DOI: 10.24143/2072-9502-2017-2-14-26
УДК 519.233

В. И. Денисов, А. Ю. Тимофеева

ОЦЕНИВАНИЕ ДИСПЕРСИИ ОШИБКИ ВХОДНОГО ФАКТОРА В ПОЛИНОМИАЛЬНОЙ ФУНКЦИОНАЛЬНОЙ МОДЕЛИ ПРИ НАЛИЧИИ ГОМОСКЕДАСТИЧНОСТИ[1]

Функциональные модели с ошибками в переменных не укладываются в стандартную регрессионную постановку, поскольку входные факторы в модели представлены неизвестными детерминированными величинами, на практике наблюдаемыми со случайными погрешностями. Обычно оценивание таких моделей производится с использованием дополнительной информации: о дисперсии ошибки входного фактора (метод скорректированных наименьших квадратов, разработанный специально для оценивания полиномиальных зависимостей) или о соотношении дисперсий ошибок факторов (метод общих наименьших квадратов). Их значения, как правило, задаются из априорных предположений. В работе предпринимается попытка ослабить модельные предположения, а именно исключить необходимость задавать дисперсию ошибки входного фактора благодаря возможности ее оценивания по тем же данным, по которым восстанавливается нелинейная модель, т. е. без привлечения дополнительной информации. Такая возможность появляется в случае, если ошибки измерения однородны. Тогда, если оценки ненаблюдаемых значений входного фактора близки к истинным, должна обнаруживаться гомоскедастичность ошибок, которая нарушается, как только во входном факторе нелинейной модели появляются погрешности. Это аналитически показано для полиномиальных моделей. Тем самым в рамках предлагаемого алгоритма подбирается такая оценка дисперсии ошибки входного фактора, которая минимизирует статистику критерия обнаружения гетероскедастичности. В ходе вычислительных экспериментов сравнивалась работа алгоритма при использовании различных критериев проверки гипотезы об однородности дисперсии ошибок. Кроме того, сопоставлялась точность восстановления отклика с учетом найденных оценок и с помощью обычного метода наименьших квадратов. Установлено, что разработанный алгоритм обеспечивает значительное превосходство по остаточной сумме квадратов, т. е. может быть рекомендован к применению на практике.

Ключевые слова: модель с ошибками в переменных, функциональный случай, дисперсия ошибки, входной фактор, гетероскедастичность, критерий Спирмена, критерий Бартлетта, критерий ANOVA, метод скорректированных наименьших квадратов, метод общих наименьших квадратов.

Введение

Исследователи во многих областях науки сталкиваются с задачей восстановления уравнения $Y = f(X;\theta)$ зависимости отклика Y от объясняющей переменной X по наблюдаемым данным. Функция $f(X;\theta)$, в общем случае нелинейная, предполагается заданной с точностью до вектора неизвестных параметров θ. Вычисление значений θ не составило бы труда, если бы наблюдаемые значения y, x точно воспроизводили истинные значения Y, X. Но на практике это далеко не так. В любом эксперименте возникают погрешности, обусловленные как неточностью измерения, так и невозможностью полностью исключить влияние внешних факторов.

В стандартной регрессионной постановке предполагается, что зашумлены только значения отклика, т. е. в i-м опыте вместо Y_i наблюдается

$$y_i = Y_i + \varepsilon_i = f(X_i;\ \theta) + \varepsilon_i,\ i = 1, ..., N, \qquad (1)$$

где ε_i – случайная погрешность изменения переменной Y в i-м эксперименте; N – число экспериментов (объем выборки).

Значения входного фактора обычно предполагаются заданными точно, без ошибки, либо с такими малыми погрешностями, которыми можно пренебречь. Однако не во всех приложени-

[1] Работа выполнена при поддержке Министерства образования и науки Российской Федерации, проект № 2.2327.2017/ПЧ.

ях это имеет место. Например, при проведении исследований в области химии высокомолекулярных соединений не удается в точности определить концентрацию мономеров, участвующих в реакции сополимеризации. Тем самым восстановление уравнения сополимеризации и оценивание относительных активностей мономеров с помощью стандартных подходов регрессионного анализа (метода наименьших квадратов (МНК)) становится некорректным. Это подчеркивается многими специалистами в этой области, и они предлагают включать в модель ошибки измерения входных факторов [1–3].

В предположении, что в i-м эксперименте значение объясняющей переменной x_i отражает истинное значение X_i с некоторой погрешностью

$$x_i = X_i + \delta_i, i = 1, ..., N, \tag{2}$$

модель (1)–(2) называется моделью с классической ошибкой.

При этом разделяют два случая. В функциональном случае [4] ненаблюдаемая истинная переменная X является детерминированной. Этому случаю посвящены исследования, начатые еще при зарождении этой тематики. Их обобщение можно найти в работе М. Кендалла, А. Стьюарта [5]. Там же описаны основные проблемы анализа такого рода моделей. Практический пример с постановкой, описанной неявной квадратичной функциональной моделью, можно найти в [6]. Здесь рассматривается проблема подгонки эллипсоида для произвольного множества точек, имеющая фундаментальное значение во многих областях прикладной науки: в астрономии, геодезии, цифровой обработке изображений, в робототехнике, в метрологии и др. [7].

В структурных моделях X предполагается случайной величиной с некоторым законом распределения с неизвестными параметрами [4]. Такая постановка более характерна для ситуаций, когда происходит наблюдение за естественным состоянием некоторого случайного процесса, и не вполне соответствует условиям активного эксперимента.

Часто, однако, по такому принципу разделяют не столько постановки, сколько методы оценивания [8], называя структурными методы, основанные на предположении о некотором распределении X, и функциональными – лишенные такого предположения. Тем самым функциональные методы являются более гибкими, их можно применять и в случае, если X представляет собой случайную величину, при этом они основаны на более слабых предположениях и не требуют спецификации распределения X и оценки параметров этого распределения.

Тем самым мы будем рассматривать именно функциональный случай моделей с классической ошибкой. Введем дополнительные предположения относительно ошибок в модели (1)–(2):

$$\mathrm{E}(\varepsilon_i) = \mathrm{E}(\delta_i) = 0,\ \mathrm{D}(\varepsilon_i) = \sigma_\varepsilon^2,\ \mathrm{D}(\delta_i) = \sigma_\delta^2,\ \forall i,$$

$$\mathrm{cov}(\varepsilon_i, \varepsilon_j) = \mathrm{cov}(\delta_i, \delta_j) = 0,\ \forall i \neq j,\ \mathrm{cov}(\varepsilon_i, \delta_j) = 0,\ \forall i, j. \tag{3}$$

И еще одно дополнительное предположение о нормальности распределения ошибок вводится для упрощения процедуры оценивания параметров и статистических выводов. Задача состоит в оценивании вектора неизвестных параметров θ в модели (1)–(3). Использование обычного МНК в данной постановке приводит к смещенным и несостоятельным оценкам [5], поэтому для оценивания параметров функциональных моделей предложен ряд специальных подходов [5, 6, 8]. Алгоритмы оценивания реализованы преимущественно для анализа полиномиальных зависимостей [9, 10], поэтому далее всюду будет предполагаться, что $f(X; \theta)$ выражено полиномом степени k.

