ЭКОНОМИКА 1991
И МАТЕМАТИЧЕСКИЕ МЕТОДЫ Том 27, вып. 2

ПРАКТИЧЕСКИЙ ОПЫТ

ОПЫТ ПРИМЕНЕНИЯ МЕТОДА ДЖЕКНАЙФ В РЕГРЕССИОННОМ АНАЛИЗЕ

Райская Н.Н., Френкель А.А.

(Москва)

Предлагается использовать метод джекнайф для уменьшения смещенности оценок параметров регрессии, полученных методом наименьших квадратов.

В настоящее время широко используется обработка экономической информации на основе регрессионных моделей. Их параметры находятся методом наименьших квадратов, который базируется на определенных предположениях. Невыполнение этих предположений приводит к неоптимальным, в частности смещенным, оценкам параметров. Обычно для получения надежных оценок из исходных данных исключают так называемые аномальные точки, распознаваемые по величине остатков. Однако не всегда большие отклонения рассчитанных по модели значений от фактических соответствуют действительно аномальным наблюдениям для рассматриваемой совокупности.

Следствием формального отбрасывания наблюдений могут быть недостоверные выводы по экономическим совокупностям, где каждый объект является качественной частью целого (отрасли, объединения и т.д.). Если же количество наблюдений невелико, то при статистическом оценивании параметров теряется и число степеней свободы. Это также немаловажно. Таким образом, при подобном подходе нет уверенности, что вновь полученные оценки окажутся устойчивыми. Альтернатива ему — построение модели с помощью непараметрических методов, которые позволяют получать оценки, слабо зависящие от исходных предпосылок и устойчивые при случайных изменениях информации.

Непараметрический подход для учета выборочного смещения впервые предложил М. Кенуй [1]. Идея заключалась в том, чтобы последовательно исключать из рассмотрения по одному наблюдению и проводить вычисления по оставшимся данным. Оценки по всем выборкам в сравнении с первоначальной оценкой по исходной совокупности могут дать информацию о смещении. Дж. Тьюки [2], совершенствовавший этот подход, назвал его Jackknife — "складной нож". Джекнайф дает возможность получать несмещенные и устойчивые оценки параметров при некотором снижении их эффективности в пределах содержательной постановки задачи. Этот подход использует идеи активного эксперимента. Он получил развитие только в последнее время, когда для многочисленных вычислений стали применяться ЭВМ.

Оценка любого параметра θ методом джекнайф рассчитывается следующим образом. Допустим, имеется реализация случайной величины $X(x_1, \ldots, x_{q-1}, x_q, x_{q+1}, \ldots \ldots, x_N)$, для которой определяется оценка $\hat{\theta}$. Последовательно удаляется каждая точка x_q; пересчитывается значение параметра для оставшихся $N-1$ наблюдений $\hat{\theta}_{-q} = (x_1, \ldots, x_{q-1}, x_{q+1}, \ldots, x_N)$ и вычисляется среднее из этих значений

$$\hat{\theta}_{(\cdot)} = \frac{1}{N} \sum_{q=1}^{N} \hat{\theta}_{-q}. \qquad (1)$$

Оценка по джекнайфу

$$\tilde{\theta} = N\hat{\theta} - (N-1)\hat{\theta}_{(\cdot)}. \qquad (2)$$

Разность $\hat{\theta}$ и $\tilde{\theta}$ составит

$$\Delta = (N-1)(\hat{\theta}_{(\cdot)} - \hat{\theta}). \qquad (3)$$

Для простых характеристик, таких, как средняя или дисперсия, Δ является действительной величиной смещения, и расчеты по методу джекнайф дают их строго несмещенные оценки $\tilde{\theta} = \hat{\theta} - \Delta$. Так, для $\hat{\theta} = \bar{x}$ результат вычисления $\hat{\theta}_{(\cdot)}$ по (1) равен $\bar{x}$, $\Delta = 0$ и $\tilde{\theta} = \bar{x}$; для

$$\hat{\theta} = \sigma^2 = \frac{1}{N} \sum_{q=1}^{N} (x_q - \bar{x})^2,$$

$$\Delta = -\frac{1}{N(N-1)} \sum_{q=1}^{N} (x_q - \bar{x})^2.$$

Это приводит к $\tilde{\theta} = \frac{1}{N-1} \sum_{q=1}^{N} (x_q - \bar{x})^2$, т.е. к обычной несмещенной оценке дисперсии.

