УДК 681.3

МОДЕЛИРОВАНИЕ ВАЛОВОГО РЕГИОНАЛЬНОГО ПРОДУКТА ИРКУТСКОЙ ОБЛАСТИ НА ОСНОВЕ ПРИМЕНЕНИЯ МЕТОДИКИ МНОЖЕСТВЕННОГО ОЦЕНИВАНИЯ РЕГРЕССИОННЫХ ПАРАМЕТРОВ

[1]Баенхаева А.В., [2]Базилевский М.П., [2]Носков С.И.

[1]ФГБОУ ВО «Байкальский государственный университет», Иркутск, e-mail: ayunab2000@mail.ru;
[2]ФГБОУ ВО «Иркутский государственный университет путей сообщения», Иркутск, e-mail: mik2178@yandex.ru

В статье рассматривается методика двухкритериального оценивания параметров линейной регрессии. Для этого необходимо минимизировать векторную функцию потерь, в которой первый частный критерий соответствует методу наименьших модулей, а второй – методу антиробастного оценивания. Результатом оценивания является множество Парето, которое в критериальном пространстве представляет собой объединение ребер образа многогранника. Приведены некоторые приемы, облегчающие работу с множеством оценок, в том числе его точечная характеризация, основанная на программе отсутствия мажорирования. Рассмотренная методика применена для решения задачи прогнозирования объема валового регионального продукта Иркутской области. При этом получены множественные оценки линейной по параметрам и нелинейной по переменным модели регрессии, объясняющими факторами которой являются объем потребления электроэнергии, численность безработных, масштабы строительства жилых домов, оборот розничной торговли.

Ключевые слова: линейная регрессионная модель, многокритериальная оптимизация, множество Парето, валовой региональный продукт, линейное программирование, метод наименьших модулей

MODELING OF GROSS REGIONAL PRODUCT IRKUTSK REGION ON THE BASIS OF METHODS OF MULTIPLE ESTIMATION OF REGRESSION PARAMETERS

[1]Baenkhaeva A.V., [2]Bazilevskiy M.P., [2]Noskov S.I.

[1]Baikal State University, Irkutsk, e-mail: ayunab2000@mail.ru;
[2]Irkutsk State Transport University, Irkutsk, e-mail: mik2178@yandex.ru

The article discusses the technique of two-criterion’s estimation of parameters of linear regression. To do this, it is necessary to minimize the vector-function of loss, in which the first partial criterion corresponds to the method of least module, and the second to anti-robust estimation. The result of estimation is the set of Pareto, which is an association of polyhedron edges of the image in criterion space. We present some techniques to facilitate work with a set of estimates, including the point-characterization, based on the program of the lack of domination. Our procedure is applied to solve the problem of forecasting the volume of the gross regional product of the Irkutsk region. We thus obtained multiple estimates of regression model, which is linear by the parameters and nonlinear by the variables and explanatory factors are the amount of electric power consumption, the number of unemployed, the scale of housing construction, the retail trade turnover.

Keywords: linear regression model, multicriterion optimization, set of Pareto, the gross regional product, linear programming, least absolute deviations

Вопросам оценивания неизвестных параметров регрессионных уравнений посвящена весьма обширная литература [2, 3, 8]. В настоящей статье рассматривается проблема двухкритериального оценивания параметров линейного уравнения регрессии. Стоит отметить, что данная работа методически основана на материале монографии [5].

Главной целью статьи является применение разработанной методики множественного оценивания параметров для решения актуальной задачи моделирования валового регионального продукта Иркутской области.

Краткое описание методики множественного оценивания параметров

Рассмотрим линейное регрессионное уравнение

$$y_k = \sum_{i=1}^{m} \alpha_i x_{ki} + \varepsilon_k, \quad k = \overline{1,n}, \qquad (*)$$

где y – эндогенная (объясняемая, зависимая, выходная), а x_i – i-ая экзогенная (объясняющая, независимая, входная) переменные; α_i – i-й подлежащий оцениванию параметр; ε – ошибки аппроксимации; k – номер наблюдения; n – число наблюдений (длина выборки).