1. Обзор методов оценивания функциональных моделей

Если задано соотношение дисперсий ошибок $\gamma = \sigma_\delta^2 / \sigma_\varepsilon^2$ или известна его оценка, при наличии априорной информации о виде распределения случайных ошибок оценки параметров могут быть найдены методом максимального правдоподобия [5]. В предположении о совместном нормальном распределении ошибок ε_i, δ_i задача сводится к минимизации выражения

$$G = \sum_{i=1}^{N} \frac{1}{\gamma}\left(x_i - X_i\right)^2 + \left(y_i - f(X_i, \theta)\right)^2 \quad (4)$$

по ненаблюдаемым значениям переменной X и вектору неизвестных параметров θ. Для численной оптимизации удобно воспользоваться итерационным алгоритмом, описанным в [11]. В работе [10] предложена реализация итерационного алгоритма для оценивания полиномов заданной степени *k*, состоящая в следующем.

Пусть истинная функция, описывающая зависимость между переменными, представлена полиномом *k*-й степени:

$$Y = f\left(X;\ \theta\right) = \theta_0 + \theta_1 X + \theta_2 X^2 + \ldots + \theta_k X^k . \quad (5)$$

Это упрощает поиск значений $\hat{X}_i$, обеспечивающих минимум функции G при заданном векторе θ, поскольку в этом случае частная производная функции G по X_i будет многочленом степени $p = 2k - 1$:

$$\frac{1}{2}\frac{\partial G}{\partial X_i} = -\frac{1}{\gamma}\left(x_i - X_i\right) + \left(\theta_0 - y_i + \theta_1 X_i + \ldots + \theta_k X_i^k\right)\left(\theta_1 + 2\theta_2 X_i + \ldots + k\theta_k X_i^{k-1}\right) . \quad (6)$$

Как известно, любой вещественный многочлен нечетной степени всегда имеет хотя бы один действительный корень, это гарантирует существование действительных значений $\hat{X}_i$, оптимизирующих функцию G при заданном векторе θ. Если вещественных корней несколько, то выбирается тот, который обеспечивает минимальное значение функции G. Для поиска всех корней полинома использовался алгоритм Дженкинса – Трауба.

Описанный подход к оцениванию полиномиальных функциональных моделей далее будем обозначать как метод общих наименьших квадратов TLS (Total least squares).

Для случая, когда вместо γ задается дисперсия ошибок объясняющей переменной σ_δ^2, в [9, 12] специально для оценивания полиномиальных зависимостей предложен метод, названный скорректированным МНК (Adjusted least squares – ALS). Опишем его основную идею.

Рассмотрим математическое ожидание *r*-й степени от x_i:

$$t_i^r = \mathrm{E}\left(x_i^r\right) = \mathrm{E}\left((X_i + \delta_i)^r\right).$$

При предположении о нормальности распределения δ_i оценки t_i^r можно определить исходя из рекурсивного соотношения

$$\hat{t}_i^{r+1} = x_i \hat{t}_i^r - r\sigma_\delta^2 \hat{t}_i^{r-1}$$

при $\hat{t}_i^{-1} = \hat{t}_i^0 = 1$.

Обозначим матрицу регрессоров уравнения (5) как *C*. Тогда каждый элемент вектора регрессоров $R = C^T y$ уравнения (5) можно оценить следующим образом:

$$\hat{R}_j = \sum_{i=1}^{N} \hat{t}_i^j y_i ,\ \ j = \overline{0, k} .$$

Для каждого элемента матрицы $P = C^T C$ получим оценку вида

$$\hat{P}_{rs} = \sum_{i=1}^{N} t_i^{r+s} ,\ \ r, s = \overline{0, k} .$$

В соответствии с МНК оценка вектора параметров уравнения (5) определяется как

$$\hat{\theta}_{ALS} = \hat{P}^{-1}\hat{R}. \tag{7}$$

Отметим, что метод ALS более простой и быстрый по сравнению с TLS.

Поскольку методы TLS и ALS основаны на разной априорной информации о распределении случайных ошибок, то для сопоставления результатов ALS с результатами TLS необходимо получить оценку соотношения дисперсий ошибок γ для метода скорректированных наименьших квадратов. При заданном значении σ_δ^2 в [9] получена оценка дисперсии ошибки отклика σ_ε^2, которая позволяет найти оценку соотношения дисперсий ошибок $\hat{\gamma}_{ALS}$ следующим образом:

$$\hat{\gamma}_{ALS} = \frac{N\sigma_\delta^2}{\sum_{i=1}^{N} y_i^2 - \hat{R}^T\hat{\theta}_{ALS}}. \tag{8}$$

Существенная проблема, сказывающая на качестве оценивания нелинейных функциональных моделей, связана с корректным определением дисперсий ошибок. В большинстве исследований на эту тему [13, 14] для оценивания привлекается дополнительная информация (инструментальные переменные, повторные наблюдения). Однако выбор подходящих инструментов достаточно сложен, а получение повторных наблюдений требует дополнительных затрат на сбор информации, поэтому актуальной остается разработка подходов к идентификации, не требующих дополнительной информации.

Некоторые из таких подходов предполагают различия в распределениях ошибок и входного фактора структурных моделей [15] и тем самым требуют фиксации законов распределения случайных величин или их характеристик. Существуют и более гибкие непараметрические подходы [16]. Они вычислительно сложнее и разработаны только для структурных моделей, в то время как функциональные модели представляются более предпочтительными для многих технических и естественнонаучных приложений, поскольку основаны на более слабых предположениях. Именно поэтому остается существенный пробел в методах идентификации функциональных моделей без привлечения дополнительной информации.