Для более сложных параметров, в том числе и для коэффициентов регрессии, оценка по методу джекнайф позволяет уменьшить смещение. Это показано в [3, 4].

Коэффициенты регрессии $\hat{\theta} = \hat{\beta}$ оцениваются по методу джекнайф так: находятся значения $\hat{\theta}_{(-q)}$ при последовательном отбрасывании каждого наблюдения, затем пересчитанные параметры заменяются на псевдооценки

$$P_q = N\hat{\theta} - (N-1)\hat{\theta}_{-q} \qquad (4)$$

и рассчитывается их средняя величина. Таким образом, оценка по этому методу

$$\tilde{\theta} = \frac{1}{N} P_q = \sum_{q=1}^{N} [N\hat{\theta} - (N-1)\hat{\theta}_{-q}]/N, \qquad (5)$$

что соответствует данному выше общему определению (2). Вычисление коэффициентов регрессии методом джекнайф через псевдозначения вместо определения (1), (2) связано с их дальнейшим статистическим анализом, при построении доверительных границ с помощью t-критерия.

Регрессионная модель представляется в виде

$$Y = X\beta + \mathrm{E}, \qquad (6)$$

где Y — вектор-столбец зависимой переменной $Y^T = (y_1, \ldots, y_N)$; X — матрица независимых наблюдений $[N \times (n+1)]$, первый столбец которой является единичным вектором; $x_q^T = (x_{q0}, \ldots, x_{qn})$; β — вектор-столбец оцениваемых параметров; E — вектор-столбец отклонений $\mathrm{E}^T = (e_1, \ldots, e_N)$, e_q — независимые, одинаково распределенные, не обязательно по нормальному закону.

Оценки β^T модели (6) методом наименьших квадратов определяются как

$$\hat{\beta} = (X^T X)^{-1} X^T Y. \qquad (7)$$

Обозначим $(X^T X) = C$. Тогда

$$\hat{\beta} = CX^T Y. \qquad (8)$$

Вектор остатков

$$\mathrm{E} = Y - X\hat{\beta} = (I - XC^{-1}X^T)Y. \qquad (9)$$

При отбрасывании q-го наблюдения, т.е. y_q и вектора x_q^T, оценки $\hat{\beta}_{-q}$ вычисляются как*

$$\hat{\beta}_{-q} = \hat{\beta} - \frac{C^{-1}x_q e_q}{1 - x_q^T C^{-1} x_q} = \hat{\beta} - \frac{C^{-1}x_q e_q}{1 - w_q}, \tag{10}$$

а псевдооценки вектора коэффициентов определяются по формуле (4) следующим образом

$$P_q = N\hat{\beta} - (N-1)\hat{\beta}_{-q} = \hat{\beta} + (N-1)\frac{C^{-1}x_q e_q}{1 - w_q}. \tag{11}$$

Среднее значение этих величин равно

$$\tilde{\beta} = N^{-1}\sum_{q=1}^{N} P_q = \hat{\beta} + (N-1)N^{-1}C^{-1}\sum_{q=1}^{N}(1 - w_q)^{-1}x_q e_q. \tag{12}$$

Полученная в (12) величина является оценкой вектора коэффициентов регрессии, рассчитанной методом джекнайф (назовем ее оценкой ДНК или ДН-коэффициентами). Если вычислять $\tilde{\beta}$, используя (2), то результат окажется тем же.

Дисперсия коэффициентов регрессии может быть оценена с помощью псевдооценок (11). Д. Хинкли [6] доказал, что их выборочная дисперсия определяется следующим образом

$$s^2_{\tilde{\beta}} = \frac{1}{N(N-1)}\sum_{q=1}^{N}(P_q - \tilde{\beta})^2, \tag{13}$$

т.е. равна выборочной дисперсии псевдооценок, деленной на N. И величина

$$\frac{\sqrt{N}(\tilde{\beta} - \hat{\beta})}{\left[\,1/N - 1\sum_{q=1}^{N}(P_q - \tilde{\beta})^2\,\right]^{1/2}} \tag{14}$$

имеет асимптотическое T-распределение. Следовательно, можно оценить значимость ДНК регрессии и их доверительные интервалы с помощью t-критерия.

