ISSN 1814-1196
Научный вестник НГТУ
том 55, № 2, 2014, с. 67–76

http://journals.nstu.ru/vestnik
Scientific Bulletin of NSTU
Vol. 55, No. 2, 2014, pp. 67–76

УДК 519.237.5

Оценивание параметров регрессионных моделей в условиях гетероскедастичности неизвестной формы*

А.В. ФАДДЕЕНКОВ

630073, РФ, г. Новосибирск, пр. Карла Маркса, 20, Новосибирский государственный технический университет, к. т. н., доцент, e-mail: fadd@fb.nstu.ru

В статье рассмотрена задача оценивания параметров линейных регрессионных моделей в условиях гетероскедастичности. При построении модели предполагается, что дисперсия случайной ошибки зависит от некоторого фактора (в роли такого фактора может выступать один из входных факторов модели или какой-то внешний фактор). Предполагается, что совокупность исходных данных может быть разбита на участки с одинаковой дисперсией. Предложен подход, позволяющий свести регрессионную модель с такими предположениями к модели со структурированной ошибкой. При построении модели предполагается, что ковариационная матрица ошибок является линейной комбинацией известных матриц с неизвестными коэффициентами. Данные известные матрицы имеют диагональный вид и формируются исходя из структуры исходных данных. Оценивание параметров этой модели предлагается проводить с использованием квадратичных оценок минимальной нормы (MINQE). Разработан алгоритм определения границ интервалов области исходных данных с одинаковой дисперсией. В алгоритме проверка гипотез о различии дисперсий на разных участках проводится с использованием критерия дисперсионного анализа (ANOVA-критерия). Средствами статистического имитационного моделирования проведены вычислительные эксперименты, показывающие работоспособность разработанного алгоритма. Даны рекомендации по выбору наилучших параметров алгоритма.

Ключевые слова: регрессионный анализ, линейная регрессионная модель, метод наименьших квадратов, гомоскедастичность, гетероскедастичность, модели компонент дисперсии, квадратичные оценки минимальной нормы, статистическое моделирование, вычислительный эксперимент

ВВЕДЕНИЕ

Регрессионный анализ является одним из самых распространенных на практике видов статистического анализа. Однако его корректное применение требует соблюдения ряда условий. Одним из предположений классической регрессионной модели является предположение постоянства дисперсии случайной ошибки (гомоскедастичности). Именно в этом случае оценки неизвестных параметров, полученные методом наименьших квадратов, являются оптимальными с точки зрения теоремы Гаусса – Маркова [1, 3, 8]. На практике часто возникают ситуации, когда данное предположение оказывается нарушено. Причиной нарушения могут служить различные факторы как внешнего, так внутреннего характера. Если исследователю известны причины появления неоднородностей в исходных данных, а также последствия (дисперсии наблюдений при каждом изменении), то для оценивания параметров регрессионных уравнений в таких условиях лучше использовать обобщенный метод наименьших квадратов. На практике исследователь такой информацией располагает крайне редко. Следовательно, необходимы специальные алгоритмы и критерии, позволяющие определять наличие гетероскедастичности, а также оценивать ее форму. Исследованию критериев обнаружения гетеро-

* *Статья получена 10 января 2014 г.*
Работа выполнена при финансовой поддержке РФФИ в рамках научного проекта №13-07-00299а

скедастичности была посвящена работа [5]. В данной же статье предлагается алгоритм, позволяющий при ряде допущений определять форму гетероскедастичности исходных данных.

1. ПОСТАНОВКА ЗАДАЧИ

Рассмотрим линейную модель следующего вида:

$$Y = X\beta + \varepsilon, \tag{1}$$

где $Y = (y_1, \ldots, y_N)^T$ – вектор, состоящий из N наблюдений; β – вектор неизвестных параметров; X – матрица значений независимых переменных, соответствующих неизвестным параметрам; $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_N)^T$ – случайный вектор.