Широкий класс методов оценивания параметров уравнения (*) связан с поиском так называемых L_ν-оценок [2] посредством минимизации функций потерь вида

$$J_\nu(\alpha) = \sum_{k=1}^{n} \left|\varepsilon_k\right|^\nu.$$

Каждая из этих оценок характеризуется реакцией на так называемые выбросы, то есть наблюдения, не согласующиеся со всей выборкой в целом. При этом чем больше значение ν, тем сильнее L_ν-оценка реагирует на выбросы. В регрессионном анализе методы оценивания, слабо реагирующие на выбросы или вообще их игнорирующие, называют робастными.

Методом оценивания параметров уравнения (*), соответствующим $\nu = 2$, является всем хорошо известный и наиболее популярный в регрессионном анализе метод наименьших квадратов (МНК). При $\nu = 1$ это метод наименьших модулей (МНМ), соответствующий городскому расстоянию, а при $\nu \to \infty$ – метод антиробастного оценивания (МАО), соответствующий расстоянию Чебышева. В упомянутых выше источниках описаны способы расчета вектора параметров α в соответствии с каждым из трех методов. Использование МНК приводит к аналитическому выражению $\alpha = (X^TX)^{-1}X^Ty$, а для отыскания вектора α по МНМ и МАО можно использовать аппарат линейного программирования (ЛП). При этом МНМ- и МАО-оценки являются своего рода антиподами: первая вообще игнорирует выбросы, вторая к ним тяготеет.

Это обстоятельство наталкивает на мысль оценивания параметров уравнения (*) одновременно по двум критериям – $J_1(\alpha)$ и $J_\infty(\alpha)$, то есть по векторному критерию $J(\alpha) = (J_1(\alpha), J_\infty(\alpha))$. Это позволило бы максимально увеличить информативность процедуры оценивания, извлечь из выборки всю заключающуюся в ней информацию при построении уравнения (*).

Задача минимизации векторного критерия $J(\alpha)$ относится к классу задач многомерного линейного программирования (МЛП). Под решением такой многокритериальной задачи обычно понимают множество Парето P, которое характеризуется тем, что ни одно паретовское решение не может быть улучшено по какому-либо одному критерию без ухудшения значения другого.

Таким образом, решением задачи оценивания параметров регрессии (*) по двум критериям $J_1(\alpha)$ и $J_\infty(\alpha)$ одновременно будет множество оценок. Назовем его L-множеством по аналогии с L_ν-оценками [4]. Существует фундаментальная работа американских математиков P.L. Yu и M. Zeleny [9], где изложен так называемый многокритериальный симплекс-метод решения задач МЛП.

Обозначим через P^* множество паретовских вершин, через $J(P)$ – образ множества P в критериальном пространстве. В [9] описаны также два способа формирования множества P^*. Первый из них имеет итерационный конструктивный характер. Второй же предполагает применение приема последовательного свертывания критериев. Рассмотрим его более подробно.

Сформируем линейную свертку критериев J_1 и J_∞:

$$J_\gamma(z) = \gamma J_1(z) + (1-\gamma)J_\infty(z), \quad \gamma \in (0,\, 1).$$

Построим на интервале (0, 1) равномерную ε-сеть:

$$0 < \gamma_1 < \gamma_2 < ... < \gamma_l < 1.$$

Для каждого узла γ_i, $i = \overline{1, l}$ решим обычную, со скалярной целевой функцией, задачу ЛП:

$$\min_{z \in Z} J_\gamma(z).$$

В [9] доказано, что ее решением является паретовская вершина. При достаточно мелкой сети, таким образом, формируется все множество P^*.

Безусловно, с такой формой задания модели, в которой параметры определены неявным образом, работать трудно. Поэтому представляется целесообразным иметь какие-то конструктивные приемы, облегчающие эту работу. Рассмотрим некоторые из них.

1. В работе [6] описан способ точечной характеризации множества Парето, позволяющий оперировать не со всем множеством, а с неким его «полномочным представителем», который в какой-то степени отражает в себе свойства всего множества. Таким представителем может быть, например, центр тяжести J^0 множества $J(P)$, характеризующий его конфигурацию. Он рассматривается как выпуклая комбинация паретовских вершин многогранника $J(P)$ с равными коэффициентами:

$$J^0 = \frac{1}{g}\sum_{i=1}^{g} J^i.$$

Очевидно, что J^0 не будет являться паретовской точкой многогранника $J(P)$. Для определения точки J^*, максимально улучшающей J^0 по обоим критериям одновременно, необходимо воспользоваться программой отсутствия мажорирования [5].

