DOI: 10.18372/2310-5461.44.14321
УДК 656.7.071: 656.7.052.002.5 (045)

V. M. Kuzmin, PhD
National Aviation University
orcid.org/0000-0003-4461-9297
e-mail: kuzmin_vn@i.ua;

M. Yu. Zaliskyi, PhD, associate professor
National Aviation University
orcid.org/0000-0002-1535-4384
e -mail: maximus2812@ukr.net;

Yu. V. Petrova, PhD, associate professor
National Aviation University
orcid.org/0000-0002-3768-7921
e -mail: panijulia.p@gmail.com;

I. V. Cheked, PhD, associate professor
National Aviation University
orcid.org/0000-0002-2702-6812
e -mail: igrnew@ukr.net

COMPARATIVE ANALYSIS OF TWO METHODS FOR TAKING INTO ACCOUNT HETEROSKEDASTICITY DURING MATHEMATICAL MODELS BUILDING

Introduction

In econometrics problems, it is often necessary to analyze empirical data with the subsequent construction of optimal mathematical models. The technique of heteroskedasticity detecting and accounting in the form of appropriate tests has recently been actively used when constructing such models [1–3].

Heteroskedasticity is characterized by different values of variance for data in one sample. Since the heteroskedasticity detection is not easy task, until recently it was believed that the standard deviation is equal to a constant for all statistical data under analysis. Generally, taking into account heteroskedasticity is a new direction in the empirical data processing, but standardized rules and methods have not been established. According to [4], the criteria for choosing the best model are:

1. The smallest number of model parameters.
2. The most simple form.
3. Physical validity.
4. The minimum sum of squared deviations.
5. Minimal variance.

Analysis of the latest research and publications

The literature analysis shows that enough attention is paid to the problems of mathematical models construction with regard to heteroskedasticity [5]. Estimates of the unknown coefficients of the approximating function become less effective if one disregards heteroskedasticity (if it actually occurs).

The ordinary least squares method and the weighted least squares method are used as a tool for mathematical models building [3].

There are different approaches to the calculation of weight coefficients in situations of heteroskedasticity [3]. Classic procedures for heteroscedasticity verification of empirical data have been proposed by Goldfeld-Quandt and Glaser and were demonstrated in [1; 2]. However, these procedures have disadvantages – they do not give a specific value of heteroscedasticity coefficient assessment [6; 7].

The analysis shows that there are a large number of tests for heteroskedasticity, but there is no unified quantitative measure in the current literature that indicates the presence of heteroskedasticity and could be used to directly calculate weightibg coefficients. The presence of heteroskedasticity can lead to approximation accuracy decreasing in case of ordinary least squares method utilization [8].

Therefore, this article solves an actual scientific and technical problem: the substantiation of heteroskedasticity quantitative measures and comparative analysis.

Problem Statement

Define a problem of research mathematically. Suppose, for a set of two-dimensional statistics

(x_i, y_i), there is a set of approximation functions $\hat{y}_i = f_n(x_i, \vec{a}_{m,n})$, where $\vec{a}_{m,n}$ is a vector of m parameters for the approximation function, n is a number of approximation functions. For each approximation function standard deviation σ between real values y_i and evaluation $\hat{y}_i$ can be calculated. In this case, selection of the best mathematical model will be carried out in accordance with the following criterion

$$n = \inf\left(s \in \mathrm{N} \,\forall j : \sigma(f_s(x_i, \vec{a}_{m,s})) \le \sigma(f_j(x_i, \vec{a}_{m,j}))\right).$$

The aim of the paper

The aim of this paper is comparative analysis of two methods for heteroskedasticity detection and accounting during mathematical models building.

To achieve the aim of the research, the following tasks were solved:

– analysis of experimental data and model building according to the ordinary least squares method;

– calculation of weighting coefficients for heteroskedasticity using the direct method;

– calculation of weighting coefficients for heteroskedasticity using optimal basic function;

– estimation of numerical value of heteroskedasticity index;

– result comparison in case of two methods implementation for taking into account heteroskedasticity.