Корректное оценивание вектора неизвестных параметров уравнения (1) методом наименьших квадратов возможно, если относительно случайных ошибок справедливы следующие предположения [2, 6]:

$$E(\varepsilon_i) = 0,\ D(\varepsilon_i) = \sigma^2,\ E(\varepsilon_i \varepsilon_j) = 0,\ \forall i \neq j, \tag{2}$$

или в матричной форме:

$$E(\varepsilon) = 0,\ D(\varepsilon) = \sigma^2 I_N.$$

В случае гетероскедастичной ошибки наблюдения нарушается предположение о неизменности дисперсии ошибки.

Допустим, что дисперсия ошибок не постоянна и можно разбить всю область исходных данных на r подгрупп таким образом, чтобы выполнялось условие

$$E[\varepsilon_i] = 0,\ D[\varepsilon_i] = \sigma_k^2 \text{ при } i \in S_k,\ i = 1, \ldots, N, \tag{3}$$

где S_k – k-я подгруппа наблюдений, $k = 1, \ldots, r$; $S_1 \cup S_2 \cup \cdots \cup S_r = \{1, \ldots, N\}$; $S_m \cap S_n = \varnothing$, при $m \neq n$.

Если дисперсии σ_k^2, $k = 1, \ldots, r$ известны, то оценивание параметров модели (1), (3) возможно по обобщенному методу наименьших квадратов. При неизвестных дисперсиях необходимо проводить их оценивание.

Решение этой проблемы возможно в рамках модели со структурированной ошибкой путем введения следующих обозначений:

$$F_k(i, j) = \begin{cases} 0, & i \neq j, \\ 0, & i = j, i \notin S_k, \\ 1, & i = j, i \in S_k, \end{cases} \quad \theta_k = \sigma_k^2,\ i = 1, \ldots, N,\ \ j = 1, \ldots, N,$$

где F_k – матрица, составленная из элементов $F_k(i, j)$, $k = 1, \ldots, r$. Тогда

$$E(\varepsilon) = 0,\ D(\varepsilon) = F_\theta = \theta_1 F_1 + \ldots + \theta_r F_r. \tag{4}$$

Модель (1), (4) является частным случаем модели со структурированной ошибкой [7].

2. ОЦЕНИВАНИЕ КОВАРИАЦИОННОЙ МАТРИЦЫ ОШИБОК

Для оценивания вектора неизвестных параметров $\theta = (\theta_1, \ldots, \theta_r)^T$ воспользуемся квадратичными оценками минимальной нормы (MINQ-оценками) [7]. Эти оценки являются несмещенными, не требуют знания законов распределения случайных составляющих модели, явля-

ются локально оптимальными в смысле минимума дисперсии. MINQ-оценка для параметра θ_j может быть вычислена следующим образом:

$$\hat{\theta}_j = Y^T A_j Y , \tag{5}$$

где $A_j = \sum_{i=1}^{r} H_{ji}^{-1} R\ F_i R^T$, H – матрица, составленная из элементов $tr\left(R\ F_i R^T F_j\right)$ ($i = 1,\cdots,r$, $j = 1,\cdots,r$); $R = T^{-} M$; $T = F_\alpha + XX^T$, $M = I - P$, $P = X\left(X^T T^{-} X\right)^{-} X^T T^{-}$, $F_\alpha = \alpha_1 F_1 + \ldots + \alpha_r F_r$, $\alpha = (\alpha_1, ..., \alpha_r)^T$ – вектор априорных значений неизвестных параметров $\theta = (\theta_1, ..., \theta_r)^T$.

