UDC 519.688+517.9
JEL Classification: F59
DOI: 10.15587/2312-8372.2019.182109

Kubiv S.

CHOICE OF THE ORDER OF THE REGRESSION MODEL FOR FORECASTING OF RANDOM NON-STATIONARY ECONOMIC PROCESSES

Об'єктом дослідження є гетероскедастичні процеси, які впливають на виробництво товарів військового призначення країн-експортерів. На сьогоднішній день збройні конфлікти є найбільш значущим фактором, який впливає на обсяги виробництва та експорту озброєння, оскільки передбачає наявність у сторін необхідної кількості озброєння та є в певному сенсі стохастичним процесом. Робота присвячена прогнозуванню стохастичних впливів на виробничі процеси товарів військового призначення країн-експортерів. В якості прикладу розглянуто економічну систему зі стохастичними впливами та проблемами вузьких місць у виробничих підрозділах. Модель процесу виходу продукції представлено у виді випадкового процесу з повільною нестаціонарністю (гетероскедастичного процесу). В ході дослідження використовувалися методи прогнозування нестаціонарних випадкових процесів. Досліджено задачу вибору та обґрунтування математичної моделі прогнозу гетероскедастичного процесу, що розглядається. Доведено, що найбільш спроможним методом короткострокового прогнозу є метод наближення Паде. Показано, що метод Паде, по суті, є методом апроксимації аналітичними (дрібно-раціональними) функціями, тому його можна інтерпретувати як метод побудови моделі авторегресії та ковзного середнього (АРКС). Розглянуті модифікації моделі АРКС, такі як модель авторегресії та інтегрованого ковзного середнього або авторегресії та фрактального інтегрованого ковзного середнього. Розроблено модифікований метод вибору порядку авторегресійної моделі за інформаційним критерієм Акаіке та за байєсівським інформаційним критерієм. Проаналізовано модельну задачу та приклади експериментальних залежностей. Запропоновано ефективну методику вибору порядку регресійних моделей, що застосовуються при практичному прогнозуванні стохастичних процесів, яка заснована на канонічних розкладах випадкової функції. Для розбиття функції розподілу на нееквідистантні інтервали з постійними інтенсивностями потоку використовується економічний рекурентний алгоритм. Результати розрахунків можуть бути використані для оптимального вибору порядку регресійної моделі, якою апроксимується реальний процес виробництва у вигляді часового ряду з випадковими зовнішніми впливами.

Ключові слова: *гетероскедастичність, дискретні часові ряди, модель авторегресії, стохастична система, апроксимація Паде, регресійна модель, порядок моделі, виробнича система.*

Received date: 17.06.2019
Accepted date: 05.07.2019
Published date: 30.10.2019

1. Introduction

One of the urgent tasks in the general problem of offset policy planning is forecasting the impacts of the implementation of offset agreements on subcontractors of exporting countries. Forecasting is carried out mainly through the use of mathematical models and methods and risk measurement.

The presented work is devoted to a very relevant and specific area of financial and economic activity – forecasting stochastic effects on the production processes of military goods of exporting countries.

The problems of analyzing time series in economics and production in recent years have attracted considerable attention. When building econometric models, risk managers most often use a standard indicator – return on assets. At the same time, the field of econometrics is experiencing various new opportunities, especially in the field of short-term forecasting, stochastic variability, and the availability of powerful specialized applications.

Time series analysis concerns the theory and practice of evaluating production capabilities over time. In a certain sense, it is an empirical discipline, but as in other scientific fields, the basis for drawing conclusions and making decisions is theory. However, the key feature that distinguishes the analysis of time series in the economy from other varieties of time series analysis. And the theoretical front-end, empirical time series contain a noticeable element of uncertainty.

When researching the time series of the characteristics of the production system, as a rule, various competing models are obtained, especially in production conditions with stochastic output associated with problems of bottlenecks. So, the choice of the best model that describes the production system becomes difficult and critical, since some models that most closely match the observed data may incorrectly predict future values due to the complexity and ambiguity of the model. In this work, the author seeks to demonstrate the procedure for selecting a model in a production system with stochasticity using the determination of the coefficient of determination, the Bayesian information criterion and the Akaike information criterion. The obtained results of estimating the production volume serve as initial data for calculating the functions of autocorrelation and multiple correlation and choosing the order of the corresponding autoregression models (AR models), ARMI, and AR models with integrated moving average (ARIMA). The model parameters are evaluated, used for forecasts and compared with the original and converted data to obtain the sum of squared errors (SSE). Model adequacy assessment is usually carried out using the Bayesian Information Criterion (BIC) and the Akaike Information Criteria (AIC). Among competing models, the ARIMA model better explains the variance of data sets and has low BIC and AIC values. Therefore, it is most often chosen as a model that represents the studied production system [1].