2. Исследование свойств ошибок в функциональных моделях

Предлагаемый подход к оцениванию дисперсий ошибок по эмпирическим данным без априорных предположений и дополнительной информации основан на различии свойств ошибок в исходной модели, содержащей неизвестные детерминированные величины X_i:

$$y_i = f(X_i;\theta) + \varepsilon_i,$$

и модели, построенной с учетом некоторых случайных оценок этих величин:

$$y_i = f(\hat{X}_i;\theta) + \varepsilon_i. \tag{9}$$

Пусть $\hat{X}_i$ является несмещенной оценкой X_i, т. е. $\mathrm{E}(\hat{X}_i) = \mathrm{E}(X_i)$. Тогда представим оценку как $\hat{X}_i = X_i + \Delta_i$, где Δ_i – случайная погрешность оценивания истинных значений входного фактора, $\mathrm{E}(\Delta_i) = 0$. Будем предполагать, что дисперсия погрешности постоянна: $\mathrm{D}(\Delta_i) = \sigma_\Delta^2$ для любого i.

Тогда для модели (5) оцениваемое уравнение будет иметь вид

$$y_i = \theta_0 + \theta_1(X_i + \Delta_i) + \theta_2(X_i + \Delta_i)^2 + \ldots + \theta_k(X_i + \Delta_i)^k + \varepsilon_i.$$

После раскрытия скобок получим модель

$$y_i = \theta_0 + \theta_1 X_i + \theta_2 X_i^2 + \ldots + \theta_k X_i^k + \nu_i \tag{10}$$

со сложной ошибкой, выражаемой соотношением

$$\nu_i = \theta_1\Delta_i + 2\theta_2 X_i\Delta_i + \theta_2\Delta_i^2 + \ldots + \theta_k\sum_{j=1}^{k} C_k^j X_i^{k-j}\Delta_i^j + \varepsilon_i .$$

Видно, что при отсутствии смещения у оценки $\hat{X}_i$, ошибка ν_i имеет нулевое математическое ожидание. Теперь выразим дисперсию ошибки ν_i:

$$\mathrm{D}(\nu_i) = \theta_1^2\sigma_\Delta^2 + 4\theta_2^2 X_i^2\,\sigma_\Delta^2 + \theta_2^2\,\mathrm{D}\left(\Delta_i^2\right) + \ldots + \theta_k^2\sum_{j=1}^{k}\left(C_k^j X_i^{k-j}\right)^2\mathrm{D}\left(\Delta_i^j\right) + \sigma_\varepsilon^2 .$$

Из полученного выражения видно, что в случае, если хотя бы один из параметров $\theta_2,\ldots,\theta_k$ ненулевой, дисперсия $\mathrm{D}(\nu_i)$ будет зависеть от X_i. Для линейной зависимости дисперсия будет постоянной и равной $\mathrm{D}(\nu_i) = \theta_1^2\sigma_\Delta^2 + \sigma_\varepsilon^2$. В иных ситуациях будет наблюдаться непостоянство дисперсии, так называемая гетероскедастичность.

Если же погрешность оценки истинных значений входного фактора настолько мала, что можно считать $\sigma_\Delta^2 \approx 0$, то дисперсия ошибки будет постоянной: $\mathrm{D}(\nu_i) \approx \sigma_\varepsilon^2$. Значит, с одной стороны, если оценки $\hat{X}_i$ близки к истинным значениям входного фактора, то должна наблюдаться гомоскедастичность, т. е. однородность дисперсии ошибки в модели (10). С другой стороны, оценки $\hat{X}_i$ в описанном выше методе TLS зависят от заданного значения соотношений дисперсий ошибок. Обозначим заданное исследователем значение как $\tilde{\sigma}_\delta^2$, оно, вообще говоря, может отличаться от истинного σ_δ^2. Таким образом, основная идея состоит в том, чтобы подобрать такое $\tilde{\sigma}_\delta^2$, при котором в результате оценки модели (10) будет обнаружена гомоскедастичность.

3. Критерии выявления гетероскедастичности

Проверка гипотезы о постоянстве дисперсии ошибки (гомоскедастичности) может быть осуществлена с помощью ряда критериев [17, 18]. Выбраны основные из них: критерий Спирмена, Бартлетта и дисперсионного анализа (ANOVA), предлагаемый в работе [18].

Формально нулевая гипотеза об однородности дисперсий ошибок во всех наблюдениях может быть записана как

$$H_0 : \mathrm{D}(\nu_i) = \mathrm{const}\ \forall i . \tag{11}$$

В качестве альтернативной гипотезы рассматривается предположение о том, что дисперсия ошибки зависит от входного фактора. Если эта зависимость описывается функцией $g(\cdot)$, то $H_1 : \mathrm{D}(\nu_i) = g(X_i)$.

На основе критерия Спирмена [17] проверяется гипотеза об отсутствии ранговой корреляции между входным фактором и остатками модели, взятыми по модулю. Вывод делается на основе обычной *t*-статистики вида

$$t = \sqrt{N-2}\,\frac{\left|r\left(\hat{X}_i,|e_i|\right)\right|}{\sqrt{1-r^2\left(\hat{X}_i,|e_i|\right)}}, \tag{12}$$

где $r\left(\hat{X}_i,|e_i|\right)$ – коэффициент корреляции Спирмена между входным фактором и модулями остатков модели (9).

Критерий Спирмена можно считать самым простым и наименее затратным в плане времени вычислений. Это серьезное преимущество, поскольку проверку гипотезы об однородности дисперсий ошибок требуется производить для каждого предполагаемого значения $\tilde{\sigma}_\delta^2$, т. е. многократно.

Еще одно достоинство критерия: он не требует каких-либо дополнительных предположений относительно вида функции $g(\cdot)$. Однако, если функция имеет одну точку оптимума, соответствующую примерно середине отрезка, на котором находится область определения входной переменной, то ранговый коэффициент корреляции будет близок к нулю. В работе [18] это названо симметричной формой гетероскедастичности, и она не выявляется с помощью критерия Спирмена.

Критерии Бартлетта и ANOVA основаны на сравнении выборочной дисперсии остатков на разных участках области определения входной переменной. Для разбиения области изменения входного фактора на K непересекающихся интервалов (классов однородности) при использовании критериев число интервалов можно определить, например, по правилу Стёрджеса [19]:

$$K = \left[1 + \log_2 N\right],$$

где $[u]$ – целая часть числа u. Исходя из заданного числа интервалов, можно производить разбиение на равноотстоящие (одинаковой ширины) и равновероятные (с одинаковым числом наблюдений) интервалы.

Статистика Бартлетта определяется как

$$B = N\left(1 + \frac{1}{3(K-1)}\left(-\frac{1}{N} + \sum_{i=1}^{K}\frac{1}{n_i}\right)\right)^{-1} \ln\left(\frac{1}{N}\sum_{i=1}^{K} n_i s_i^2 \left(\prod_{i=1}^{K} s_i^{2n_i}\right)^{-\frac{1}{N}}\right), \tag{13}$$

где n_i – количество наблюдений, попавших в i-й интервал; s_i^2 – оценка дисперсии остатков, попавших в i-й интервал.