В результате матрица F_θ может быть оценена следующим образом:

$$\hat{F}_\theta = \hat{\theta}_1 F_1 + \ldots + \hat{\theta}_r F_r , \tag{6}$$

а вычисление оценок фиксированных параметров проводится с использованием обобщенного метода наименьших квадратов:

$$\hat{\beta} = (X^T \hat{F}_\theta^{-1} X)^{-} X^T \hat{F}_\theta^{-1} Y.$$

Естественно предположить, что наилучшие результаты достигаются, если априорные значения $\alpha_1, ..., \alpha_r$ максимально близки к истинным значениям параметров $\theta_1, ..., \theta_r$. Снижение зависимости результатов от исходных априорных значений возможно при переходе к итерационной процедуре, в которой оценки, найденные на предыдущем шаге, используются на следующем как априорные значения. Оценки, полученные таким образом, получили названия итерационных MINQ-оценок (IMINQ-оценки) [7].

3. АЛГОРИТМ ОПРЕДЕЛЕНИЯ ГРАНИЦ ПОДМНОЖЕСТВ С ОДИНАКОВОЙ ДИСПЕРСИЕЙ

Построение матрицы (6) не вызывает особых трудностей, если известны все подмножества S_i, для которых выполняется условие (3). Если эти подмножества неизвестны, то их необходимо каким-либо образом выделить. Дальнейшие рассуждения будут основываться на предположении, что гетероскедастичность определяется зависимостью дисперсии ошибки наблюдений от некоторого фактора. Таким фактором может послужить один из регрессоров модели (1) или некоторый выделенный внешний фактор, определенный из особенностей моделируемого объекта.

В дальнейшем будем считать, что фактор, влияющий на величину дисперсии ошибки, определен и все наблюдения отсортированы в соответствии с этим фактором. Задача поиска подмножеств S_i сводится к разбиению имеющегося диапазона исходных данных на некоторое количество отрезков. При этом будем фиксировать не границы подмножеств S_i, а количество точек, попавших в эти подмножества, с учетом проведенной сортировки.

Обозначим через n_k количество наблюдений в *k*-й подгруппе наблюдений (или в *k*-м интервале), $k = 1, \ldots, r$. Естественно, должно выполняться условие

$$\sum_{k=1}^{r} n_k = N.$$

Рассмотрим предлагаемый автором алгоритм определения границ интервалов, содержащих данные с одинаковой дисперсией.

Шаг 1. Определяем максимально возможное количество интервалов $r_{\max}$, минимальное количество элементов в одном интервале $n_{\min}$, единицу приращения размера интервала Δ. Задаем начальное разбиение исходных данных $\{n_k,\ k=1,\ldots,r_{\max}\}$ как равномерное

$$n_k=\frac{N}{r_{\max}},\ k=1,\ldots,(r_{\max}-1),\ \ n_{r_{\max}}=N-\sum_{k=1}^{(r_{\max}-1)}n_k\ .$$

Текущее количество интервалов определяем как $r=r_{\max}$. Для текущего разбиения $\{n_k,\ k=1,\ldots,r\}$ вычисляем значение критерия оптимальности F. В качестве такого критерия может послужить статистика любого критерия на гетероскедастичность, при построении которого используется разбиение исходной области данных на интервалы [5].

Например, в качестве критерия оптимальности может быть использована статистика критерия Бартлетта:

$$F=\frac{N}{1+L}\ln B_s\ , \tag{7}$$

где

$$B_s=\frac{\frac{1}{N}\sum_{k=1}^{r}n_k s_k^2}{\left(\prod_{k=1}^{r}s_k^{2n_k}\right)^{1/N}},\ \ L=\frac{1}{3(r-1)}\left(-\frac{1}{N}+\sum_{k=1}^{r}\frac{1}{n_k}\right),$$

s_k^2 – оценка дисперсии остатков уравнения (1), попавших в *k*-й интервал.

Кроме этого, можно воспользоваться ANOVA-критерием, предложенным в [5]. Для этого строится вспомогательная однофакторная модель дисперсионного анализа

$$y_{ij}=\mu+\alpha_i+u_{ij}\ , \tag{8}$$

где в качестве значений зависимой переменной y_{ij} выступают модули (или квадраты) остатков исходного уравнения (1), α_i – главные эффекты, число которых определяется количеством интервалов, μ – генеральное среднее, u_{ij} – случайная компонента, i – номер интервала, j – номер наблюдения в интервале.

