EURASIAN JOURNAL OF LAW, FINANCE AND APPLIED SCIENCES
Innovative Academy Research Support Center
UIF = 8.3 | SJIF = 5.961 www.in-academy.uz

AN ECONOMETRIC FORECASTING OF DEMAND AND THE ALGORITHM FOR CONSTRUCTION OF LINEAR MODEL BASED ON PANEL DATA

Oripova Gulirano Nodirovna
Undergraduate of master degree in TUIT named afer M. al-Khwarezmi
Tel: +998(97) 718 12 08
grahmatova1996@mail.ru
https://doi.org/10.5281/zenodo.7214391

ARTICLE INFO

Received: 27th September 2022
Accepted: 01st October 2022
Online: 17th October 2022

KEY WORDS

econometric forecasting, retail, customer demand for goods, panel data, statistical methods, multilinear regression, SKU(Stock Keeping Unit)

ABSTRACT

The role of demand forecasting in the automation of retail processes is considered. A description of possible solutions to the problem of forecasting in the form of a small task of regression analysis is given. This article discusses the selection and stages of data selection and development process, expressing a comparative analysis in adopting a simple model of retail forecasting with a basic mathematical model and its results available.

Introduction

Today, activity in any field of economics (management, finance and credit, marketing, accounting, auditing) requires a specialist to know modern methods of work, the achievements of world economic thinking, to understand the scientific language. Most of the new methods are based on econometric models, concepts and techniques. The language of economics is becoming more and more a mathematical language, and economics is increasingly being called one of the most mathematical sciences. The main results of economic theory are qualitative in nature, and econometrics incorporates them empirically. It is not possible to make any reliable forecast without econometric methods.

In economics, it is natural to distinguish three types of scientific and practical activities as a discipline at the intersection of economics (including management) and statistical analysis (according to the degree of specificity of the methods associated with immersion in certain problems):

a) development and research of econometric methods (applied statistical methods) taking into account the specific features of economic data;

b) development and study of econometric models in accordance with the specific needs of economic science and practice;

c) use of econometric methods and models for statistical analysis of accurate economic data.

Currently, statistical data processing is, as a rule, carried out using panel data and related software products. Panel data, or longitudinal data, are multidimensional data used in the social sciences and econometrics obtained as a result of a series of measurements or observations over a period of time for the same companies or people. The stage of development of the Uzbek economy, world management and information technology is focused on the simple calculation of the vector, the collection of economic data, etc., and on this basis the development of complex computational processes, automation of retail business using arithmetic operations.

It is difficult to list all the tasks that need to be solved:

• creation of referral services, in which the product is automatically offered to a potential buyer;

• creation of automatic plans and schedules, various groups / positions (solving commercial / merchandizer issues);

• product screen optimization: frequency, volume of calculations and time spent on this work by responsible managers;

• forecasting the demand for each product group and a specific brand position, etc.

The purpose of these tasks is to automate retail business processes, i.e. to reduce the cost of performing similar work by hired professionals. It is not only about maintaining the quality of work, but also about improving it because of the limited capacity of the so-called “human factor”.

This article considers one of the main tasks in the field of modern retail business, as well as in the field of "data mining" and econometric modeling, that is, forecasting the demand for a particular brand position. Forecasting the demand for goods on SKU(Stock Keeping Unit) is to improve the management of inventory goods, optimization of wages and the development of tactical plans for the movement of goods. Thus, the demand forecast is the basis for further development of retail trade, the quality of which determines the efficiency of business processes of the company.

Qualitative solution of the problem of forecasting requires not only a large amount of collected data, but also specialized knowledge in the field of mathematical modeling and statistical data processing, the use of specialized software systems and programming languages. Commodity demand forecasting is usually taken as a quantitative response forecast taking into account the nature of the problem and the data under consideration[1]. At present, there are a large number of mathematical methods and models that solve this problem on their own, including regression analysis methods - various linear regression, solution tree, neural networks, etc.; as well as methods for analyzing temporal series - exponential alignment models, various models and combinations of autoregression. In general, simply using one method will not give the desired result. Successful prediction requires the following:

1. first, to prepare preliminary data that reflect the logic of the processes that take place during the sale of a particular product;
2. second, to test certain methods and determine the best ones in terms of functional quality;
3. third, create an ensemble of models[2] or identify mechanisms for specific flexible models to reduce the risk of errors increasing over time[3].