По критерию ANOVA проверка гипотезы (11) осуществляется путем построения модели дисперсионного анализа:

$$\left|e_{ij}\right| = \mu + \alpha_i + u_{ij}, \tag{14}$$

где $\left|e_{ij}\right|$ – абсолютные значения остатков модели (9); α_i – эффект i-го интервала, $i = 1, ..., K$; μ – генеральное среднее; u_{ij} – случайная ошибка. Тем самым модель (14) наряду с константой μ включает K факторов, представляющих бинарные индикаторы принадлежности наблюдения i-му интервалу μ.

Известно, что в моделях вида (14) оценить можно только так называемые функции, допускающие оценку. Чаще всего такие функции представляют в виде парных сравнений главных эффектов $\left(\alpha_j - \alpha_i\right)$, где $j \neq i$. Модель (14) оценивается с помощью МНК, и для парных сравнений определяются t-статистики $t_{\alpha_j - \alpha_i}$. Итоговая статистика находится как

$$t_V = \max_{i=1,...,K,\, j=1,...,K:\, i\neq j} \left|t_{\alpha_j - \alpha_i}\right|. \tag{15}$$

В работе [18] с помощью вычислительных экспериментов проведено сравнение мощности критериев Бартлетта и ANOVA. Выявлено, что результаты зависят от используемого правила разбиения на интервалы и от выбранного числа интервалов. При малом числе наблюдений, попадающих в интервал, оценка дисперсии ошибки будет сильно варьироваться. Тем самым критерии будут скорее отвергать гипотезу о гомоскедастичности. Это проблема будет усугубляться при наличии аномальных наблюдений. Однако уменьшение числа интервалов и, соответственно, увеличение числа наблюдений, приходящихся на один интервал, может привести к тому, что существующая гетероскедастичность не будет выявлена. В этой связи нужно очень осторожно подходить к выбору правила разбиения на интервалы и исследовать несколько вариантов. Это, однако, увеличивает время расчетов, поэтому в целом критерии Бартлетта и ANOVA можно считать вычислительно более сложными по сравнению с критерием Спирмена.

4. Алгоритм оценивания дисперсии ошибки входного фактора

С учетом того, что необходимо многократно переоценивать функциональную модель и проверять гипотезу (11), предлагается комбинировать оценивание на основе методов скорректированных и общих наименьших квадратов. Метод скорректированных наименьших квадратов обладает таким важным преимуществом, как высокая скорость вычислений. Кроме того, на входе он использует значение дисперсии ошибки входного фактора, которое проще варьировать, чем соотношение дисперсий ошибок.

Действительно, в силу (2), вариация наблюдаемых значений входного фактора складывается из вариации истинных значений и дисперсии ошибки их измерения. Тогда можно ограничить интервал значений дисперсии ошибки входного фактора следующим образом:

$$0 \le \tilde{\sigma}_{\delta}^{2} \le s_x^2,$$

где s_x^2 – выборочная дисперсия входного фактора.

Следовательно, если задать шаг изменения значений дисперсии $\Delta\tilde{\sigma}_{\delta}^{2} = \frac{s_x^2}{l}$, где l – число интервалов сетки, то можно найти значение $\tilde{\sigma}_{\delta}^{2}$, при котором гипотеза об однородности дисперсий ошибок не отвергается с наибольшей вероятностью, т. е. при наименьшем значении статистик (12), (13) или (15). Выбор значения l влияет на точность определения оценки дисперсии: чем больше будет интервалов, тем точнее получится сетка. Однако увеличение числа l приведет к росту времени вычислений, поэтому необходимо выбрать некоторый компромиссный вариант, либо сначала использовать более грубую сетку, а затем производить расчеты на более мелкой сетке в районе найденного грубого приближения.

Метод скорректированных наименьших квадратов не позволяет, однако, найти оценки истинных значений входного фактора для расчета остатков и проверки гипотезы о гомоскедастичности. Зато это возможно сделать на основе метода общих наименьших квадратов.

С использованием представленных идей предлагается алгоритм оценивания дисперсии ошибки входного фактора в модели (1)–(3), (5).

Шаг 1. Считываются исходные данные x_i, y_i. Задаются степень полинома k, объем выборки N. Задается критерий проверки гипотезы об однородности дисперсий ошибок (выбор одного из трех вариантов: Спирмена, Бартлетта, ANOVA). Задаются K, l и способ разбиения на интервалы (равноотстоящие, равновероятные). Вычисляется $\Delta\tilde{\sigma}_{\delta}^{2} = \frac{s_x^2}{l}$. Полагается, что $\tilde{\sigma}_{\delta}^{2} = 0$.

Шаг 2. Если $\tilde{\sigma}_{\delta}^{2} = 0$, то по исходным данным с помощью МНК оценивается модель (5). Определяются остатки e_i.

Шаг 3. В противном случае, при $\tilde{\sigma}_{\delta}^{2} > 0$, модель (1)–(3), (5) по исходным данным оценивается с помощью метода скорректированных наименьших квадратов и определяются остатки на основе оценок истинных значений входного фактора, найденных с помощью метода общих наименьших квадратов. Шаг 3 включает следующие промежуточные шаги.

Шаг 3.1. По соотношению (7) находятся оценки параметров $\hat{\theta}_{ALS} = \left(\hat{\theta}_{ALS0},\ \hat{\theta}_{ALS1},\ \hat{\theta}_{ALS2},\ \ldots,\ \hat{\theta}_{ALSk}\right)$. Определяется оценка соотношения дисперсий ошибок (8). Задается переменная цикла i: = 1.

Шаг 3.2. Найденные оценки подставляются в выражение многочлена (6) вместо $\theta_0, \theta_1, \theta_2, \ldots, \theta_k$ и γ.

Шаг 3.3. С помощью алгоритма Дженкинса – Трауба находятся все корни полинома (6). Выбирается вещественный корень. Если вещественных корней несколько, выбирается тот, который обеспечивает минимальное значение функции (4). Он будет соответствовать оценке $\hat{X}_i$.

Шаг 3.4. Если $i < N$, то i: = i + 1 и возврат к шагу 3.2.

Шаг 3.5. Определяются остатки по соотношению

$$e_i = y_i - \left(\hat{\theta}_{ALS0} + \hat{\theta}_{ALS1}\hat{X}_i + \hat{\theta}_{ALS2}\hat{X}_i^2 + \ldots + \hat{\theta}_{ALSk}\hat{X}_i^k\right).$$

Шаг 4. По найденным остаткам проверяется гипотеза об однородности дисперсий ошибок (гомоскедастичности). В зависимости от выбранного критерия выделяются следующие промежуточные шаги.