В качестве функции F может использоваться объясненная сумма квадратов или коэффициент детерминации модели (8). Также в роли функции F может выступать статистика Фишера, построенная при проверке на значимость качественного фактора модели (8).

Пусть номер текущего интервала $i=1$.

Шаг 2. Для текущего интервала i проводим сдвиг правой границы вправо на величину Δ: $n_i=n_i+\Delta$, $n_{i+1}=n_{i+1}-\Delta$. Для нового разбиения вычисляем значение критерия оптимальности F_{new}. Если $F_{new}>F$, то сохраняем новое разбиение в качестве текущего и переходим на Шаг 3. В противном случае для интервала i проводим сдвиг правой границы влево на величину Δ: $n_i=n_i-\Delta$, $n_{i+1}=n_{i+1}+\Delta$. Для нового разбиения вычисляем значение критерия оптимальности F_{new}. Если $F_{new}>F$, то сохраняем новое разбиение в качестве текущего и переходим на Шаг 3. Если сдвиг правой границы интервала i не привел к улучшению критерия оптимальности, то переходим к следующему интервалу: $i=i+1$ и повторяем Шаг 2.

Если $i=r$, то оптимальное разбиение достигнуто, останавливаем алгоритм.

Шаг 3. Проводим проверку на допустимые минимальные размеры интервалов. Если для какого-либо интервала выполняется условие, что

$$n_k < n_{\min}, \; k = 1, \ldots, r,$$

то проводим объединение этого интервала с соседним (для определенности будем считать, что объединение текущего интервала проводится с предыдущим, а первого – со вторым).

Далее определяем $r = r - 1$, $i = 1$ и переходим на шаг 1. Если интервалов с количеством наблюдений меньшим $n_{\min}$ не найдено, то определяем $i = i + 1$ и переходим на шаг 1. Работоспособность описанного алгоритма напрямую зависит от определения его начальных параметров: максимального количества интервалов $r_{\max}$, минимального количества элементов в одном интервале $n_{\min}$ и единицы приращения размера интервала Δ. Естественно, что эти параметры зависят от общего количества наблюдений. Число интервалов можно определить, например, по правилу Штюргеса [4]:

$$r_{\max} \approx \left[1 + 3.32 \lg N\right].$$

Размер единицы приращения интервала Δ можно порекомендовать выбирать в объеме 1 % от исходного размера выборки, а минимальное количество элементов в одном интервале $n_{\min}$ – 5–10 %, но не менее 10 наблюдений.

4. ПРИМЕРЫ РАБОТЫ АЛГОРИТМА

Рассмотрим в качестве примера модель парной регрессии

$$y = 1 + x + \varepsilon, \tag{9}$$

в которой случайная ошибка ε подчиняется нормальному закону распределения $N(0, \sigma_\varepsilon^2)$ с переменным параметром σ_ε^2. Значения независимой переменной распределялись равномерно на отрезке, т. е. $x \in [1; 5]$.

В ходе проведенных вычислительных экспериментов рассматривалась гетероскедастичность, при которой дисперсия случайной ошибки принимала два значения: $\sigma_{\min}^2$ и $\sigma_{\max}^2$. Величина $\sigma_{\min}^2$ во всех экспериментах принималась равной 0.5, а $\sigma_{\max}^2 = 3$.

В первой серии экспериментов были определены четыре участка одинакового размера, на каждом из которых фиксировалась своя дисперсия ошибки. График остатков модели, полученных при такой форме гетероскедастичности, представлен на рис. 1.