Метод джекнайф использовался при построении модели производительности труда в сахарной промышленности по всей совокупности предприятий (80 заводов). Моделируемый показатель определялся как затраты рабочей силы на переработку 100 т сахарной свеклы (чел/дни). В качестве факторов, влияющих на трудоемкость, взяты следующие восемь показателей: x_1 – установленная мощность завода (тонн свеклы в сутки); x_2 – среднегодовая стоимость промышленно-производственных фондов (тыс.руб); x_3 – длительность хранения свеклы (сут); x_4 – фондовооруженность (руб/т); x_5 – коэффициент использования мощности; x_6 – удельный вес рабочих в общей численности персонала (%); x_7 – потери свеклы при хранении и транспортировке (т); x_8 – потери сахара в производстве к весу переработанной свеклы (%).

По этим переменным была рассчитана регрессионная модель методами наименьших квадратов (МНК) и джекнайф. В табл. 1 приведены полученные оценки параметров, их среднеквадратические ошибки и общие характеристики моделей. Величины коэффициентов регрессии, вычисленных по МНК и ДНК, различны для переменных $x_3 \div x_7$. Это подтверждает предположение о смещении. У переменной x_8 изменилась не только величина, но и знак коэффициента регрессии, что говорит о явной неустойчивости оценки МНК. Среднее квадратическое отклонение коэффициентов регрессии, оцененных по методу джекнайф, оказалось больше соответствующих величин, рассчитанных для МНК-коэффициентов (за исключением β_{x_5}). Ошибка коэффициента регрессии при переменной x_8 в 5–7 раз превышает величину самого коэффициента в обеих моделях, что свидетельствует о необоснованном включении переменной в модель. Коэф-

*Вывод формулы (10) приведен в [5].

Таблица 1

Оценки линейных регрессионных моделей производительности труда

Показатель	МНК		ДНК	
	значение коэффициента регрессии	среднеквадратическое отклонение	значение коэффициента регрессии	среднеквадратическое отклонение
Переменная				
x_0	14,700	16,291	10,957	26,350
x_1	−0,005	0,0006	−0,005	0,0008
x_2	−0,0004	0,0002	−0,0004	0,0002
x_3	−0,039	0,0252	−0,044	0,0287
x_4	−0,091	0,0585	−0,099	0,0808
x_5	−0,099	0,0525	−0,088	0,0518
x_6	0,303	0,1652	0,335	0,2881
x_7	0,332	0,2273	0,376	0,2357
x_8	−0,434	1,9901	0,199	1,5785
Средняя ошибка аппроксимации	16,05		16,20	
Коэффициент множественной детерминации	0,687		0,685	
Коэффициент множественной корреляции	0,829		0,828	

Таблица 2

Границы доверительных интервалов оценок МНК и ДНК

Показатель	МНК		ДНК		Разница в величине концов доверительных интервалов	
	слева	справа	слева	справа	слева	справа
Переменная						
x_1	−0,006	−0,003	−0,006	−0,003	0	0
x_2	−0,0007	0,000	−0,0007	0,000	0	0
x_3	−0,089	0,011	−0,102	0,013	0,013	−0,002
x_4	−0,208	0,025	−0,261	0,061	0,053	−0,036
x_5	−0,204	0,005	−0,191	0,015	−0,010	−0,010
x_6	−0,026	0,632	−0,239	0,909	0,213	−0,267
x_7	−0,120	0,785	−0,090	0,845	−0,030	−0,060
x_8	−4,403	3,533	−2,948	3,346	−1,455	0,187

фициент множественной корреляции для модели МНК равен $R = 0{,}829$; для модели ДНК $R = 0{,}828$, т.е. связь переменных с моделируемым показателем остается такой же тесной. Средняя ошибка аппроксимации осталась на том же уровне.

Были получены доверительные интервалы, которые оказались различными для оценок параметров регрессии МНК и ДНК. В табл. 2 приведены значения левых и правых концов этих интервалов для каждого метода и разница их величин. В прежних границах остались только коэффициенты при x_1 и x_2. Изменение левых и правых границ интервалов остальных коэффициентов не является пропорциональным, что свидетельствует о смещении оценок по МНК.