1. Selecting input data and pre-processing them

According to research, the main external and internal factors affecting brand demand are as follows:

• internal dynamics (characteristics) of product sales;

• active customer flow at the point of sale;

• availability of calendar holidays and other calendar effects (days of the week).

Sampling is done in retail stores for all products in a particular group of a

[1] Баринова О.В., Вальков А.С., Воронцов К.В., Громов С.А., Ефимов А.Н., Чехович Ю.В. Система прогнози-
рования потребительского спроса Goods4Cast. Вычислительный центр им. А.А. Дородницына РАН.
М., 2015.

[2] Business Data Analytics: Ансамбли моделей. URL: http://businessdataanalytics.ru/ModelEnsembles.htm.

[3] Лукашин Ю.П. Адаптивные методы краткосрочного прогнозирования временных рядов. М.: Финансы и статистика, 2003. 416 с.

EURASIAN JOURNAL OF LAW, FINANCE AND APPLIED SCIENCES
Innovative Academy Research Support Center
UIF = 8.3 | SJIF = 5.961 www.in-academy.uz

particular retail store. The form of data used in the analysis is in the form of "panel data", indicating the data label (code), the presence of a target variable in the form of sales of a particular product, and another link parameters are provided. The sample will take at least 2.5 years to sell; The data used to produce accurate results cover the period from 01.10.2019 to 30.09.2020. It should be noted that this period is determined by the presence of an annual season, which should be noted in the data for a better forecast.

Before describing the data logic, it is necessary to determine the initial assumptions about the data, as well as the general principles on which they are based:

• initial assumptions about the structure and content of the data are based on research from the following sources;

• according to the type, variables are divided into quantitative, nominal and orderly types in the standard form.

During the analysis, the nature of the initial parameters changed based on certain assumptions - locking of the base model and quality criteria. The basic model in determining the specifications and properties of data and their interrelationships is a general view of multifactor regression:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_m x_m + \varepsilon,$$

where m is the number of variables under consideration, and y is the target (dependent) variable, $x_1, x_2, \ldots, x_m$- independent variables and $\beta_1, \beta_2, \ldots, \beta_m$ - coefficients of arbitrary variables calculated by the least squares method, β_0- the free coefficient of the model, ε - random error of the model.

The changes in the specification under consideration are based on the heuristic method of finding the smallest value of the quality index(*Medium square error*) of the model, as well as the standard deviation:

$$MSE = \frac{1}{n} \times \sum_{i}^{n} (y_i - \hat{y}_i)^2$$

where y_i- the actual value of the sale (demand), $\hat{y}_i$- this is the approximate value of the estimated value, n- is the number of elements in the sample. *MSE* is strictly calculated based on the test results.

• training and test samples will be formed naturally, strictly in 2 time intervals, in the ratio of 80% and 20%, respectively. Researchers make primary assumptions; the composition and type of data, the MSE coefficient in the test sample for the given data set of model coefficients are evaluated. The base model MSE ratio is analogously compared to the values obtained by the forecasting algorithm available in retail activities.:

$$y_t = 0{,}4 \times y_{t-7} + 0{,}3 \times y_{t-14} + 0{,}2 \times y_{t-21} + 0{,}1 \times y_{t-28},$$

where y_{t-7} , y_{t-14} , y_{t-21}, y_{t-28}- the values of demand for the product for 7, 14, 21 and 28 days, respectively;

0.4, 0.3, 0.2 va 0.1 - model coefficients to compare previous values in terms of increasing the impact of new data on the old. The method is basically a moving algorithm of average weight.

The process of adding and changing variables is carried out until the following case:

$$MSE_{im,} < MSE_{MA},$$

where $MSE_{im,}$- the mean square error in the test sample of the base applied in the base linear regression, MSE_{MA}- this is the average square error for the algorithm that exists and is used in the enterprise.

The task of the researcher is to form a variable that most fully reflects the processes being analyzed and to develop a basic model that has a higher quality category than the current model.