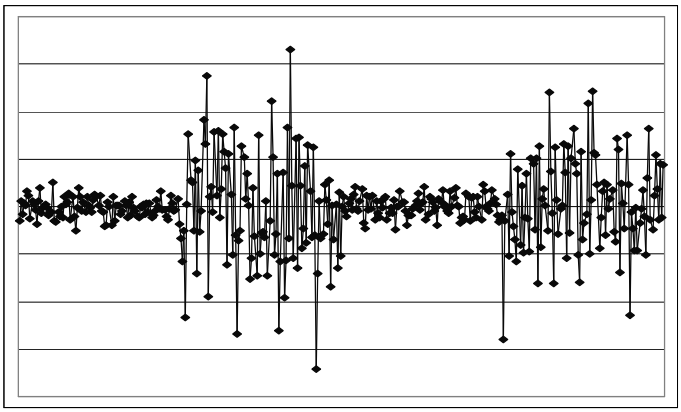

Рис. 1. Остатки модели (9) в первой серии экспериментов

На рис. 2–4 приведены результаты работы алгоритма при использовании в качестве критерия оптимальности ANOVA-критерия. Как видно из рисунков, во всех случаях были получены интервалы, близкие к истинным.

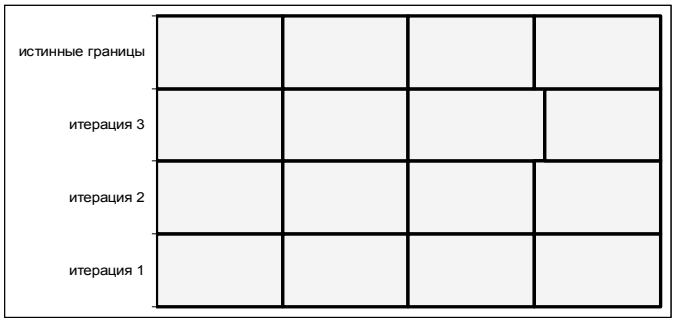

Рис. 2. Последовательность изменения интервалов в первой серии экспериментов при $r_{\max} = 4$

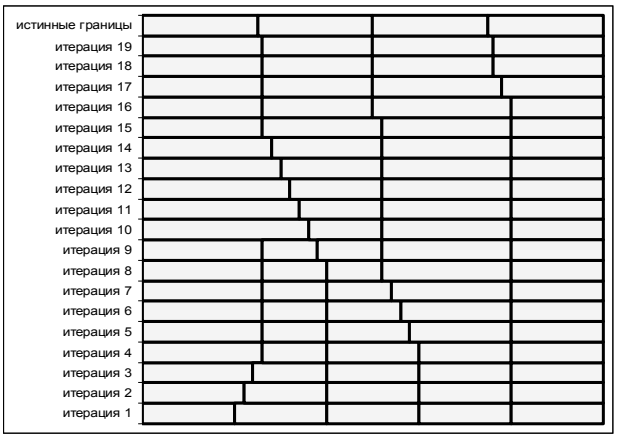

Рис. 3. Последовательность изменения интервалов в первой серии экспериментов при $r_{\max} = 5$

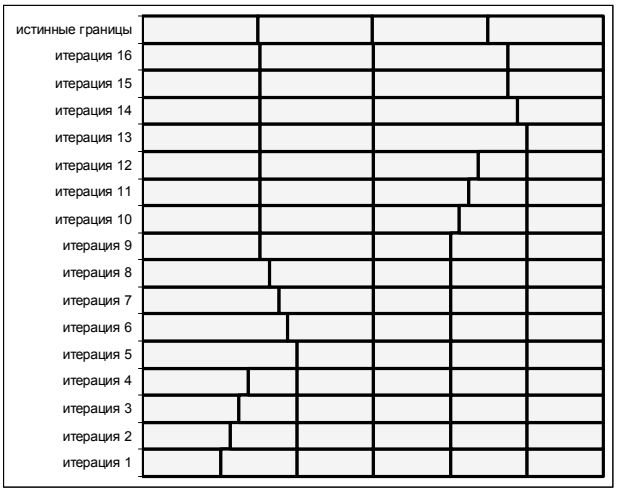

Рис. 4. Последовательность изменения интервалов в первой серии экспериментов при $r_{\max} = 6$

Во второй серии экспериментов были определены три участка разного размера, на каждом из которых фиксировалась своя дисперсия ошибки. График остатков модели, полученных при такой форме гетероскедастичности, представлен на рис. 5.