Можно сказать, что использование оценок ДНК в регрессионной модели позволило

Таблица 3

Отсев переменных по t-критерию

Метод оценки параметров	Номер шага	Расчетное значение t-критерия для параметров*							
		β_{x_1}	β_{x_2}	β_{x_3}	β_{x_4}	β_{x_5}	β_{x_6}	β_{x_7}	β_{x_8}
МНК	1	7,40	2,07	1,54	1,56	1,90	1,83	1,46	0,21
	2	7,57	2,08	1,59	1,56	1,97	1,92	1,46	–
	3	7,40	2,05	1,43	1,54	2,95	1,83	–	–
	4	7,21	2,49	–	1,35	3,69	2,23	–	–
	5	7,13	4,04	–	–	3,80	2,55	–	–
ДНК	1	6,27	1,93	1,54	1,24	1,71	1,16	1,59	0,12
	2	6,40	1,92	1,55	1,22	1,93	1,21	1,67	–
	3	6,32	1,83	1,96	1,76	1,62	–	1,63	–
	4	6,63	1,54	2,49	1,87	–	–	2,56	–
	5	8,78	–	2,73	3,52	–	–	2,39	–

*Табличное значение $T = 2{,}00$.

избежать смещения, имевшего место при оценке по МНК, для $\hat{\beta}_{x_3}$ и $\hat{\beta}_{x_4}$ вправо, $\hat{\beta}_{x_5}$, $\hat{\beta}_{x_6}$, $\hat{\beta}_{x_7}$ – влево.

Более интересным представляется результат применения метода джекнайф для конечного вида модели, построенной с помощью отсева переменных по t-критерию, в сравнении его с конечной моделью, полученной по МНК. В табл. 3 приведены значения t-критерия при многошаговом регрессионном анализе.

Порядок исключения переменных из моделей на втором и третьем шагах оказался разным, что привело к совершенно иной модели производительности труда. Конечные модели оказались

$$\hat{y}_{\text{МНК}} = 7{,}069 - 0{,}004x_1 - 0{,}0006x_2 - 0{,}155x_5 + 0{,}399x_6, \tag{15}$$

$$\hat{y}_{\text{ДНК}} = 34{,}811 - 0{,}005x_1 - 0{,}0725x_3 - 0{,}1805x_4 + 0{,}4467x_7. \tag{16}$$

Модели почти полностью различаются по вошедшим в них переменным (за исключением x_1), хотя количество переменных не изменилось, что нельзя считать закономерным. В конечную модель, полученную методом джекнайф, вошли переменные с высокими коэффициентами вариации: $v_{x_1} = 35{,}4\%$; $v_{x_3} = 25{,}5\%$; $v_{x_4} = 30{,}4\%$; $v_{x_7} = 38{,}6\%$, т.е. существенными оказались переменные с наибольшим разбросом значений. Для сравнения в модели МНК $v_{x_5} = 11{,}4\%$; $v_{x_6} = 2{,}8\%$. Следовательно, модель по методу джекнайф позволяет учесть больше информации, содержащейся в исходных данных, ее параметры являются несмещенными и устойчивыми, а получаемая по ней оценка трудоемкости точнее. Такая модель надежна как для различных нормативных расчетов, так и для анализа резервов повышения производительности труда в сахарной промышленности.

ЛИТЕРАТУРА

1. *Quenouille M.H.* Approximate Tests of Correlation in Time Series // J. Roy Statist. Soc. Ser. B. 1949. V. 11.
2. *Мостеллер Ф., Тьюки Дж.* Анализ данных и регрессия. М.: Финансы и статистика, 1982.
3. *Quenouille M.H.* Notes and Bias in Estimation // Biometrika. 1956. V. 43. № 2.
4. *Эфрон Б.* Нетрадиционные методы многомерного статистического анализа. М.: Финансы и статистика. 1988.
5. *Mielle R.G.* An Unbalanced Jackknife // The Annals of Statisticks. 1974. V. 2. № 5.
6. *Hinkley D.V.* Jackknifing in Unbalanced Situations // Technometrics. 1977. V. 19. № 2.

Поступила в редакцию
25 V 1989