Решение этой задачи и будет являться искомой точечной характеризацией множества P. Оно может также трактоваться как компромиссное решение задачи оценивания параметров уравнения (*).

2. Рассмотрим способ повышения «осязаемости» в восприятии множества P. Это может быть сделано, в частности, посредством построения множества A

$$A = \left\{ \alpha \in R^m \middle| \alpha_i \in \left[\underline{\alpha_i}, \overline{\alpha_i} \right] \right\},$$

которому гарантированно будут принадлежать вектора $\alpha = (\alpha_1, ..., \alpha_m)^T$.

A представляет собой параллелепипед в m-мерном пространстве, в который вписана проекция $\overline{P}$ множества P на R^m.

Легко видеть, что справедливы равенства

$$\underline{\alpha}_i = \underline{\alpha_i^1} - \overline{\alpha_i^2}; \quad \overline{\alpha}_i = \overline{\alpha_i^1} - \underline{\alpha_i^2}.$$

Здесь $\underline{\alpha_i^1}$, $\underline{\alpha_i^2}$ и $\overline{\alpha_i^1}$, $\overline{\alpha_i^2}$ – соответственно минимально и максимально возможные значения положительных и отрицательных частей компонент вектора оцениваемых параметров.

Для построения множества A необходимо для каждого параметра α_i решить $2(g - 1)$ следующих задач ЛП:

$$\underline{\alpha_i^j} = \min_{z \in C^j} \left(\alpha_i^1 - \alpha_i^2 \right), \quad j = \overline{1, g-1};$$

$$\overline{\alpha_i^j} = \max_{z \in C^j} \left(\alpha_i^1 - \alpha_i^2 \right), \quad j = \overline{1, g-1}.$$

Тогда $\underline{\alpha}_i$ и $\overline{\alpha}_i$ отыщутся по формулам

$$\underline{\alpha}_i = \min_{j=\overline{1,g-1}} \underline{\alpha_i^j}; \quad \overline{\alpha}_i = \max_{j=\overline{1,g-1}} \overline{\alpha_i^j}.$$

Имея множество A, легко формировать вектора $\tilde{\alpha}$ для регрессии (*). Но, как правило, A содержит «лишние» параметры, являющиеся компонентами непаретовских векторов. Для того чтобы выявить такие вектора, необходимо всякий раз реализовать программу отсутствия мажорирования.

3. Рассмотрим проблему прогнозирования значений эндогенной переменной y регрессии (*) с множественной оценкой ее параметров.

Пусть заданы значения экзогенных переменных уравнения равны $x_i = \tilde{x}_i$, $i = \overline{1, m}$. Поскольку оценка параметров имеет множественный характер, естественно считать, что соответствующее прогнозное значение переменной y также будет принадлежать множеству – отрезку $\left[\underline{y}, \overline{y} \right]$. Ниже приведен способ расчета его границ.

Решим $2(q - 1)$ задач ЛП:

$$\underline{y}^j = \min_{\tilde{z} \in C^j} \sum_{i=1}^m \tilde{\alpha}_i \tilde{x}_i, \quad j = \overline{1, q-1};$$

$$\overline{y}^j = \max_{\tilde{z} \in C^j} \sum_{i=1}^m \tilde{\alpha}_i \tilde{x}_i, \quad j = \overline{1, q-1}.$$

Тогда получим

$$\underline{y} = \min_{j=\overline{1,q-1}} \underline{y}^j; \quad \overline{y} = \max_{j=\overline{1,q-1}} \overline{y}^j.$$

Моделирование валового регионального продукта Иркутской области

Для построения регрессионной модели валового регионального продукта (ВРП) Иркутской области с сайта Федеральной службы государственной статистики были собраны статистические данные за период с 2005 по 2014 г. по следующим переменным:

y – валовой региональный продукт, млрд руб.;

x_1 – потребление электроэнергии, млрд кВт·ч;

x_2 – численность безработных, тыс. чел.;

x_3 – строительство жилых домов, тыс. кв. м;

x_4 – оборот розничной торговли, млрд руб.