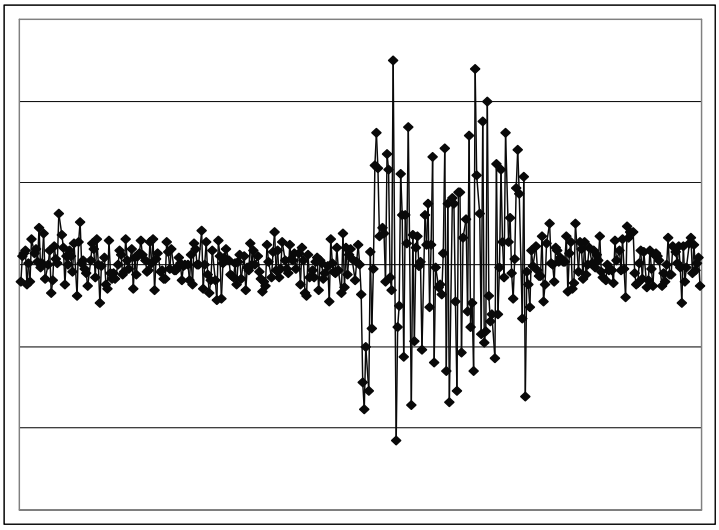

Рис. 5. Остатки модели (9) во второй серии экспериментов

На рис. 6–8 приведены результаты работы алгоритма при использовании, как и ранее, в качестве критерия оптимальности ANOVA-критерия. Как видно из рисунков, во всех случаях также были получены интервалы, близкие к истинным.

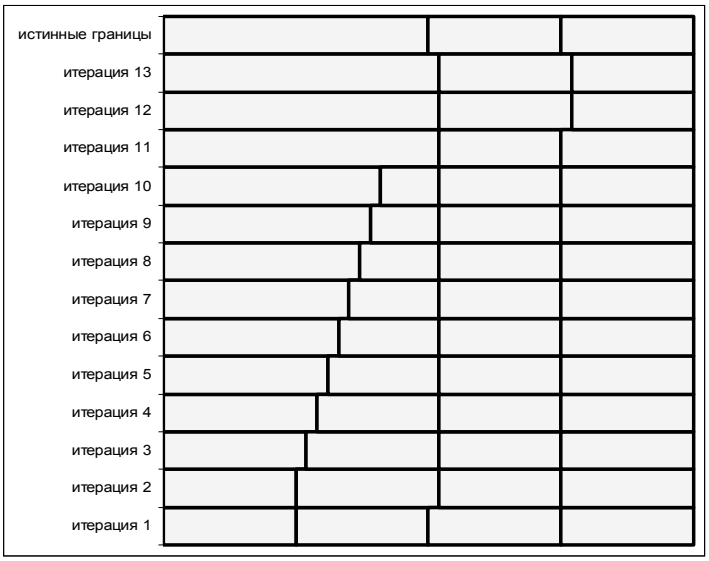

Рис. 6. Последовательность изменения интервалов в первой серии экспериментов при $r_{\max} = 4$