This procedure can be demonstrated using the block diagram in the table below:

EURASIAN JOURNAL OF LAW, FINANCE AND APPLIED SCIENCES
Innovative Academy Research Support Center
UIF = 8.3 | SJIF = 5.961 www.in-academy.uz

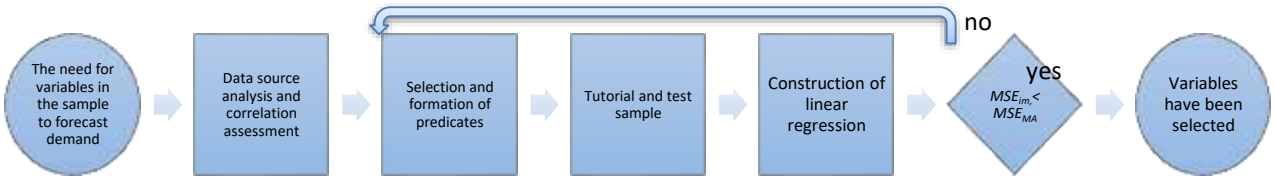

Figure 1. Heuristic search and algorithm for forming an independent variable

The following is a description of all the parameters used, the order of processing and the logic of their use in creating a demand forecasting model.

Code (identifier) and product name. The unit of commodity nomenclature is defined by the concepts of Stock Keeping Unit (SKU) or product identifier. Each SKU has a unique code and a unique name. The sample contains information on the single commodity position No. 691, its specific sales history and turnover period. Finding the right relationship between goods improves the quality of the forecast.

Date (period). The appropriate unit of time (period) to solve the problem is a day. Accordingly, the selected samples were developed for each day from 01.10.2019 to 30.09.2020.

The choice of daily parts as a unit of time (period) depends conditionally on the issue of stock replenishment and inventory management. In practice, most suppliers to retailers have a package schedule of goods no more than 1 time per day. Thus, it makes no sense to formulate a detailed review in hourly periods, as this will result in additional computational costs. It should be noted that the data of this sample are divided into training and testing sections as follows:

• educational data: from 01.10.2019 to 01.02.2020, ie about 79% of this sample;

• test data: from 01.02.2020 to 30.09.2020, which is 21% of the sample.

One of the important features of dividing the samples into periods as above is that our problem is related to the forecast values of demand for future periods. Accordingly, the prognosis may be affected by nonstationality, and such an effect can only be recorded with a type of division similar to the study and test samples.

Days of the week, calendar holidays and other features of the time. Additional variables to the parameters used in demand modeling include factors such as the days of the changing week, the availability of calendar and other holidays (e.g., February 14) and the number of days before / after the holiday, the number of days in the holiday period, and the number of days in the year. `variables are included. This calendar allows you to record important events that affect customer demand data.

Air temperature condition. Depending on the weather conditions, the demand for goods in the consumer market may change. For example, sales of ice cream in hot weather are likely to increase compared to cold weather. Therefore, the average temperature variable of the point of sale location is included in the data structure.

Number of checks in the store. The variable is included in the possible group of goods as the main measurement factor of the demand scale. Clearly, with the increase in the flow of consumers in the general situation, the demand for products will increase due to the fact that customers in the store buy different goods (in detail [7]).

Product balance at the beginning and end of the day. Variables that indicate the amount of goods available at the time of store opening and closing. These data are not directly relevant in modeling the target variable. Nevertheless, this creates additional conditions for quality, which will be discussed below.

Sale of goods. This is the primary variable in the modeling of demand variables for goods. In order to determine

EURASIAN JOURNAL OF LAW, FINANCE AND APPLIED SCIENCES
Innovative Academy Research Support Center
UIF = 8.3 | SJIF = 5.961 www.in-academy.uz

the value of direct demand in the trading process, the following conditions must be met:

$$\gamma_i = \begin{cases} NA, if\ b_i \le 0 \\ NA, if\ e_i \le 0, \\ then\ s_i \end{cases}$$

where $\gamma_i -$ the value of consumer demand for the commodity, $b_i -$ the value of the volume of the balance of goods at the beginning of the day, taken from the database of the enterprise, $e_i -$ the value of the residual volume of goods at the end of the day, taken from the database of the enterprise, $s_i -$ the value of sales of goods, the value of the non-existent target variable NA [3].

Assigning non-existent values to a target variable indicates that non-existent values can be replaced by some approximate algorithm or removed from the data source set together with a non-existent variable. Because of the complexity of replacing a variable for the target variable, it was decided to exclude the data. A representation of the resulting demand variable γ is shown in the following histogram:

Required volume histogram

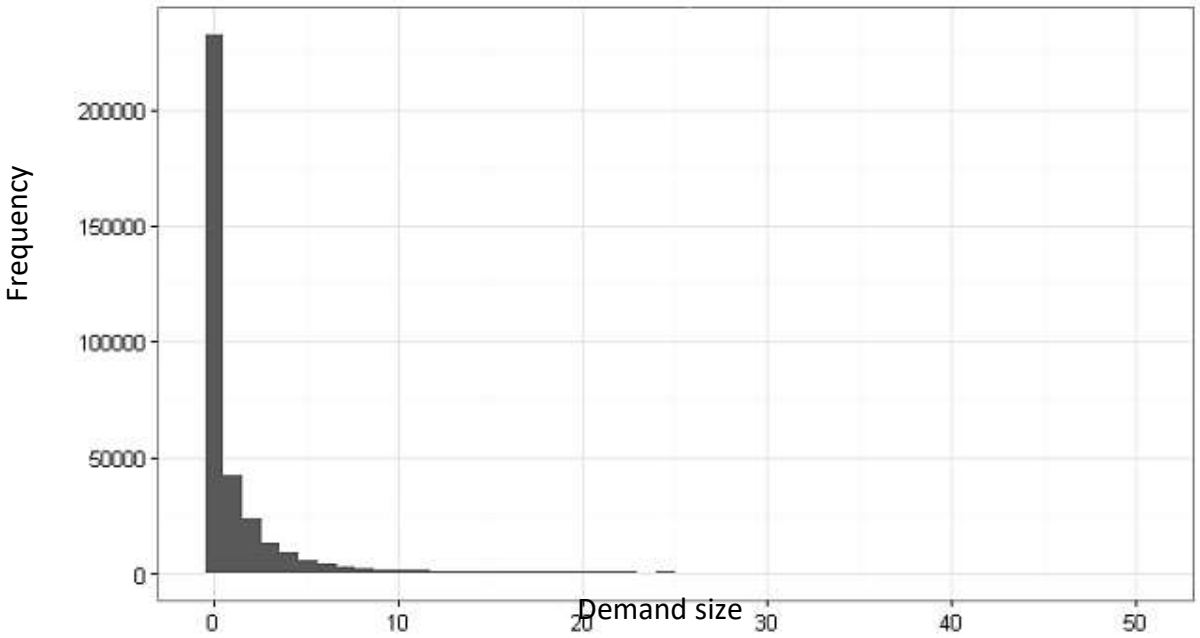

Figure 2. Demand volume distribution

The concentration of the above value is about 0, which confirms the fact that MSE quality is functionally accepted. This allows you to better take into account the quality of the trade forecast on the right side of the residual distribution of $(Y - \hat{Y})^2$.

Previous values of demand (lagging demand). An important part of the model is to include the previous value of the demand in the list of independent (independent) variables. The influence of previous periods on the sale of consumer goods is clearly visible, which means the presence of autocorrelation [1; 7]. However, the presence of lags included in the model is the last feature of the model in terms of quality.

As an example, we give a graph of the specific autocorrelation function of one of the goods in the sample (selected by SKU):

EURASIAN JOURNAL OF LAW, FINANCE AND APPLIED SCIENCES
Innovative Academy Research Support Center
UIF = 8.3 | SJIF = 5.961 www.in-academy.uz

