------------------|----------------------|------|----------------------|------|----------------------|------|----------------------|------|
|                        |                                          |                   | $X_2 \subset [0; 2]$ |      | $X_2 \subset [0; 1]$ |      | $X_2 \subset [0; 2]$ |      | $X_2 \subset [0; 1]$ |      |
|                        |                                          |                   | МОЖ                  | СКО  | МОЖ                  | СКО  | МОЖ                  | СКО  | МОЖ                  | СКО  |
| $C(\rho)$              | 5                                        | 0,6               | $<1$                 | 0,12 | $<1$                 | 0,14 | $<1$                 | 0,07 | $<1$                 | 0,08 |
|                        |                                          | 2                 | $<1$                 | 0,08 | $<0,94$              | 0,09 | $<0,98$              | 0,05 | $>0,05$              | 0,05 |
|                        | 15                                       | 0,6               | $<1$                 | 0,11 | $<0,98$              | 0,12 | $<0,93$              | 0,07 | $<0,97$              | 0,08 |
|                        |                                          | 2                 | $<1$                 | 0,07 | $<0,87$              | 0,09 | $<0,89$              | 0,06 | $<0,89$              | 0,07 |
| $\hat{C}(\rho)$        | 5                                        | 0,6               | $<1$                 | 0,15 | $<1$                 | 0,15 | $<0,98$              | 0,09 | $<0,96$              | 0,09 |
|                        |                                          | 2                 | $<1$                 | 0,09 | $<0,95$              | 0,11 | $<0,87$              | 0,05 | $>0,08$              | 0,06 |
|                        | 15                                       | 0,6               | $<1$                 | 0,14 | $<0,94$              | 0,14 | $<0,84$              | 0,10 | $<0,85$              | 0,09 |
|                        |                                          | 2                 | $<0,92$              | 0,92 | $<0,09$              | 0,11 | $<0,85$              | 0,07 | $<0,87$              | 0,09 |

2. Погрешность определения  $\gamma$  по (9) незначительно влияет на разброс точки пересечения функций  $\hat{C}(\rho)$  для моделей (1,1,1) и (1,1,0), оставляя, фактически, без изменения мощность критерия в предлагаемом для практического использования интервале значений  $\rho \in [0,2; 0,5]$  при построении моделей прогноза.

Таким образом, проведенные модельные исследования подтверждают целесообразность применения предложенного критерия на базе статистики  $\hat{C}(\rho)$  для  $\rho=0,2-0,5$  при построении регрессионных моделей для прогноза, даже если условия эксплуатации и обучения различаются по уровню шума в 2—3 раза.

Описанная процедура полностью формализована; она позволяет повысить эффективность автоматизации выбора вида регрессионных моделей.

#### ЛИТЕРАТУРА

1. Айвазян С. А., Енюков И. С., Мешалкин Л. Д. Прикладная статистика: Исследование зависимостей. М.: Финансы и статистика, 1985.
2. Ивахненко А. Г. Долгосрочное прогнозирование и управление сложными системами. Киев: Техника, 1975.
3. Себер Дж. Линейный регрессионный анализ. М.: Мир, 1980.
4. Helms R. W. The Average Estimated Variance Criterion for the Selection-of-Variables Problem in General Linear Models Technometrics. 1974. V. 16. № 3.
5. Akaike H. A. New Look at the Statistical Model Identification//IEEE Transactions on Automatic Control. 1974. AC-19.
6. Aitkin M. A. Simultaneous Inference and the Choice of Variable Subsets in Multiple Regression//Technometrics. 1974. V. 16. N 3.
7. Young A. S. The Bivar-Criterion for Selecting Regressors//Technometrics. 1982. V. 24. N 3.
8. Shelton J. T., Khuri A. I., Cornell J. A. Selecting Check — Points for Testing Lack-of-fit in Response Surface Models//Technometrics. 1983. V. 25. N 4.
9. Cardner M. J. On Using an Estimated Regression Line in a Second Sample//Biometrika. 1972. V. 59. N 2.
10. Snee R. D. Validation of Regression Models: Methods and Examples//Technometrics. 1977. V. 19. N 4.
11. Вапник В. Н. Алгоритмы и программы восстановления зависимостей. М.: Финансы и статистика, 1984.
12. Авербух Е. А. Статистический критерий проверки гипотезы о точности прогнозов одной функции//Экономика и мат. методы. 1989. Т. XXV. Вып. 1.
13. Демиденко Е. З. Линейная и нелинейная регрессии. М.: Финансы и статистика, 1981.
14. Pike M. S.//Comm. ACM. 1965. V. 8. N 10.

Поступила в редакцию  
23 VI 1989