Шаг 4.1. Если выбран критерий Спирмена, то по соотношению (12) вычисляется и сохраняется значение t-статистики.

Шаг 4.2. Если выбран критерий Бартлетта, с учетом выбранного значения K и способа разбиения на интервалы по соотношению (13) вычисляется и сохраняется значение статистики B.

Шаг 4.3. Если выбран критерий ANOVA, с учетом выбранного значения K и способа разбиения на интервалы по соотношению (15) вычисляется и сохраняется значение статистики t_V .

Шаг 5. Если $\tilde{\sigma}_\delta^2 < s_x^2$, то $\tilde{\sigma}_\delta^2 := \tilde{\sigma}_\delta^2 + \Delta\tilde{\sigma}_\delta^2$ и возврат к шагу 3.

Шаг 6. Определяется оптимальное значение дисперсии ошибки входного фактора. В зависимости от выбранного критерия выделяются следующие промежуточные шаги.

Шаг 6.1. Если выбран критерий Спирмена, то выбирается значение $\tilde{\sigma}_\delta^2$, соответствующее минимальному из сохраненных значений t-статистики.

Шаг 6.2. Если выбран критерий Бартлетта, то выбирается значение $\tilde{\sigma}_\delta^2$, соответствующее минимальному из сохраненных значений статистики B.

Шаг 6.3. Если выбран критерий ANOVA, выбирается значение $\tilde{\sigma}_\delta^2$, соответствующее минимальному из сохраненных значений статистики t_V .

Алгоритм реализован в статистической среде R [20]. При этом использовались разработанные ранее реализации алгоритмов TLS и ALS [10].

5. Результаты вычислительных экспериментов

Для проверки работы предложенного алгоритма использовались три модели и пять схем эксперимента. Общий вид модели:

$$y_i = 2\left(X_i + m - \delta_i\right)^2 + \varepsilon_i ,$$

где для j-й модели $m = j - 1$. Коэффициенты подобраны так, чтобы кривые отличались по степени нелинейности. Эти отличия хорошо видны на рис. 1. Модель 1 – это самый нелинейный случай, модель 3 соответствует самой низкой степени нелинейности, модель 2 – это промежуточный вариант.

Во всех экспериментах $\sigma_\varepsilon^2 = 0{,}25$, истинные значения входного фактора задавались как квантили стандартного нормального распределения величины S. Значения S задавались по равномерной сетке от 0,001 до 0,999. Объем выборки составлял 1000.

Дисперсия ошибки входного фактора варьировалась в зависимости от схемы эксперимента. Для схемы 1 полагалось $\sigma_\delta^2 = 0$. Далее для j-й схемы $\sigma_\delta^2 = 0{,}01\left(j+1\right)^2$.

В соответствии с каждой моделью и схемой эксперимента строились случайные выборки объемом 1000. По каждой выборке производился расчет по предложенному алгоритму. Результаты усреднялись по 500 повторениям.

Для проверки влияния на результаты работы алгоритма задаваемых параметров число интервалов варьировалось от 5 до 11. Отметим, что по правилу Стёрджеса при N = 1000 K = 11. Производилось равновероятное разбиение на интервалы во всех тестах. Значение l выбрано равным 20.

Результаты оценивания дисперсий σ_δ^2 для разных моделей представлены в таблице. Указаны средние значения по 500 выборкам, в скобках приведено стандартное отклонение. Для критериев Бартлетта и ANOVA результаты приведены только для оптимального числа интервалов K^*, при котором полученные оценки в среднем ближе к истинным значениям.

Оценки дисперсии ошибки входного фактора при использовании различных критериев выявления гетероскедастичности

Схема	Модель 1			Модель 2			Модель 3		
σ_δ^2	t	B	t_V	t	B	t_V	t	B	t_V
0	0,317 (0,15)	0,000 (0,00)	0,000 (0,00)	0,000 (0,00)	0,000 (0,00)	0,000 (0,00)	0,000 (0,00)	0,000 (0,00)	0,000 (0,00)
0,09	0,249 (0,33)	0,097 (0,11)	0,110 (0,19)	0,058 (0,01)	0,292 (0,30)	0,057 (0,01)	0,057 (0,00)	0,073 (0,07)	0,057 (0,00)
0,16	0,484 (0,42)	0,242 (0,15)	0,268 (0,28)	0,119 (0,02)	0,699 (0,08)	0,122 (0,03)	0,117 (0,02)	0,268 (0,16)	0,117 (0,02)
0,25	0,648 (0,42)	0,231 (0,13)	0,404 (0,25)	0,195 (0,04)	0,750 (0,07)	0,186 (0,04)	0,179 (0,03)	0,370 (0,17)	0,182 (0,04)
0,36	0,706 (0,40)	0,604 (0,46)	0,322 (0,12)	0,278 (0,05)	0,892 (0,12)	0,287 (0,07)	0.259 (0,04)	0,463 (0,22)	0,263 (0,04)
K^*	–	5	5	–	5	5	–	5	5

Если сравнивать результаты применения критерия Спирмена для разных моделей, то видно, что для модели 1 оценки дисперсии наиболее существенно отличаются от истинных значений при любом уровне шума. Это объясняется тем, что модель 1 соответствует упомянутой выше симметричной форме гетероскедастичности, когда этот критерий не позволяется корректно определить ее наличие. Между тем для моделей 2 и 3 использование критерия Спирмена обеспечивает примерно такие же результаты, как и использования критерия ANOVA. Однако вычисление коэффициента корреляции – гораздо более быстрая операция по сравнению с построением дисперсионной модели. Следовательно, если степень нелинейности зависимости не слишком высокая, можно рекомендовать в рамках предложенного алгоритма использовать критерий Спирмена.

Оценки дисперсии, полученные при использовании критерия Бартлетта, в большинстве случаев оказались завышенными в 2–3 раза по сравнению с истинными значениями. Наиболее приемлемый результат критерий показал на модели 1, т. е. в случае высокой степени нелинейности зависимости. И для этой модельной ситуации в большинстве случаев критерий Бартлетта дает меньшие стандартные ошибки оценок дисперсии по сравнению с критерием ANOVA.

Оптимальное число интервалов K^* везде получилось равным 5, поэтому можно рекомендовать выбирать небольшое число интервалов, а не руководствоваться правилом Стёрджеса, которое предписывает при таком объеме выборки разбивать на 11 интервалов. Кроме того, это позволит увеличить скорость расчетов.