Initial data analysis

Consider an example of experimental data on Lucerne yield dependence on the level of irrigation [9]. The data are given in Table 1. These data are characterized by the fact that in each section are the results of multiple measurements. This makes it possible to find the heteroskedasticity equation by a direct method.

The graphical representation of the initial data is shown in Fig. 1.

Table 1

Lucerne yield dependence on the level of irrigation

	Irrigation in inches							
	0	12	18	24	30	36	48	60
Lucerne yield	2.35	4.31	5.69	6.00	7.53	7.58	8.05	5.55
	2.75	4.78	6.46	6.89	7.97	8.22	8.45	7.25
	2.89	4.84	7.02	7.96	8.32	8.63	8.63	10.17
	3.85	5.83	8.02	8.32	9.43	9.33	8.83	10.70
	5.52	6.51		8.38	9.54	9.38	9.52	
	5.94	7.52		9.96	11.06	12.48	10.62	

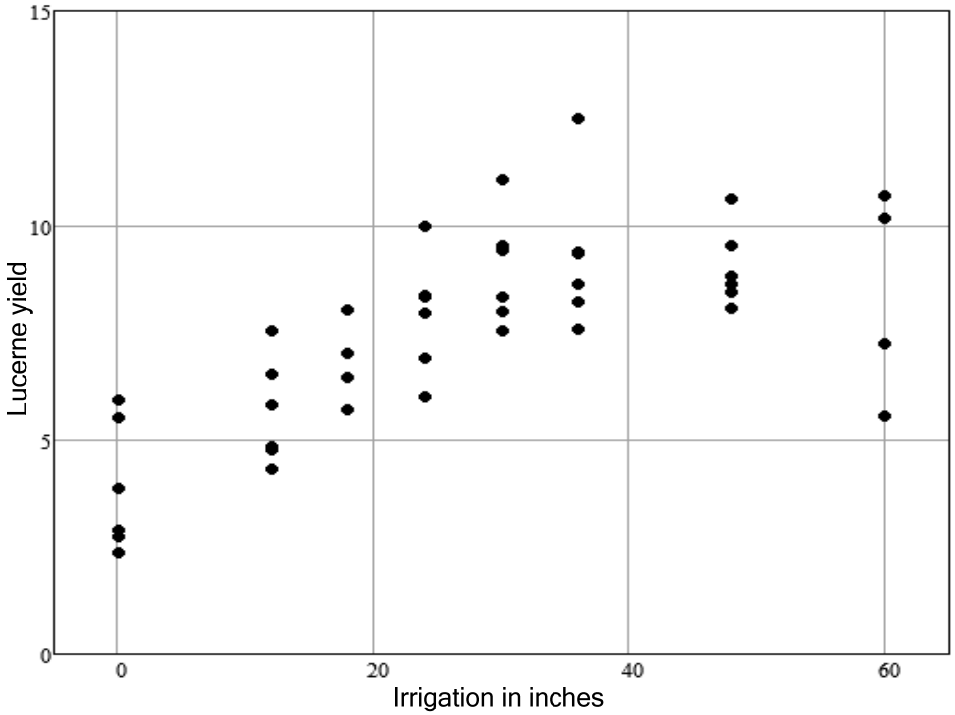

Fig. 1. The graphical representation of the initial data

In nature, the phenomenon of heteroskedasticity is quite common, but in practice it is not taken into account due to the complexity of the calculations and the lack of a single methodology.

The heteroskedasticity is very often present in the equations for the dependence of yield on the level of water use and fertilizer consumption.

In this experiment, the task was to obtain a fragmentary section of a mathematical model in order to determine the most accurate maximum yield value and the conditions for its achievement.

Carrying out the multiple experiments require a lot of material and time resources.

The first step in the construction of the mathematical model is to approximate the data by a second order parabola using the ordinary least squares method.

The resulting equation has the following form:

$$y(x) = 3.543 + 0.252x - 2.823 \cdot 10^{-3} x^2 .$$

The standard deviation is 1.406.

Then the equation with confidence limits is given by

$$y(x) = 3.543 + 0.252x - 2.823 \cdot 10^{-3} x^2 \pm 2.812.$$

A graphic representation of the approximation by a parabola with confidence limits is shown in Fig. 2.