Значения этих переменных представлены в табл. 1, а их динамика – на рис. 1 и 2.

Таблица 1

Статистические данные

Год	y	x_1	x_2	x_3	x_4
2005	258,1	52,5	127,5	303,0	104,3
2006	330,8	53,6	108,3	331,0	128,0
2007	402,7	53,3	104,9	575,0	151,3
2008	438,9	55,1	109,9	585,0	192,1
2009	458,8	52,4	137,9	602,2	191,4
2010	546,1	54,3	127,3	629,5	197,3
2011	634,6	56,7	114,9	755,2	225,8
2012	738,0	58,0	97,8	871,4	250,0
2013	805,2	56,6	104,4	829,2	266,5
2014	907,4	56,3	109,7	716,9	285,9

Корреляционная матрица переменных представлена в табл. 2.

По матрице видно, что между зависимой переменной y и каждым из объясняющих факторов x_1, x_3, x_4 имеется сильная положительная корреляция, а между y и x_2 – умеренная отрицательная, что вполне укладывается в каноны регрессионного анализа.

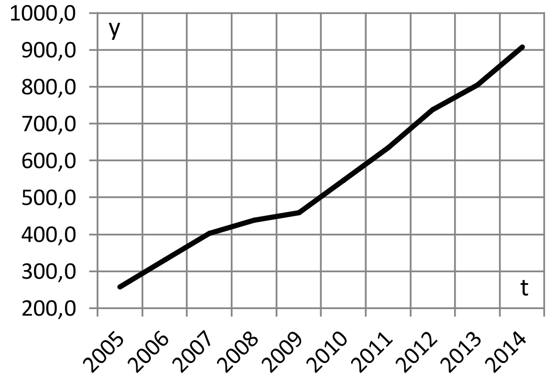

Рис. 1. Динамика ВРП Иркутской области

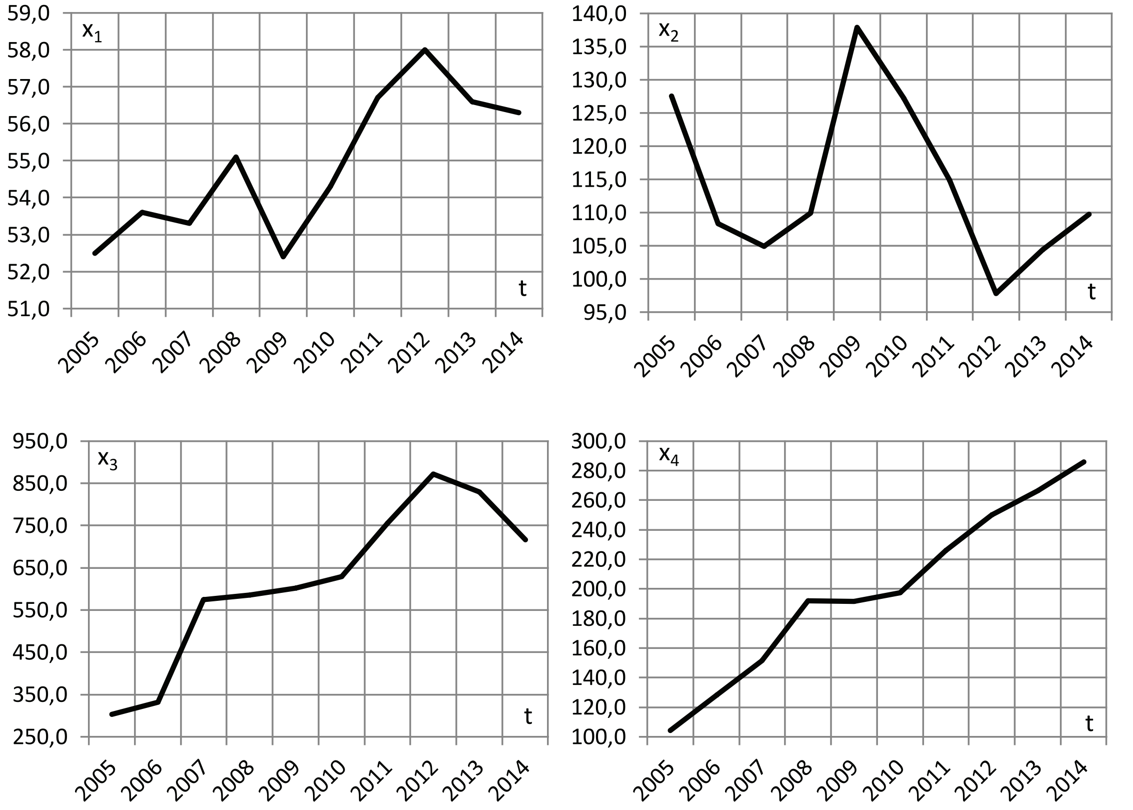

Рис. 2. Динамика влияющих на ВРП показателей

Таблица 2

Корреляционная матрица

	y	x_1	x_2	x_3	x_4
y	1	0,8248	–0,4153	0,8647	0,9780
x_1		1	–0,6748	0,8277	0,8200
x_2			1	–0,4069	–0,3610
x_3				1	0,9088
x_4					1

По корреляционной матрице также видно, что некоторые из объясняющих переменных тесно коррелируют между собой. Следствием этого будет возникновение эффекта мультиколлинеарности при попытке построения, например, линейной модели множественной регрессии. Поэтому для выбора спецификации была реализована технология «конкурса» моделей [1, 7]. В результате была выбрана и оценена по МНК следующая нелинейная по факторам, но линейная по параметрам регрессия:

$$y = \underset{(5,141)}{203,795} + \underset{(10,5)}{4,244 \cdot 10^{-7}}\, x_1 x_4^3 + \underset{(2,123)}{0,881} \frac{x_3}{\ln x_2}.$$

Под коэффициентами этого уравнения записаны значения t-статистик. Для уровня значимости 10 % все коэффициенты значимы. Коэффициент детерминации этой модели $R^2 = 0,985$, критерий Фишера $F = 231$. Статистика Дарбина – Уотсона $DW = 1,96$, что говорит об отсутствии автокорреляции в ошибках модели.

Множественное оценивание модели ВРП Иркутской области

Для реализации процедуры множественного оценивания регрессионных моделей в среде программирования Delphi был разработан специализированный программный комплекс. Все представленные ниже результаты получены с его помощью.

Множество Парето $J(P^*)$, состоящее из шести вершин, представлено в табл. 3.

Таблица 3

Паретовские вершины

J_1	J_∞
245,6	44,23
235,7	44,36
226,2	44,61
218,4	44,9
197,1	46,18
190,6	53,56

Множество Парето в критериальном пространстве представлено на рис. 3.

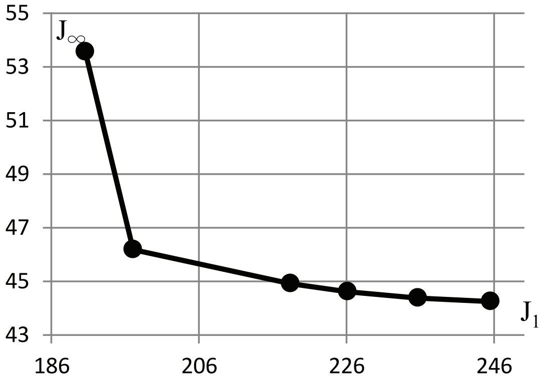

Рис. 3. Множество Парето

1. Точечная характеризация множества P:

$$y = 239,057 + 4,382 \cdot 10^{-7} x_1 x_4^3 + 0,601 \frac{x_3}{\ln x_2}.$$

2. Множество A:

$$\alpha_1 \in [132,1;\ 251,1];$$

$$\alpha_2 \in [3,68 \cdot 10^{-7};\ 4,513 \cdot 10^{-7}];$$

$$\alpha_3 \in [0,4096;\ 1,665].$$

3. Для получения прогноза ВРП Иркутской области на 2015 г. были использованы следующие значения объясняющих переменных: $x_1^* = 52,7$; $x_2^* = 103,1$; $x_3^* = 963,7$; $x_4^* = 290,8$. Интервальный прогноз имеет вид $y \in [921,1;\ 962,6]$.