It is also found that the adjusted determination coefficient in combination with the BIC and AIC criteria is an adequate tool for model selection in the study of time series, especially in the presence of stochastic effects [2].

In practical terms, a very attractive feature of ARIMA-type models is their acceptable accuracy in predicting and in the absence of unlimited disagreement of the extrapolating function (extrapolants). This feature is due to the presence of the asymptotic properties of the extrapolant as such, which is finely rational by definition, while no polynomial has either horizontal or any other asymptote [3].

The importance of calculating quantitative measures of forecasting accuracy is well covered in the literature [4, 5]. But specific recommendations on the quantitative distinction of qualitative forecasts from unsuccessful, as a rule, are absent. For this, standard classical measures of forecasting errors are usually used, such as:

– mean absolute deviation (MAD);

– mean square error (MSE) or average (percentage) absolute error (mean Absolute Percentage Error – MAPE).

For such measures, lower values indicate better forecasting models. However, these measures may not always ensure the accuracy of forecast models in practical applications. This leads to the fact that users can't understand the consequences of forecasts for their activity [5]. In this paper, a simple and practical indicator of the accuracy of forecasting the model is proposed – the percentage error of the forecast (Percent Forecast Error – PFE).

Researchers in the thematic field for decades have proposed various methods for choosing a model, for example, in [6] it is argued that the performance of a model is a function of its intended ability, and its selection is extremely necessary because it controls the quality choice of the selected model. Improving the performance of the model obtained by choosing a model provides a reliable forecast of the future system [7]. In [8], it is noted that in addition to testing the adequacy of the model, the goal of model selection includes the search for a good forecast algorithm that describes the system, and the AIC criterion is the main method of model selection [9]. In [9], a regularized information criterion (RIC) is proposed for the Kullback-Leibler divergence, which is an extension of BIC and AIC criteria, and then used to select a model. In [10], various methods are used to select a model, including hypothesis testing, diagnostic tests, correspondence methods, Bayesian approaches, and forecast estimation methods. In [11], the SURE-Autometrtcs model selection algorithm is used; it is claimed that the method worked well. In [12], hypothesis tests and selection criteria were proposed using the final prediction error (FPE) for model selection.

Thus, the various methods of model selection procedures are presented above, but they can be applied in rather narrow, specific situations, different from the industry studied here. This also indicates that there is no single method for choosing a model, and some procedures recommended in the literature are quite complex, time-consuming and rather abstract, which narrows the scope of their practical application. So, *the object of research* is heteroskedastic processes that affect the production of military goods of exporting countries. *The aim of this research* is development of an effective methodology for choosing the order of regression models used in the practical forecasting of processes.

2. Methods of research

The key variables necessary to select a model using a consistent determination coefficient R^2 are the number of model parameters and the sum of the squared errors. The results of measuring the volume of output obtained from the manufacturing organization serve as input to the autocorrelation function and the calculation of the partial autocorrelation function. The values of the parameters are used to predict and compare them with the original and converted data to obtain the sum of the squares of the errors.

Let's consider the simplest linear regression equation:

$$y = ax + b,$$

where y – the dependent variable; x – the independent variable; a, b – least squares estimation coefficients.

The adequacy of the model is evaluated using the adjusted determination coefficient R^2, which gives an idea of how many data points fall into the regression line to study the relationship in the data set. In other words, this is a certain proportion of the variance σ_y^2 of the dependent variable y. This proportion appears in the total variance due to the influence of the independent variable x:

$$\sigma_{res}^2 = \sigma_{x|y}^2 + \sigma_y^2,$$

where $\sigma_{x|y}^2$ – the conditional (according to the influence of the variable x) variance of the dependent variable, that is, the variance of the random model error; σ_y^2 – the actual variance of the dependent variable y.

The adjusted coefficient of determination R^2 shows the proportion of the variation of the variable, which is explained by the influence of the independent variable on the dependent variable.

The equation for the adjusted determination coefficient is given by the expression:

$$\overline{R}^2 = 1 - \frac{\sum\limits_{k=1}^{n} \frac{\left(y_k - \overline{y}_k\right)^2}{\left(n-k+1\right)}}{\sum\limits_{k=1}^{n} \frac{\left(y_k - \overline{y}_k\right)^2}{\left(n-k+1\right)}}, \ \overline{y}_k = \frac{1}{n}\sum_{k=1}^{n} y_k,$$

where k – the number of parameters; n – the number of independent variables; $SSE=\sum_{k=1}^{n}\left(y_k-\overline{y}_k\right)^2$ – the sum of the squares of the regression residues; y_k, $\overline{y}_k$ – actual and estimated values of the explanatory variable.