It should be noted that confidence intervals within the variance band are established on the basis of the assumption of their Gaussian nature. Therefore, a histogram of the distribution can be constructed. The sequence of plotting the histogram for the case of multiple measurements in each section is following:

1. At the first stage, it is a need to determine the section onto which all experimental points will be projected. In this case, this section corresponds to the abscissa $x = 36$. This section was chosen because it contains the maximum value. The selected section will be called basic.

2. At the second stage, each experimental value will be projected onto the base section along the trajectories of the second order parabola that are equidistant to the main parabola obtained by the ordinary least squares method.

3. At the third stage, the histogram is constructed for all points that were reduced to the base section.

The obtained histogram for the considered example is shown in Fig. 3.

Visual analysis of Fig. 3 shows that the obtained histogram can be described by a distribution with positive asymmetry.

So, the analysis of the constructed histogram allows us to clarify the statistical nature of the empirical data variance within confidence limits.

Methods of taking into account heteroskedasticity

It is known, that a clear methodology for calculating weighting coefficients is required to use the weighted least squares method [10]. Let us take into account heteroskedasticity based on the direct method. To do this, in each section the average values and standard deviations need to be calculated. The calculation results are shown in table 2.

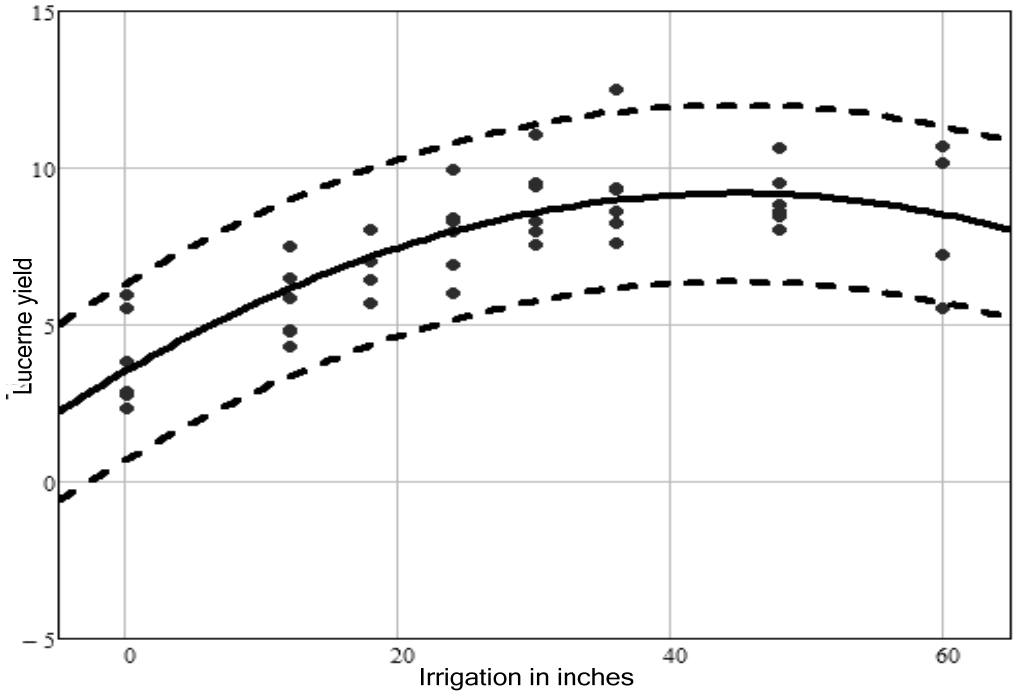

Fig. 2. Approximation of the initial data by a second order parabola with confidence limits