Выводы

1. Рассмотрена технология множественного оценивания параметров линейных регрессионных моделей.
2. С помощью технологии организации «конкурса» моделей выбрана структурная спецификация модели ВРП Иркутской области.
3. Проведена процедура множественного оценивания параметров выбранной модели регрессии, с помощью которой получен интервальный прогноз ВРП Иркутской области на 2015 г.
4. Из материала работы следует, что аппарат множественного оценивания параметров является весьма эффективным при моделировании сложных систем и более гибким (мягким) по сравнению с традиционных регрессионным анализом, в рамках которого возможно построение только так называемых точечных (жестких) оценок.

Список литературы

1. Базилевский М.П., Носков С.И. Технология организации конкурса регрессионных моделей // Информационные технологии и проблемы математического моделирования сложных систем. – Иркутск, 2009. – Вып. 7. – С. 77–84.

2. Демиденко Е.З. Линейная и нелинейная регрессии. – М.: Финансы и статистика, 1981. – 302 с.

3. Дрейпер Н., Смит Г. Прикладной регрессионный анализ. – М.: Финансы и статистика, 1981. – т. 1. – 366 с., т. 2. – 351 с.

4. Носков С.И. L–множество в многокритериальной задаче оценивания параметров регрессионных уравнений // Информационные технологии и проблемы математического моделирования сложных систем. – Иркутск, 2004. – № 1. – С. 64–71.

5. Носков С.И. Технология моделирования объектов с нестабильным функционированием и неопределенностью в данных. – Иркутск: Облинфорпечать, 1996. – 320 с.

6. Носков С.И. Точечная характеризация множества Парето в линейной многокритериальной задаче // Современные технологии. Системный анализ. Моделирование. – Иркутск, 2008. – № 17. – С. 99–102.

7. Носков С.И., Базилевский М.П. Программный комплекс автоматизации процесса построения регрессионных моделей // Международный журнал прикладных и фундаментальных исследований. – М., 2010. – № 1. – С. 93–94.

8. Себер Дж. Линейный регрессионный анализ. – М.: Мир, 1980. – 456 с.

9. Yu L., Zeleny M. The set of all nondoinated solutions in linear cases and multicriteria simplex method // J. of Math. Anal. and Applic. – 1975. – Vol. 49. – № 2. – P. 430–468.

References

1. Bazilevskij M.P., Noskov S.I. Tehnologija organizacii konkursa regressionnyh modelej. Informacionnye tehnologii i problemy matematicheskogo modelirovanija slozhnyh system, Irkutsk, 2009, no. 7, pp. 77–84.

2. Demidenko E.Z. Linejnaja i nelinejnaja regressii. M.: Finansy i statistika, 1981, 302 p.

3. Drejper N., Smit G. Prikladnoj regressionnyj analiz. M.: Finansy i statistika, 1981, t.1, 366 p., t. 2, 351 p.

4. Noskov S.I. L–mnozhestvo v mnogokriterialnoj zadache ocenivanija parametrov regressionnyh uravnenij. Informacionnye tehnologii i problemy matematicheskogo modelirovanija slozhnyh system, Irkutsk, 2004, no. 1, pp. 64–71.

5. Noskov S.I. Tehnologija modelirovanija obektov s nestabilnym funkcionirovaniem i neopredelennostju v dannyh, Irkutsk: Oblinforpechat, 1996, 320 p.

6. Noskov S.I. Tochechnaja harakterizacija mnozhestva Pareto v linejnoj mnogokriterialnoj zadache. Sovremennye tehnologii. Sistemnyj analiz. Modelirovanie, Irkutsk, 2008, no. 17, pp. 99–102.

7. Noskov S.I., Bazilevskij M.P. Programmnyj kompleks avtomatizacii processa postroenija regressionnyh modelej. Mezhdunarodnyj zhurnal prikladnyh i fundamentalnyh issledovanij, Moskva, 2010, no. 1, pp. 93–94.

8. Seber Dzh. Linejnyj regressionnyj analiz, M.: Mir, 1980, 456 p.

9. Yu L., Zeleny M. The set of all nondoinated solutions in linear cases and multicriteria simplex method. J. of Math. Anal. and Applic., 1975, no. 2, pp. 430–468.