The corresponding equations for the Akaike criteria and for the Bayesian criterion are:

$$AIC=m\ln\left(SSE\right)+2k;$$

$$BIC=m\ln\left(SSE\right)+k\ln m,$$

where m – the number of observations in the series.

Let's show that at least one factor (randomness) affects the accuracy and can explain the differences in the relative effectiveness of different methods. Moreover, it is assumed that the accuracy of the forecasting method depends on several factors, and these factors can be identified and quantified, as well as measured their impact. Perhaps when high randomness is present in a series of data, more advanced methods, such as the ARIMA, can overwhelm the model with this data. This allowance may occur when the mean square error is minimized, that is, when stationarity is achieved by either differentiating for the season, or when the selected ARIMA model (*p*, *q*). For example, the absence of random events in residuals does not always mean better forecasting results.

Finally, the difference between model matching and forecasting, as well as the type of loss function, should be noted. For example, when there is an unlimited standardized function of the quadratic losses, the seasonal analysis methods give the same results as the methods using data, are adjusted for the season. However, for forecasting this is not so. Best practices vary depending on the accepted loss function and the number of random variables present in the series.

A decision maker using these series and using a single forecasting method would get very different results depending on which loss function would be minimized and it would like to minimize errors in fitting the model or in the forecasting phase. However, in general, this can also be done using simpler methods. Here, an important role is played by smoothing the experimental data by introducing some finite weight function. Let's consider the problem of optimizing such a smoothing function.

3. Research results and discussion

One of the most universal methods of the constructed analytical models of a random function is the canonical decompositions of V. S. Pugachev [13]. Let's consider the canonical decomposition technique for sampling a fixed volume.

In a real situation, in the collection and processing system, a priori data on the statistical characteristics of the process, as a rule, are either completely absent or only partially, of a very general nature. Therefore, when constructing a decomposition algorithm for sampling a variable (increasing) volume, it is necessary to simultaneously evaluate the necessary characteristics, taking into account the newly obtained data. Such characteristics include expectation, variance, the correlation function of the process, the distribution density of the expansion coefficients V_v.

Taking into account the continuity property of random variables V_v, which follows from the continuity of the bypass random function $X(t)$, one can apply the non-parametric Parzen estimation [14] of the form:

$$F_N(V)=\frac{1}{N\,d_f}\sum_{k=1}^{N}g\left(u_k\right),$$

where d_f – a constant called the blur coefficient; $g\left(u_k\right)$ – weight smoothing function or core function; $u_k=\left(V-V_k\right)/d_f$, V_k – k-th implementation of a random variable V.

Having a set of realizations $V_k,\ k=\overline{1,N}$ and asking in any way d_f and $g(u)$, it is possible to uniquely determine the probability density of a random variable V.

A practical technique for choosing a smoothing core and blur coefficients is proposed in [15]. It is shown that with the symmetry of the core $g(u)=g(-u)$, the structure of the form:

$$g(u)=\begin{cases}a-bu^2, & |u|\le c,\\ 0, & |u|>c,\end{cases}$$

is optimal by the criterion of minimum integral mean square error (MSE) approximation. Here a, b, c are some constants that are selected based on the characteristics of the problem being solved. It is known that the choice of function smooths the cores, comes to the edges, almost always gives the result that is closest to the optimal one. At the same time, for this task, deviations of the shape of the nucleus from the above are not very critical. For example, when using a simple – rectangular function of the core, the MSE approximation increases by only 6 %. This significantly reduces the complexity of the calculations that must be performed in real time.

Here are some considerations for choosing a coefficient d_f – a parameter that determines the interval of nonzero values of the core. If select it too large, the estimate will be too smooth, insensitive to rapid deviations of a random variable. If the value d_f is too small, the estimate will not be smoothed out enough, it will be «noisy». The optimal between these extreme positions for a rectangular core function is the choice:

$$d_f=0.5\sup_k\left|V_k-V_{k-1}\right|,\ V_k\ge V_{k-1},\ k=\overline{2,N},$$

half the largest distance between two adjacent members of a random sequence. Moreover, the coefficient d_f obviously depends on the parameters of the sample, which guarantees the absence of gaps in the domain of definition of the estimate, that is, the absence of noise, and the minimum «smoothing» of the estimate, not on average for the entire set of samples, but for each specific sample.