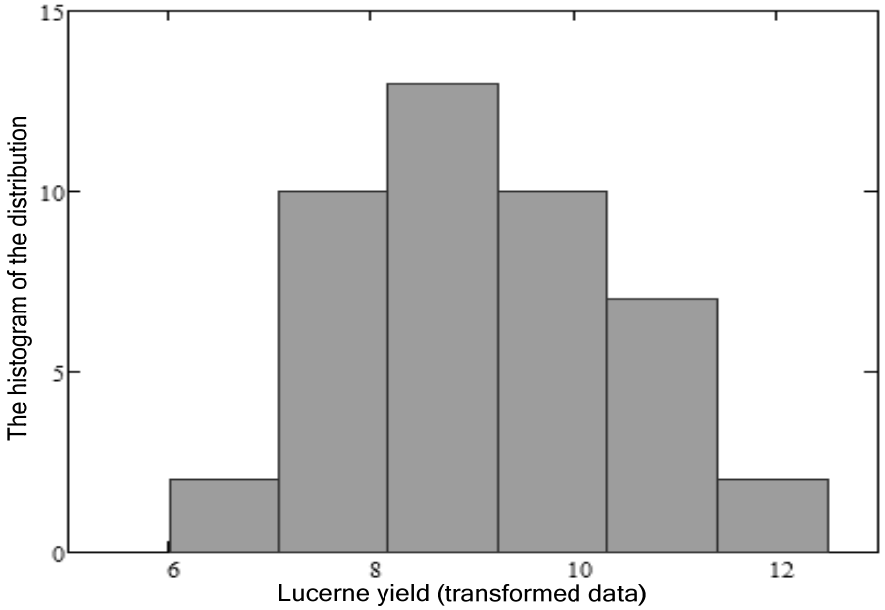

Fig. 3. The histogram of the distribution in the base section

Table 2

Mathematical expectations and standard deviations calculated for each section

n	1	2	3	4	5	6	7	8
m	3.883	5.632	6.698	7.918	8.975	9.27	9.017	8.418
σ	1.519	1.223	0.981	1.363	1.296	1.714	0.923	1.89

Based on the data from table 2 the equation of heteroskedasticity can be obtained. To do this, we approximate the dependence σ(*m*) by a linear function using the ordinary least squares method. As a result, the equation can be found:

$$\sigma(m) = 1.274 + 0.012m .$$

The graphical dependence σ(*m*) is shown in Fig. 4.

The calculation of heteroskedasticity coefficients is performed according to the formula

$$W_i = \left(\frac{\overline{\sigma}}{\sigma(m_i)} \right)^2 ,$$

where $\overline{\sigma}$ is an average standard deviation.

The weighting coefficients are presented in table 3. These weights are used to obtain a second order parabola using the weighted least squares method. The resulting equation has the form

$$y(x) = 3.557 + 0.251x - 2.8 \cdot 10^{-3} x^2 .$$

Consider a new method of taking into account heteroskedasticity. In this case, the weighting coefficients are calculated by the formula [11; 12].

$$W_i = \left(\frac{\overline{y}}{y(x_i)} \right)^h ,$$

where $\overline{y}$ is the average value of all empirical data, $y(x_i)$ is the value obtained by the equation of the second order parabola using the ordinary least squares method (base parabola), h is the heteroskedasticity index.

To find the optimal value of the heteroskedasticity index, we consider five options with heteroskedasticity indexes $h = \{-1; -0.5; 0; 0.5; 1\}$. The weighting coefficients for different heteroskedasticity indexes are presented in table 3.

The corresponding parabola equation was obtained for each value of the heteroskedasticity index. For each variant of the parabola, a weighted sum of squared deviations S were found. The obtained discrete dependence $S(h)$ is approximated by a parabola of the second order by the ordinary least squares method. The resulting equation has the form

$$S(h) = 81.04 - 0.075h + 0.295h^2 .$$

The minimum of this equation corresponds to the optimal heteroskedasticity index $h_{\text{opt}} = 0.126$.

As can be seen, heteroskedasticity index is small. However, with the growth of computational capabilities and scientific and technological progress, its accounting will allow to obtain more accurate and adequate mathematical models.