Использование критерия ANOVA в рамках предложенного алгоритма можно считать компромиссным вариантом: он обеспечивает достаточно приемлемое качество оценивания во всех рассмотренных экспериментах.

Для того чтобы понять, как сказывается точность определения значения σ_δ^2 на результатах оценивания модели в целом, воспользуемся стандартным критерием качества восстановления прогнозных значений отклика:

$$ESS = \frac{1}{2}\sum_{i=1}^{N}\left(y_i - f\left(X_i;\hat{\theta}\right)\right)^2 .$$

Значения ESS усреднены по всем 500 модельным выборкам для каждой схемы эксперимента. Результаты после применения критерия ANOVA представлены на рис. 2–4 сплошной линией. Для сравнения остаточные суммы квадратов посчитаны для случаев, когда модели оцениваются по МНК (пунктирная линия на рис. 2–4) и ALS с известным истинным значением σ_δ^2 (штриховая линия).

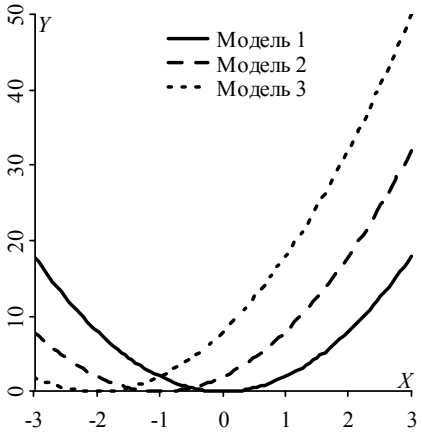

Рис. 1. Модельные кривые

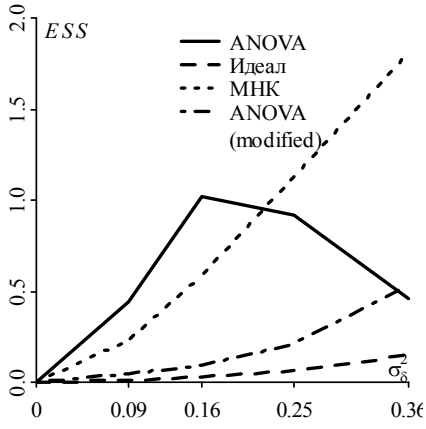

Рис. 2. Остаточная сумма квадратов для модели 1, $K^* = 5$

В целом, с ростом дисперсии ошибки входного фактора, использование предложенного подхода дает значительный выигрыш в точности восстановления кривой $Y = f\left(X;\theta\right)$ по сравне-

нию с оцениванием обычным МНК. Неудовлетворительные результаты применения критерия ANOVA для модели 1 (рис. 2) побудили нас к более подробному исследованию поведения статистики (15). По полученным в этих экспериментах значениям оценок дисперсии построены ящики с усами (рис. 5), на которых хорошо видна неоднородность значений статистики. Для ее объяснения проанализирована зависимость статистики от предполагаемого исследователем значения $\tilde{\sigma}_\delta^2$. Оказывается, что здесь обнаруживается несколько локальных минимумов. Это иллюстрирует рис. 6, построенный для модели 1 схемы 2 при $K^* = 5$. Следовательно, для разных выборочных реализаций оценка попадает то в один, то в другой локальный оптимум.

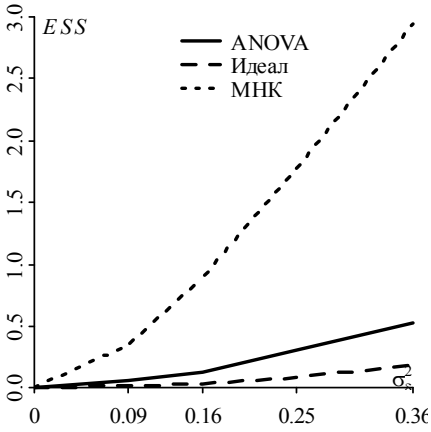

Рис. 3. Остаточная сумма квадратов для модели 2, $K^* = 5$

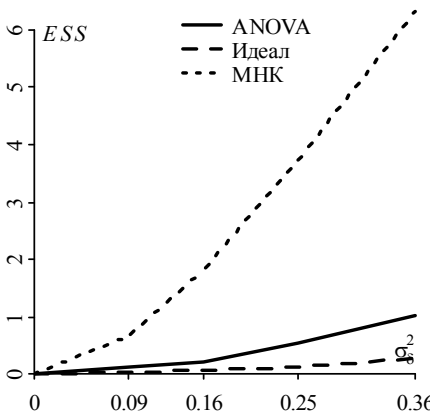

Рис. 4. Остаточная сумма квадратов для модели 3, $K^* = 5$

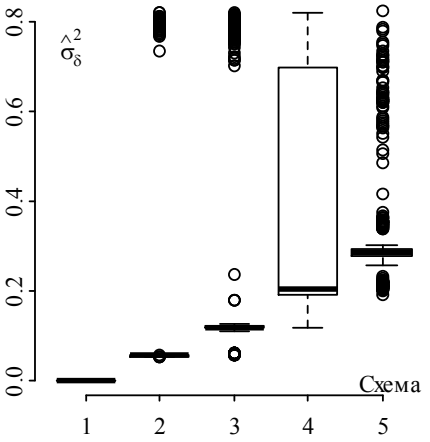

Рис. 5. Оценки дисперсии ошибки в модели 1, $K^* = 5$

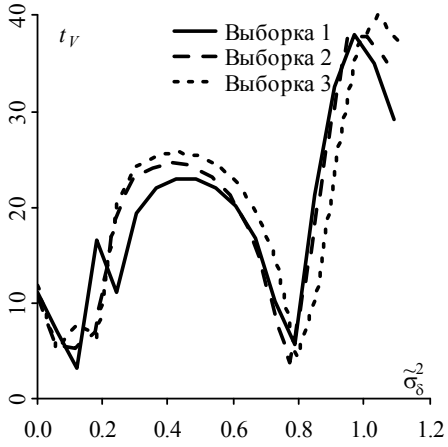

Рис. 6. Зависимость статистики t_V от заданного значения σ_δ^2

Исходя из исследования свойств ошибки в модели (9), следует ожидать, что при $\sigma_\delta^2 \neq 0$ использование в качестве $\hat{X}_i$ наблюдаемых значений x_i (т. е. предположение о нулевой дисперсии ошибки входного фактора) приведет к выявлению гетероскедастичности, т. е. гипотеза (11) будет отвергнута. По мере приближения к истинному значению σ_δ^2 значение статистики t_V должно уменьшаться, пока не достигнет минимума в точке, наиболее приближенной к ситуации отсутствия гетероскедастичности, в связи с чем нас интересует первый минимум статистики критерия выявления гетероскедастичности, а не минимум на всем промежутке $0 \le \tilde{\sigma}_\delta^2 \le s_x^2$. Поэтому предложенный выше алгоритмы был модифицирован.