Thus, when using the rectangular smoothing function and the above rule for calculating the coefficients, the expression for the estimate has the form:

$$F_N(V)=\frac{1}{N}\sum_{k=1}^{N}g_u(V),$$

where

$$g_u(V)=\begin{cases}1/2d_f, & V_k-d_f\le V\le V_k+d_f,\\ 0, & \left|V_k-V\right|>d_f,\end{cases}\quad k=\overline{1,N}.$$

In the practical implementation of the technique in question on a computer with limited speed and memory capacity, accounting for the duration of the aftereffect of real processes is quite an important task. An effective estimate of the duration of the aftereffect is the normalized correlation coefficient of the process, expressed in terms of the coordinate functions of the canonical decomposition:

$$r_{\nu x}(\nu,k)=\frac{\sigma_\nu\varphi_\nu(k)}{\sigma_x(k)},\ \nu=\overline{1,N},\ k=\overline{1,N}.$$

It is shown in [13] that a function of some special form of a random variable $r_{\nu x}(\nu,k)$ can be considered distributed according to the Gaussian law with the corresponding mathematical expectation and dispersion. The area of acceptance of the hypothesis that the correlation coefficient is equal to zero is also defined there. Thus, using the test of this hypothesis against a simple alternative about the inequality of the correlation coefficient to zero in parallel with the general data processing algorithm, it is possible at each stage of calculating the decomposition parameters to determine the necessary amount of information stored.

4. Conclusions

A methodology for choosing a model in a production system with stochasticity is proposed in this paper. The technique is based on the canonical layouts of a random function. To partition the distribution function into non-equidistant intervals with constant flow intensities, an economic recurrence algorithm is used that is easily implemented on a computer.

The calculation results can be used to optimally select the order of the regression model, which approximates the real production process – a time series with random external influences.

References

1. McQuarrie, A., Tsai, C.-L. (1998). *Regression and Time Series Model Selection*. World Scientific Publishing Co. Pte. Ltd, 455.
2. Tsay, R. S. (2010). *Analysis of Financial Time Series*. Hoboken: John Wiley & Sons, Inc., 677. doi: http://doi.org/10.1002/9780470644560
3. Kendall, M. (1976). *Time-series*. London: Charles Griffin, 197.
4. Lewis, C. D. (1982). *Industrial and business forecasting methods: a practical guide to exponential smoothing and curve fitting*. London, Boston: Butterworth Scientific, 143.
5. Klimberg, R., Ratick, S. (2018). Development of a Practical and Effective Forecasting Performance Measure. *Advances in Business and Management Forecasting. Vol. 12*. Emerald Publishing Ltd, 103–118. doi: http://doi.org/10.1108/s1477-407020170000012007
6. Hastie, T., Tibshirani, R., Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference and Prediction*. Springer Series in Statistics. New York: Springer-Verlag, 764.
7. Ferreira, E. (2015). Model Selection in Time Series Machine Learning Applications. *Academic dissertation of Technology and Natural Sciences of the University of Oulu*. Linnanmaa, 87.
8. Rao, C. R. (Ed.) (2013). *Handbook of Statistics. Vol. 30*. Time Series Analysis: Methods and Applications. Kidlington, Oxford: Elsevier, The Boulevard, 777.
9. Hurvich, C. M., Tsai, C.-L. (1989). Regression and time series model selection in small samples. *Biometrika, 76 (2),* 297–307. doi: http://doi.org/10.1093/biomet/76.2.297
10. Hannan, E. J., Krishanaiah, P. R., Rao, M. M. (1985). Various Model Selection Techniques in Time Series Analysis. *Handbook of Statistics. Vol. 5*. Elsevier B.V., 179–187. doi: http://doi.org/10.1016/s0169-7161(85)05008-8
11. Norhayati, Y. (2016). *SURE-Autometrtcs Algorithm for Model Selection in Multiple Equations*. Utara, 94.
12. Liebscher, E. (2012). A Universal Selection Method in Linear Regression Models. *Open Journal of Statistics, 2 (2),* 153–162. doi: http://doi.org/10.4236/ojs.2012.22017
13. Pugachev, V. S. (1962). *Teoriia sluchainykh funkcii*. Moscow: Fizmatgiz, 784.
14. Parzen, E. (1962). On Estimation of a Probability Density Function and Mode. *The Annals of Mathematical Statistics, 33 (3),* 1065–1076. doi: http://doi.org/10.1214/aoms/1177704472
15. Epanechnikov, V. A. (1969). Neparametricheskaia ocenka mnogomernoi plotnosti veroiatnosti. *Teoriia veroiatnostei i ee primeneniia, 1,* 156–161.

Kubiv Stepan, *PhD, Associate Professor, First Vice Prime Minister of Ukraine – Minister of Economic Development and Trade of Ukraine, Cabinet of Ministers of Ukraine, Kyiv, Ukraine, ORCID: http://orcid.org/0000-0002-1110-2024, e-mail: sikubiv@ukr.net*