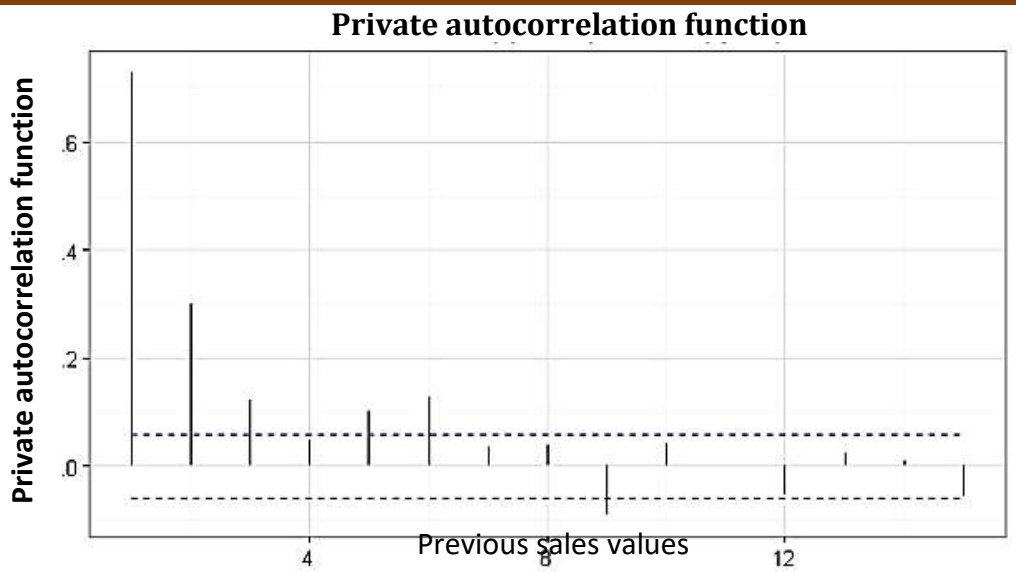

Figure 3. Private autocorrelation function of product sales

The fading structure of autocorrelation effects is specific to the sale of consumer goods, so it makes sense to use the first seven lags when simulating variables. When the autoregression components are introduced on a private basis, the general shape of the linear model looks like this:

$$\boldsymbol{y = \beta_0 + (\alpha_1 y_{t-1} + \alpha_2 y_{t-2} + \cdots + \alpha_k y_{t-k}) + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_m x_m + \varepsilon},$$

where $y -$ target (dependent) variable, $y_{t-1}, y_{t-2}, \ldots, y_{t-k} -$ lagging values of the sequence, $\alpha_1, \alpha_2, \ldots, \alpha_k -$ autoregression coefficients of the model, m- the number of independent variables, independent variables and, ..., the coefficients of the dependent variables, $x_1, x_2, \ldots, x_m$- independent variables and $\beta_1, \beta_2, \ldots, \beta_m$ - coefficients of arbitrary variables calculated by the least squares method, β_0- the free coefficient of the model, ε - random error of the model. The composition of the entered parameters may vary depending on the maximum increase in model quality.

Product price indicators. One of the main features of the product is its price.

The value of goods is measured by the elasticity of demand, the perception by the buyer as a feature of the product and other factors that affect the level of demand for consumption. At the same time the price of the goods (in the example - the unit of time of day), as well as the price discount is included if the store announced a discount with a decrease in price during the sale. The discount rate was calculated as follows:

$$l = \left| \frac{C_b - C_p}{C_b} \right|$$

where l- discount rate, discounted price of goods in the previous (base) period, product price in the promoactive period.

To set the number of clusters, the value of the MSE linear model in the test set is used. Depending on the characteristics of the cluster centers obtained, a meaningful analysis of the obtained groups is possible.

2. Basic linear model: use in data generation

In the algorithm stage in Figure 1, after determining all the variables and their sizes, the coefficients of multi-line regression with the additional effect are evaluated. Table 1 shows the estimates of the partial model coefficients (there are 60 variables in the total model):

Table 1.

Coefficients of the linear regression model

Indicators	Evaluation of the coefficient	Standard error	t-value	p-value
β_0 coefficient	-1,467	0,175	–8,369	0,000
Delay of demand 1	0,371	0,002	208,677	0,000
…	…	…	…	…
Delay of demand 7	0,344	0,002	195,087	0,000
Product cluster 2	0,127	0,047	2,671	0,008
Product cluster 3	0,114	0,051	2,254	0,024
Product cluster 4	0,158	0,084	1,880	0,060
…	…	…	…	…
Availability of stock	3,566	0,105	34,061	0,000
Existing stock delay 1	–2,638	0,058	–45,763	0,000
Discount rate	5,068	0,382	13,255	0,000
Average temperature	0,001	0,001	1,957	0,050
Number of days in a year	0,000	0,000	0,651	0,515
…	…	…	…	…
Size (weight)	–0,139	0,019	–7,250	0,000
Number of checks	0,000	0,000	18,742	0,000
Tuesday	0,265	0,028	9,632	0,000
…	…	…	…	…
Sunday	0,239	0,029	8,357	0,000
Easter	–0,384	0,172	–2,227	0,026
14 february	–0,319	0,181	–1,769	0,077
23 february	–0,012	0,149	–0,081	0,936
…	…	…	…	…
Weekends	–0,348	0,149	–2,334	0,020
The number of SKUs that depend on x in terms of price	–0,008	0,001	–11,765	0,000
The number of SKUs depending on x at the stock price	–0,015	0,004	–4,309	0,000

R^2= 0.7549 is explained in the model, MSE_{im} equals to10.64 in the test sample. According to the obtained model, it is clear that most of the coefficients are significant enough (p-value is equal to 0). The initial assumption of this coefficient is confirmed by the indicator "Number of interchangeable SKUs by price" (total and stock), which are negative and valuable, so the demand for a particular product decreases with increasing number of SKUs .