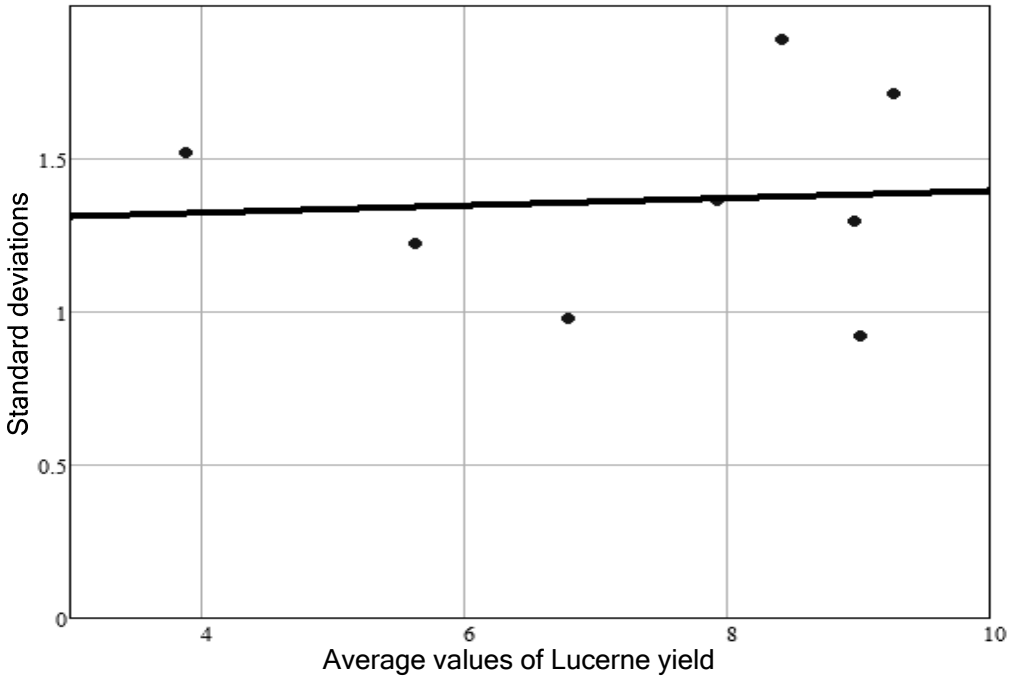

Fig. 4. The graphical dependence σ(*m*)

Table 3

The weighting coefficients for different values of heteroskedasticity indexes

Heteroskedasticity indexes	Sample section							
	1	2	3	4	5	6	7	8
Direct method	1.066	1.033	1.012	0.993	0.974	0.969	0.974	0.984
$h = -1$	0.474	0.824	0.959	1.066	1.146	1.199	1.224	1.14
$h = -0.5$	0.667	0.904	0.979	1.035	1.074	1.099	1.108	1.063
$h = 0$	1	1	1	1	1	1	1	1
$h = 0.5$	1.453	1.101	1.021	0.968	0.934	0.913	0.904	0.937
$h = 1$	2.062	1.21	1.044	0.941	0.875	0.837	0.818	0.874
$h_{opt} = 0.128$	1.093	1.024	1.006	0.993	0.984	0.978	0.975	0.983

The optimal equation obtained taking into account heteroskedasticity according to the second method has the form

$$y(x) = 3.564 + 0.251x - 2.798 \cdot 10^{-3} x^2 .$$

The initial data and two types of approximations with taking into account heteroskedasticity according to the first and second method are shown in Fig. 5. As can be seen, both equations practically coincide. This indicates about accounting reliability of heteroskedasticity by the second method.

Let us calculate the conditions for achieving the maximum possible yield according to the three approximation methodology (without and taking into account heteroskedasticity by the first and second methods).

The calculation results are shown in table 4.

As can be seen from table 4, the direct and the new methods for heteroskedasticity accounting give very close results that indicates about reliability and adequacy of the new method.

Table 4

Maximum yield and conditions for its achievement

	Without heteroskedasticity	Direct method for heteroskedasticity accounting	A new method for heteroskedasticity accounting
x_{opt}	44.695	44.8	44.792
y_{opt}	9.183	9.177	9.178