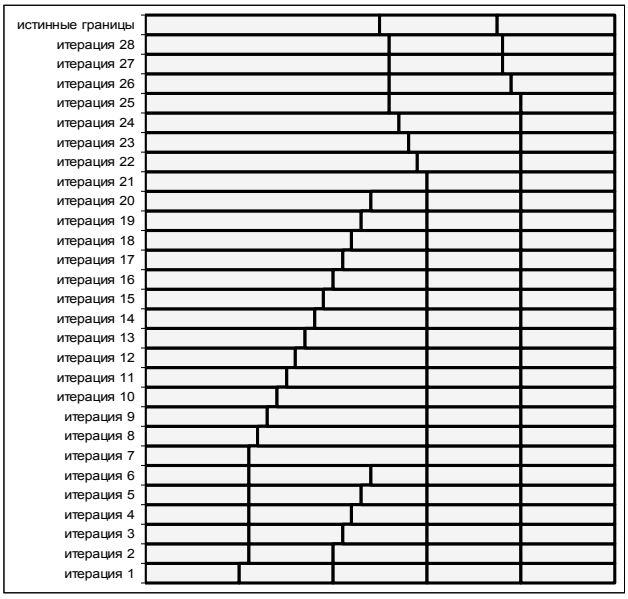

Рис. 7. Последовательность изменения интервалов в первой серии экспериментов при $r_{\max} = 5$

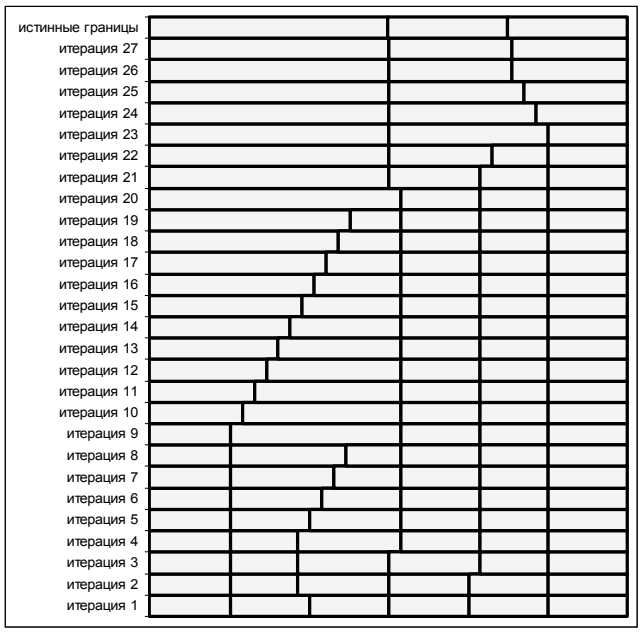

Рис. 8. Последовательность изменения интервалов в первой серии экспериментов при $r_{\max} = 6$

Отметим, что все примеры работы алгоритма определения границ интервалов приведены для параметров $\Delta = 0.01N$, $n_{\min} = 0.1N$. Такие значения параметров алгоритма дают наилучшие результаты в большинстве случаев. Границы интервалов, определенные этим алгоритмом, используются далее для определения совокупности подмножеств S_k, $k = 1, \cdots, r$ и далее для определения вида модели (1), (4).

ЗАКЛЮЧЕНИЕ

Предложенный автором алгоритм определения границ интервалов области исходных данных с одинаковой дисперсией позволяет решить проблему идентификации линейных регрессионных моделей в условиях гетероскедастичности. Однако следует помнить, что данный алгоритм применим только в случае, если дисперсия случайной составляющей модели зависит от некоторого фактора (в роли такого фактора может выступать один из входных факторов модели или какой-то внешний фактор), значения которого поддаются определению для каждого наблюдения.

СПИСОК ЛИТЕРАТУРЫ

[1] **Айвазян С.А., Мхитарян В.С.** Прикладная статистика и основы эконометрики. Т. 2. – М.: Юнити, 2001. – 432 с.

[2] **Доугерти К.** Введение в эконометрику. – М.: МГУ, 1999. – 402 с.

[3] **Дрейпер Н.Р., Смит Н.** Прикладной регрессионный анализ. – М.: Статистика, 1973. – 392 с.

[4] **Львовский Е.Н.** Статистические методы построения эмпирических формул: учеб. пособие для втузов. – М.: Высш. шк., 1988. – 239 с.

[5] **Тимофеев В.С., Фаддеенков А.В.** Исследование критериев обнаружения гетероскедастичности в регрессионных моделях // Науч. вестн. НГТУ. – Новосибирск: Изд-во НГТУ, 2007. – № 4 (29). – С. 3–14.