It can be seen that without any nonlinear variation, the MSE of the multi-line regression cannot perform the task of obtaining the best model on the quality criterion (based on the algorithm in Figure 1): $MSE_{im,}<MSE_{MA}$, where MSE_{MA} = 10,55. A further modification of the linear regression formula should be performed to improve the result within the model. During the changes, the algorithm can be repeated as in Figure 1.

The final model is explained by the following formula:

$$y = \beta_0 + (\alpha_1 y_{t-1} + \alpha_2 y_{t-2} + \cdots + \alpha_k y_{t-k}) \times (\beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_{m-l} x_{m-l}) + \cdots + \beta_m x_m + \varepsilon,$$

where $y -$ target (dependent) variable, $y_{t-1}, y_{t-2}, \ldots, y_{t-k} -$ lagging values of the sequence, $\alpha_1, \alpha_2, \ldots, \alpha_k -$ autoregression coefficients of the model, m- the number of independent variables, independent variables and, ..., the coefficients of the dependent variables, x_1,x_2, ...,x_m- independent variables and β_1, β_2 ,..., β_m - coefficients of arbitrary variables calculated by the least squares method, β_0- the free coefficient of the model, ε - random error of the model.

We can see that the specification of the model has changed slightly - there is a multiplicative relationship between autoregression functional and several variables from the original set. It should be noted that 194 coefficients were estimated during the simulation, which is 3.23 times the number of variables in the first version of the regression model. Nevertheless, the addition of a nonlinear relationship between the variables allowed the following result to be achieved: MSE_{im} = 9.09, which is less than MSE_{MA}.

In this model R^2 = 0,8038, MSE_{im} = 9.09. Therefore, the selection of source variables and their modified variants according to the given algorithm is completed.

Conclusion

The application of the heuristic algorithm, as shown in Figure 1, allows the researcher to achieve the following basic result:

1) identification of key variables derived from the initial data of the information system, as well as radically new modified data that allows to take into account the specific features of decision-making by the buyer;

2) to create a basis for an alternative to the current method, based on a moving average value, which does not have sufficient prognostic ability.

However, it should be noted that the algorithm in question should consider a number of issues to improve the quality of the task of forecasting the demand for goods:

• Like many other methods, it requires prior analysis of the following variables: economic, statistical, and so on. It has not been analyzed in detail in the context of this article, but its methodology is presented in the sources below [3; 5-7];

• The simulation result is a linear model with a large number of coefficients. The next step in solving the problem of demand prediction is the use of more advanced regression analysis models, in which the explanatory capacity of the model decreases, but its accuracy increases. Compositional models should also be considered;

• Although there is a “soft” condition for the accuracy of the algorithm (Figure 1), conceptual changes in the data and enrichment of the original sample are possible; advanced methods of mass extension - support vector machine, it is proposed to use as a technological basis for subsequent modeling to work on random forest and artificial neural networks.

References:

1. Баль А.В., Логиновский О.В. Автоматизированный заказ высокооборачиваемых товаров с низкими сроками годности с использованием почасовых продаж // Вестн. Южно-Уральского гос. ун-та. Сер. Компьютерные технологии, управление, радиоэлектроника. 2015. Т. 15, Вып. 1. С. 21-25.

2. Баринова О.В., Вальков А.С., Воронцов К.В., Громов С.А., Ефимов А.Н., Чехович Ю.В. Система прогнози-рования потребительского спроса Goods4Cast. Вычислительный центр им. А.А. Дородницына РАН. М., 2015.
3. Пивкин К.С. Алгоритм построения линейной модели на панельных данных как этап эконометрического прогнозирования товарного спроса. Автореферат. Т. 27, вып. 2. 2017.
4. Джалилова У.Т. Моделирование спроса на продовольственные товары на основе учета его особенностей // Вестн. Таджик. гос. ун-та права, бизнеса и политики. Серия гуманитарных наук. 2014. № 1 (57). С. 159-164.
5. Пивкин К.С. Корреляционный анализ факторов влияния на покупательский спрос розничного магазина как этап формирования модели прогнозирования и управления запасами // Вестн. Удм. ун-та. Сер. Экономика и право. 2017. Вып. 3. С. 40-50.
6. Крок Г.Г., Сысоева С.В. Большая книга директора магазина 2.0. Новые технологии. СПб.: Питер, 2016. 464 с.
7. Стрижов В.В., Крымова Е.А. Методы выбора регрессионных моделей. М.: ВЦ РАН, 2010. 60 с.