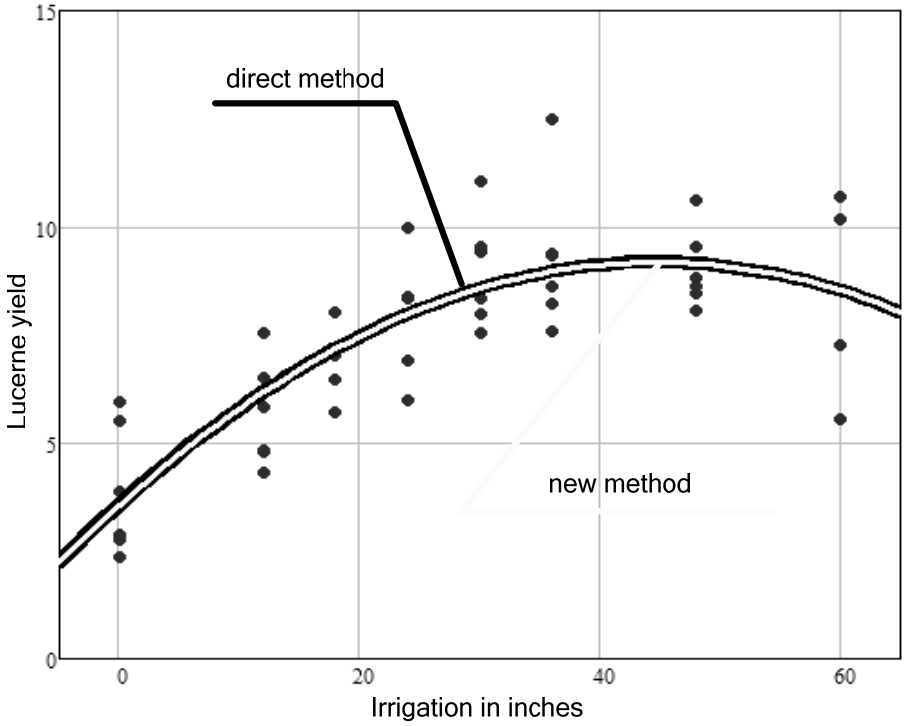

Fig. 5. Approximation of the initial data by the second degree parabola taking into account heteroskedasticity

Conclusion

The article is devoted to the problems of constructing mathematical models for empirical data taking into account heteroskedasticity. The analysis of data with multiple measurements in all sections is performed. Such data made it possible to apply the direct method for constructing the heteroskedasticity equation. A new method was proposed to bring them to a single section for a more correct determination of the probabilistic law of these data description.

The article describes a new method of heteroskedasticity accounting. A comparative analysis with a direct method showed approximately the same approximation results.

Thus, the new method of heteroskedasticity accounting allows us to construct a mathematical model, without carrying out multiple expensive measurements in each section, which can significantly reduce the time and resource costs by several times.

REFERENCES

1. **Goldfield S. M.,** Quandt R. E. Some tests for homoskedasticity. *Journal of the American Statistical Association*. 1965. Vol. 60. Pp. 539–547.

2. **Glejser H.** A new tests for heteroskedasticity. *Journal of the American Statistical Association*. 1969. Vol. 64. Pp. 316–323.

3. **Johnston J.** Econometric methods. New York: McGraw Hill, 1984. 568 p.

4. **Himmelblau D. M.** Process analysis by statistical methods. New York: John Wiley and Sons, 1970. 958 p.

5. **Бородич С. А.** Эконометрика. Минск: Новое знание, 2001. 408 с.

6. **Граббер Дж.** Эконометрика. Том 1. Введение в эконометрику. К.: Астарта, 1996. 398 с.

7. **Догерти К.** Введение в эконометрику. Москорова: ИНФРА-М, 2001. 402 с.

8. **Sushchenya L. M.,** Trubetskova I.L., Kuzmin V.N. A mathematical model of daphnia nutrition rate at different temperatures and food concentrations // Reports of the Academy of Sciences of the BSSR. – 1986. – Vol. XXX, № 4. – P. 376–379. (In Russian).

9. **Миллс Ф.** Статистические методы. М.: Государственное статистическое издательство, 1958. 800 с.

10. **Кузьмин В. М.,** Лапач С.М. Полигональная регрессия для проявлений гетероскедастичности в экономических задачах. *Экономика и управление*. 2007. № 1. С. 81-86.

11. **Kuzmin V. N.** The Statistical Analysis of Econometric Data under Heteroskedasticity. *Computer data analysis and modeling*. Proceedings of the Sixth International Conference (8 – 12 June 1998, Minsk). 1998. Vol. 2. Pp. 37–42.