[6] **Green W.H.** Econometric analysis. – 6th ed. – New Jersey, Prentice Hall, 2007. – 1216 p.

[7] **Rao C.R., Kleffe J.** Estimation of variance components and applications. – Amsterdam: Elsevier Science, 1988. – 374 p.

[8] **Searle S.R.** Linear models. – New York, John Wiley & Sons, Inc., 1971. – 532 p.

Фаддеенков Андрей Владимирович, кандидат технических наук, доцент кафедры теории рынка. Основное направление научных исследований – разработка и исследование методов и алгоритмов анализа многофакторных объектов со структурированной ошибкой. Имеет более 30 публикаций, в том числе один учебник. E-mail: fadd@fb.nstu.ru

Estimation of regression model parameter under heteroskedasticity of an unknown form[*]

A.V. FADDEENKOV

Novosibirsk State Technical University, 20, K. Marx Prospect, Novosibirsk, 630073, Russian Federation, PhD (Eng.), associate professor, e-mail: fadd@fb.nstu.ru

The paper considers the problem of estimating linear regression model parameters under heteroscedasticity conditions. Constructing the model assumes that the variance of a random error depends on several factors (e.g. one of the model input factors or some external factor). It is assumed that all the source data can be split into sections of equal dispersion. The proposed approach permits us to reduce the regression model with these assumptions to a structured error model. Constructing the model assumes that the covariance matrix of an error is a linear combination of known matrices with unknown coefficients. These known matrices are diagonal matrices and are formed based on the structure of the input data. The estimation of the model parameters is proposed using the minimum norm quadratic estimates. An algorithm for determining the boundaries of the area of the source data intervals with equal variance is proposed. In this algorithm the test of hypotheses of variance differences at different areas is carried out by using the variance analysis criterion (ANOVA-criterion). Computational experiments showing the efficiency of the algorithm have been conducted by statistical simulation methods. Advice on selecting the best algorithm parameters is given.

Keywords: regression analysis, linear regression model, least squares method, homoscedasticity, heteroscedasticity, variance component model, minimum norm quadratic estimation, statistical modeling, computational experiment

[*] *Manuscript received January 10, 2014.*

Work is executed at financial support of the Russian Foundation for basic research within the framework of a research project №13-07-00299a

REFERENCES

[1] **Ayvazyan S.A., Mhitaryan V.S.** *Prikladnaya statistika i osnovyi ekonometriki* [Applied statistics and econometrics]. Moscow, Yuniti Publ., 2001. 432 p.

[2] **Dougherti K.** *Introduction to econometrics*. New York, Oxford University Press, 1992. 399 p. (Russ. ed.: Dougerti K. *Vvedenie v ekonometriku*. Moscow, MGU Publ., 1999. 402 p.).

[3] **Draper N., Smit N.** *Applied regression analysis*. New York, Wiley, 1968. 407 p. (Russ. ed.: Dreiper N., Smit N. *Prikladnoy regressionnyiy analiz*. Moscow, Statistika Publ., 1973. 392 p.).

[4] **Lvovskiy E.N.** *Statisticheskie metodyi postroeniya empiricheskih formul* [Statistical methods for constructing empirical formula]. Moscow, High school Publ., 1988. 239 p.

[5] **Timofeev V.S., Faddeenkov A.V.** Issledovanie kriteriev obnaruzheniya geteroskedastichnosti v regressionnyih modelyah [The research of tests for heteroscedasticity in regression models]. *Nauchnyi vestnik NGTU* [Science Bulletin of Novosibirsk State Technical University], 2007, no. 29, vol. 4, pp. 3-14.

[6] **Greene W.H.** Econometric analysis. 6th edition. New Jersey, Prentice Hall, 2007. 1216 p.

[7] **Rao C.R., Kleffe J.** Estimation of variance components and applications. Amsterdam, the Netherlands, Elsevier Science, 1988. 374 p.

[8] **Searle S.R.** Linear models. New York, John Wiley & Sons, Inc., 1971, 532 p.