На *шаге 4* текущее k-е значение статистики сравнивалось с предыдущим $(k-1)$-м, и если оно оказывалось больше, то значение $\widetilde{\sigma}_\delta^2$, соответствующее $(k-1)$-му значению статистики, считалось искомой оценкой $\hat{\sigma}_\delta^2$. На этом алгоритм заканчивался. Такая реализация позволила снизить остаточную сумму квадратов, что отображено на рис. 2 штрихпунктирной кривой ANOVA (modified).

Таким образом, предложенная идея оценки нелинейных функциональных моделей без дополнительной информации на основе выявления гомоскедастичности обеспечивает гораздо меньшую (в 3–8 раз) остаточную сумму квадратов, чем МНК, и в 3–4 раза большую, чем в идеальном случае, когда известно истинное значение σ_δ^2. Поэтому в ситуациях, когда исследователь не располагает точным значением погрешности измерения входного фактора, разработанный подход может быть рекомендован к применению.

Заключение

Основная проблема оценивания функциональных моделей заключается в том, что, как правило, значения погрешностей измерения точно неизвестны. Существующие подходы к оцениванию привлекают дополнительную информацию (инструментальные переменные, повторные наблюдения), которая не всегда доступна на практике. Нами предложена новая идея оценивания дисперсии ошибки входного фактора с использованием свойств ошибки, а именно однородности ее дисперсии при правильном выборе параметра $\widetilde{\sigma}_\delta^2$. Разработанный на этой основе алгоритм оценивания полиномиальных моделей протестирован в ходе вычислительных экспериментов. Выявлено, что в результате обеспечивается гораздо меньшая остаточная сумма квадратов, чем при использовании МНК, т. е. подход может быть рекомендован к использованию на практике. На его основе могут быть разработаны методы оценивания многофакторных моделей более сложной структуры, например уравнения полимеризации в задачах синтеза высокомолекулярных соединений.

СПИСОК ЛИТЕРАТУРЫ

1. *Polic L., Duever T. A., Penlidis A.* Case Studies and Literature Review on the Estimation of Copolymerization Reactivity Ratios // J. Polym. Sci., Part A: Polym. Chem. 1998. Vol. 36. P. 813–822.

2. *Kazemi N., Duever T. A., Penlidis A.* A Powerful Estimation Scheme with the Error-In-Variables-Model for Nonlinear Cases: Reactivity Ratio Estimation Examples // Comput. Chem. Eng. 2013. Vol. 48. P. 200–208.

3. *Zwanzig S.* On the criteria for experimental design in nonlinear error-in-variables models // Advances in Stochastic Simulation Methods. Birkhäuser Boston, 2000. P. 153–163.

4. *Fuller W. A.* Measurement error models. New York: John Wiley and Sons, Inc., 1987. 440 p.

5. *Кендалл М., Стьюарт А.* Статистические выводы и связи. М.: Наука, 1973. 900 с.

6. *Kukush A., Markovsky I., Van Huffel S.* Consistent estimation in an implicit quadratic measurement error model // Computational statistics & data analysis. 2004. Vol. 47, no. 1. P. 123–147.

7. *Bektas S.* Least squares fitting of ellipsoid using orthogonal distances // Boletim de Ciências Geodésicas. 2015. Vol. 21, no. 2. P. 329–339.

8. *Schneeweiss H., Augustin T.* Some recent advances in measurement error models and methods. In: Modern Econometric Analysis / O. Hübler, J. Frohn. Berlin: Springer, 2006. P. 183–198.

9. *Cheng C.-L., Schneeweiss H.* Polynomial regression with errors in the variables // Journal of the Royal Statistical Society: Series B. 1998. Vol. 60. P. 189–199.

10. *Денисов В. И., Тимофеева А. Ю., Хайленко Е. А., Бузмакова О. И.* Устойчивое оценивание нелинейных структурных зависимостей // Сибирский журнал индустриальной математики. 2013. № 4. С. 47–60.

11. *Грешилов А. А., Стакун В. А., Стакун А. А.* Математические методы построения прогнозов. М.: Радио и связь, 1997. 112 с.

12. *Kukush A., Markovsky I., Van Huffel S.* Consistent fundamental matrix estimation in a quadratic measurement error model arising in motion analysis // Comput. Statist. Data Anal. 2002. Vol. 41, no. 1. P. 3–18.

13. *Schennach S. M.* Estimation of nonlinear models with measurement error // Econometrica. 2004. Vol. 72, no. 1. P. 33–75.

14. *Hausman J., Newey W., Ichimura H., Powell J.* Measurement errors in polynomial regression models // Journal of Econometrics. 1991. Vol. 50, no. 3. P. 273–295.

15. *Макарова Т. А., Тырсин А. Н.* Идентификация линейных регрессионных моделей при наличии погрешностей во входных и выходных данных // Автометрия. 2012. Т. 48, № 1. С. 56–62.

16. *Schennach S. M., Hu Y.* Nonparametric identification and semiparametric estimation of classical measurement error models without side information // Journal of the American Statistical Association. 2013. Vol. 108, no. 501. P. 177–186.

17. *Тимофеев В. С., Фаддеенков А. В., Щеколдин В. Ю.* Эконометрика. Новосибирск: Изд-во НГТУ, 2015. 354 с.

18. *Тимофеев В. С., Фаддеенков А. В.* Исследование критериев обнаружения гетероскедастичности в регрессионных моделях // Науч. вестн. Новосибирск. гос. техн. ун-та. 2007. № 4 (29). С. 3–14.

19. *Sturges H. A.* The choice of a class interval // Journal of the American Statistical Association. 1926. Vol. 21, no. 153. P. 65–66.

20. *R Core Team*. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, 2016. URL: http://www.R-project.org/.

Статья поступила в редакцию 20.03.2017

ИНФОРМАЦИЯ ОБ АВТОРАХ

Денисов Владимир Иванович – Россия, 630073, Новосибирск; Новосибирский государственный технический университет; д-р техн. наук, профессор; советник проректора по научной работе; videnis@nstu.ru.

Тимофеева Анастасия Юрьевна – Россия, 630073, Новосибирск; Новосибирский государственный технический университет; канд. экон. наук; доцент кафедры экономической информатики; a.timofeeva@corp.nstu.ru.