12. **Kuzmin V.**, Zaliskyi M., Asanov M. Three-dimensional mathematical model in heteroskedasticity conditions in control systems. *IEEE 3rd International Conference on Methods and Systems of Navigation and Motion Control (MSNMC 2014)*. Kyiv: NAU, 2014, Proceedings. Pp. 139–142.

Кузьмин В. М., Заліський М. Ю., Петрова Ю. В., Чекед І. В.

ПОРІВНЯЛЬНИЙ АНАЛІЗ ДВОХ МЕТОДІВ УРАХУВАННЯ ГЕТЕРОСКЕДАСТИЧНОСТІ ПІД ЧАС ПОБУДОВИ МАТЕМАТИЧНИХ МОДЕЛЕЙ

У статті розглянуто задачу порівняльного аналізу двох методів урахування гетероскедастичності під час побудови математичних моделей. Урахування гетероскедастичності є новим напрямом під час аналізу емпіричних даних. Гетероскедастичність характеризується різними значеннями дисперсії для даних в одній вибірці. Наявність гетероскедастичності може призвести до зниження точності апроксимації у разі використання звичайного методу найменших квадратів. Тому в цій статті розглядається задача урахування гетероскедастичності під час аналізування емпіричних даних. Першим етапом побудови математичної моделі є апроксимація даних з використанням звичайного методу найменших квадратів. При цьому попередньо обирається апроксимуюча функція, виходячи із візуального аналізу структури статистичних даних. Наступним етапом побудови математичної моделі є урахування гетероскедастичності. Існують різні тести для виявлення гетероскедастичності. У цій статті розглянуто прямий метод побудови рівняння гетерокедастичності та новий метод, який порівнюється з прямим. Прямий метод заснований на обчисленні середніх значень та стандартних відхилень для кожного перетину початкової вибірки. Вагові коефіцієнти гетероскедастичності розраховуються відповідно до апроксимаційної залежності стандартних відхилень від середніх значень для статистичних даних з використанням лінійної функції. Такий метод має суттєвий недолік: він потребує кратних вимірювань для кожного перетину вибірки. Під час вирішення задач синтезу нового алгоритму виявлення та урахування гетероскедастичності автори пропонують нову кількісну міру гетероскедастичності. Оцінка запропонованого індексу гетероскедастичності виконується у такій послідовності: 1) для декількох варіантів можливих значень індексу гетероскедастичності розраховують відповідні апроксимаційні функції; 2) для кожної отриманою функції розраховують зважену суму квадратів відхилень; 3) визначають індекс гетероскедастичності, для якого зважена сума квадратів відхилень є мінімальною. У статті також розглянуто унікальний приклад емпіричних даних з кратними вимірюваннями у кожному перетині. Аналіз таких даних дозволив обґрунтувати надійність та адекватність нового методу виявлення та урахування гетероскедастичності. Новий метод урахування гетероскедастичності дозволяє побудувати математичну модель без проведення декількох вимірювань для кожного перетину.

Ключові слова: апроксимація; зважений метод найменших квадратів; гетероскедастичність; порівняльний аналіз; показник гетероскедастичності.

Kuzmin V. M., Zaliskyi M. Yu., Petrova Yu. V., Cheked I. V.

COMPARATIVE ANALYSIS OF TWO METHODS FOR TAKING INTO ACCOUNT HETEROSKEDASTICITY DURING MATHEMATICAL MODELS BUILDING

The article deals with the problem of comparative analysis of two methods of taking into account heteroskedasticity during mathematical models building. Heteroskedasticity accounting is a new trend for the empirical data analysis. Heteroskedasticity is characterized by different values of variance for data in one sample. The presence of heteroskedasticity can lead to approximation accuracy decreasing in case of ordinary least squares method utilization. Therefore, this article concentrates on the problem of heteroskedasticity accounting in case of empirical data analysis. The first step during the mathematical model building is to approximate the data using the ordinary least squares method. In this case, the approximation function is pre-selected in advance based on visual analysis of the statistical data structure. The next step in mathematical model building is to take into account heteroscedasticity. There are different tests for heteroskedasticity detection. This article discusses the direct method of the heteroskedasticity equation construction and the new method that is compared with the direct one. The direct method is based on the calculation of average values and standard deviations for in each section of initial sample. Heteroskedasticity weighting coefficients are calculated according to the approximation of standard deviations dependence on the average values for statistics using a linear function. This method has a significant weakness: it requires multiple measurements for each sample section. During solving the problems of synthesizing a new algorithm for heteroskedasticity detecting and accounting, the authors propose a new quantitative measure of heteroskedasticity. The estimation of the proposed heteroskedasticity index is performed in the following sequence: 1) for several options of possible values of the heteroskedasticity index, the corresponding approximation functions are calculated; 2) the sum of squared deviations is calculated for each obtained function; 3) the heteroskedasticity index is equal to value for which the sum of squared deviations is minimal. A unique example of empirical data with multiple measurements in each section is considered. The analysis of such data allowed justifying the reliability and adequacy of the new method for heteroskedasticity detecting and accounting. The new method of heteroskedasticity accounting allows us to construct the mathematical model without carrying out multiple expensive measurements in each section.

Keywords: approximation; weighted least squares method; heteroskedasticity; comparative analysis; heteroskedasticity index.

Кузмин В. Н., Залисский М. Ю., Петрова Ю. В., Чекед И. В.

СРАВНИТЕЛЬНЫЙ АНАЛИЗ ДВУХ МЕТОДОВ УЧЕТА ГЕТЕРОСКЕДАСТИЧНОСТИ ВО ВРЕМЯ ПОСТРОЕНИЯ МАТЕМАТИЧЕСКИХ МОДЕЛЕЙ

В статье рассмотрена задача сравнительного анализа двух методов учета гетероскедастичности при построении математических моделей. Учет гетероскедастичности является новым направлением при анализе эмпирических данных. Гетероскедастичность характеризуется различными значениями дисперсии для данных в одной выборке. Наличие гетероскедастичности может привести к снижению точности аппроксимации при использовании обычного метода наименьших квадратов. Поэтому в этой статье рассматривается задача учета гетероскедастичности во время анализа эмпирических данных. Первым этапом построения математической модели является аппроксимация данных с использованием обычного метода наименьших квадратов. При этом предварительно выбирается аппроксимирующая функция, исходя из визуального анализа структуры статистических данных. Следующим этапом построения математической модели является учет гетероскедастичности. Существуют различные тесты для выявления гетероскедастичности. В этой статье рассмотрены прямой метод построения уравнения гетероскедастичности и новый метод, который сравнивается с прямым. Прямой метод основан на вычислении средних значений и стандартных отклонений для каждого сечения выборки. Весовые коэффициенты гетероскедастичности рассчитываются в соответствии с аппроксимационной зависимостью стандартных отклонений от средних значений для статистических данных с использованием линейной функции. Такой метод имеет существенный недостаток: он требует кратных измерений для каждого сечения выборки. При решении задач синтеза нового алгоритма обнаружения и учета гетероскедастичности авторы предлагают новую количественную меру гетероскедастичности. Оценка предложенного индекса гетероскедастичности выполняется в следующей последовательности: 1) для нескольких вариантов возможных значений индекса гетероскедастичности рассчитывают соответствующие аппроксимирующие функции; 2) для каждой полученной функции рассчитывают взвешенную сумму квадратов отклонений; 3) определяют индекс гетероскедастичности, для которого взвешенная сумма квадратов отклонений является минимальной. В статье также рассмотрен уникальный пример эмпирических данных с кратными измерениями в каждом сечении. Анализ таких данных позволил обосновать надежность и адекватность нового метода обнаружения и учета гетероскедастичности. Новый метод учета гетероскедастичности позволяет построить математическую модель без проведения нескольких измерений для каждого сечения.

Ключевые слова: аппроксимация; взвешенный метод наименьших квадратов; гетероскедастичность; сравнительный анализ; индекс гетероскедастичности.

Стаття надійшла до редакції 25.10.2019 р.
Прийнято до друку 24.12.2019 р.