V. I. Denisov, A. Y. Timofeeva

ESTIMATING DISPERSION OF INPUT FACTOR ERROR IN THE POLYNOMIAL FUNCTIONAL MODEL IN THE PRESENCE OF HOMOSCEDASTICITY

Abstract. The functional error-in-variable models don't fit within standard regression formulation for the reason that input factors are unknown determinate variables which in practice have random errors. Usually, estimation of such models is performed using additional information: about input factor error variance (adjusted least squares estimator, developed specifically for estimating the polynomial dependencies) or the relation of the factor error variances (total least squares estimator). Their values are typically given by a priori assumptions. The paper attempts to weaken the model assumptions, namely to eliminate the need to set the input factor error dispersion due to the possibility of its estimation on the same data, for which the non-linear model is recovered, i.e. without additional information. This possibility occurs when the measurement errors are homogeneous. Then, if the estimates of unobservable input factor values are close to the true, homoscedasticity of errors should be detected, which is broken as soon as the input factor in the nonlinear model contains errors. In this paper it is shown analytically for polynomial models. Thus, in the proposed algorithm, such an estimate of the dispersion of the input factor error is selected, which minimizes test statistic of heteroskedasticity detection. In the computational experiments the algorithm outputs were compared by different criteria to test the hypothesis of homogeneity of error variance. Besides, the approximation accuracy was compared based on found estimates and using a usual least squares estimator. It was found that the developed algorithm provides a significant advantage for the residual sum of squares and thus can be recommended for use in practice.

Key words: errors-in-variables model, functional case, error dispersion, input factor, heteroskedasticity, Spearman criterion, Bartlett criterion, ANOVA criterion, adjusted least squares estimator, total least squares estimator.

REFERENCES

1. Polic L., Duever T. A., Penlidis A. Case Studies and Literature Review on the Estimation of Copolymerization Reactivity Ratios. *J. Polym. Sci., Part A: Polym. Chem*., 1998, vol. 36, pp. 813-822.

2. Kazemi N., Duever T. A., Penlidis A. A Powerful Estimation Scheme with the Error-In-Variables-Model for Nonlinear Cases: Reactivity Ratio Estimation Examples. *Comput. Chem. Eng.*, 2013, vol. 48, pp. 200-208.

3. Zwanzig S. *On the criteria for experimental design in nonlinear error-in-variables models. Advances in Stochastic Simulation Methods*. Birkhäuser Boston, 2000. P. 153-163.

4. Fuller W. A. *Measurement error models*. New York: John Wiley and Sons, Inc., 1987. 440 p.

5. Kendall M. G., Stuart A. Inference and Relationship. Griffin, 1973. 900 p. (Russ. ed.: Kendall M., St'iuart A. Statisticheskie vyvody i sviazi. Moscow, Nauka Publ., 1973. 900 p.).

6. Kukush A., Markovsky I., Van Huffel S. Consistent estimation in an implicit quadratic measurement error model. *Computational statistics & data analysis*, 2004, vol. 47, no. 1, pp. 123-147.

7. Bektas S. Least squares fitting of ellipsoid using orthogonal distances. *Boletim de Ciências Geodésicas*, 2015, vol. 21, no. 2, pp. 329-339.

8. Schneeweiss H., Augustin T. *Some recent advances in measurement error models and methods*. In: Modern Econometric Analysis / O. Hübler, J. Frohn. Berlin: Springer, 2006. P. 183-198.

9. Cheng C.-L., Schneeweiss H. Polynomial regression with errors in the variables. *Journal of the Royal Statistical Society: Series B*, 1998, vol. 60, pp. 189-199.

10. Denisov V. I., Timofeeva A. Iu., Khailenko E. A., Buzmakova O. I. Ustoichivoe otsenivanie nelineinykh strukturnykh zavisimostei [Robust estimation of nonlinear structural dependences]. *Sibirskii zhurnal industrial'noi matematiki*, 2013, no. 4, pp. 47-60.

11. Greshilov A. A., Stakun V. A., Stakun A. A. *Matematicheskie metody postroeniia prognozov* [Mathematical methods of forecast modelling]. Moscow, Radio i sviaz' Publ., 1997. 112 p.

12. Kukush A., Markovsky I., Van Huffel S. Consistent fundamental matrix estimation in a quadratic measurement error model arising in motion analysis. *Comput. Statist. Data Anal*., 2002, vol. 41, no. 1, pp. 3-18.

13. Schennach S. M. Estimation of nonlinear models with measurement error. *Econometrica*, 2004, vol. 72, no. 1, pp. 33-75.

14. Hausman J., Newey W., Ichimura H., Powell J. Measurement errors in polynomial regression models. *Journal of Econometrics*, 1991, vol. 50, no. 3, pp. 273-295.

15. Makarova T. A., Tyrsin A. N. Identifikatsiia lineinykh regressionnykh modelei pri nalichii pogreshnostei vo vkhodnykh i vykhodnykh dannykh [Identification of linear regression models in the presence of errors in input and output data]. *Avtometriia*, 2012, vol. 48, no. 1, pp. 56-62.

16. Schennach S. M., Hu Y. Nonparametric identification and semiparametric estimation of classical measurement error models without side information. *Journal of the American Statistical Association*, 2013, vol. 108, no. 501, pp. 177-186.

17. Timofeev V. S., Faddeenkov A. V., Shchekoldin V. Iu. *Ekonometrika*. Novosibirsk, Izd-vo NGTU, 2015. 354 p.

18. Timofeev V. S., Faddeenkov A. V. Issledovanie kriteriev obnaruzheniia geteroskedastichnosti v regressionnykh modeliakh [Analysis of criteria of heteroscedasticity finding in the regression models]. *Nauchnyi vestnik Novosibirskogo gosudarstvennogo tekhnicheskogo universiteta*, 2007, no. 4 (29), pp. 3-14.

19. Sturges H. A. The choice of a class interval. *Journal of the American Statistical Association*, 1926, vol. 21, no. 153, pp. 65-66.

20. *R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing*. Vienna, Austria, 2016. Available at: http://www.R-project.org/.

The article submitted to the editors 20.03.2017

INFORMATION ABOUT THE AUTHORS

Denisov Vladimir Ivanovich – Russia, 630073, Novosibirsk; Novosibirsk State Technical University; Doctor of Technical Sciences, Professor; Advisor to the Pro-rector on Scientific Work; videnis@nstu.ru.

Timofeeva Anastasiia Yurievna – Russia, 630073, Novosibirsk; Novosibirsk State Technical University; Candidate of Economics; Assistant Professor of the Department of Computer Science in Economics; a.timofeeva@corp.nstu.ru.