Choosing the Optimal Quantity of Factors for Prediction the Severity of Bronchial Asthma in Children Using Linear Regression Models

Oleh Pihnastyi [1], Olga Kozhyna [2] and Tetiana Kulik[2]

[1]National Technical University "Kharkiv Polytechnic Institute", 2 Kyrpychova, Kharkiv, 61002, Ukraine
[2]Kharkiv National Medical University, 4 Nauky Avenue, Kharkiv, 61022, Ukraine

Abstract

The severity of the course of bronchial asthma depends on many factors. Clinical and laboratory studies were carried out on 90 children aged 6 to 18. 70 children with bronchial asthma of various degrees of severity as well as 20 healthy school-aged children were included into the main group. 142 predictors were studied, 11 factors were selected from the bottom in accordance with the selection method. Multivariate linear regression models have been developed and analyzed to predict the severity of bronchial asthma disease. The dependence of the forecast quality of the observed value on the number of model regressors is analyzed. The MSE value was used as a characteristic of forecast quality. An estimate of the number of regressors required for a significant increase in the forecast quality is presented. The law of distribution of the error in predicting the severity of bronchial asthma disease in a multifactorial linear regression model has been substantiated. The visual representation of multivariate models is made using the residual plot.

Keywords 1

Bronchial asthma, child, severe asthma, prediction, MSE, regression model, residual plot

1. Introduction

Bronchial asthma is a severe heterogeneous chronic lung disease. Numerous studies have revealed that the prevalence of bronchial asthma does not depend on the level of wealth of the country and is 4 – 10 % among the adult population [1, 2].The severity of the course of bronchial asthma depends on a sufficiently large number of factors that, presumably, have the same effect on the clinical manifestations of the disease. The first clinical symptoms may appear already in early childhood, very often similar to the symptoms of other childhood diseases [3, 4]. The relationship between the early age at which the first manifestations of the disease appeared and the severity of the course in the patient's adult life has been proven [5].

Despite similar symptoms of bronchial asthma among patients, the result of treatment and further prognosis of the disease is very different. Investigations have confirmed the presence of various phenotypes of the disease and the influence of a huge number of factors on the occurrence of bronchial asthma and the peculiarities of its course [6, 7].

Currently, the tactics of treatment and observation of patients have been developed to increase the level of disease control [8], based on the use of stepwise therapy. However, there is a fairly large category of patients who are characterized by an uncontrolled or severe course of the disease [9], which confirms the presence of different pathogenetic mechanisms of occurrence of bronchial asthma [10, 11].

ITTAP’2021: 1nd International Workshop on Information Technologies: Theoretical and Applied Problems, November 16–18, 2021, Ternopil, Ukraine

pihnastyi@gmail.com (A. 1); olga.kozhyna.s@gmail.com (A. 2); tv.kulik@knmu.edu.ua (A. 3)
ORCID: 0000-0002-5424-9843 (A. 1); 0000-0002-4549-6105 (A. 2); 0000-0002-8842-892X (A. 3)

Studying not only the factors, but also determining their relationship with each other, is an important step in understanding the course of the disease in each individual case. Analysis of the multifactorial nature of bronchial asthma underlies the prediction of the disease and its course [12]. Numerous studies have examined various categories of factors. Commonly used factors include age, gender of the child, whistling breath, allergic sensitization, Ig E [13, 14].

To assess the prognosis of a severe course of bronchial asthma, both linear and nonlinear multifactorial models are used, containing a different number of regressors (Table 1), aimed at determining predictors that are unambiguously significant in determining the severity of the course of bronchial asthma disease [15].

In this case, the values of the regressors are determined by both quantitative and qualitative values. Table 1 presents data on some common types of models for predicting the severity of bronchial asthma and on the number of predictors in these models.

Table 1
The number of regressors in models for predicting the severity of bronchial asthma

Number of regressors	Linear	Logistic	Machine learning
2-3	[24]	[22]	-
4-7	-	[17, 21]	-
8-10	-	[16, 19, 20]	-
>10	-	-	[18, 23, 25]

It should be noted that the models presented in Table 1 with the same number of regressors are used to analyze the prediction of the severity of bronchial asthma by different initial factors.

2. Formulation of the problem

The presence of a large number of models with different numbers of regressors makes the issue of choosing both the type of model and the number of regressors in the model relevant. In this research, we analyze the dependence of the forecast quality on the number of model regressors.

The process of building a linear regression model with a large number of regressors is quite laborious. The computational complexity of the algorithm for constructing a regression model grows in proportion to the square of the number of regressors in the model. Therefore, when analyzing the severity of bronchial asthma, linear models are usually used with the number of regressors, the number of which does not exceed 5-7 (Table 1). Linear models with a small number of regressors can be considered as a tool for preliminary analysis of a set of experimental data. A deeper analysis requires an increase in the number of regressors in the model. Due to the fact that the results of predicting the severity of the course of bronchial asthma depend on a sufficiently large number of weakly dependent factors with approximately the same scale of formation of the explained value, an increase in the number of regressors in the model leads to a slight increase in the quality of the prediction of the observed value. On the other hand, the presence of a large number of weakly dependent factors with approximately the same scale of formation of the value of the explained quantity leads to the fact that linear regression models are a good tool for predicting the severity of the progression of bronchial asthma disease due to the fact that the distribution of the prediction error for the quantity of the regressors $K \geq 10$ satisfies the normal distribution law. However, the issue arises, how many regressors the model should contain and how much the prediction accuracy of the model will be estimated to increase with an increase in the number of regressors. This work is devoted to the analysis of this problem.

The regression model that allows you to determine the severity of the course of the bronchial asthma can be summarized as follows:

$$Y_i = F(X_1, X_2, \ldots, X_k) + \varepsilon_i \quad (1)$$

where X_m is the value of m-the regressor; Y_i is the numerical value of the characterizing the severity of the course of the disease of bronchial asthma; ε_i is an error in predicting the numerical value for the i – the test.

To analyze the influence of the number of regressors on the quality of predicting the severity of the disease of bronchial asthma, we will use the data set formed during the examination of 90 children with a diagnosis of bronchial asthma aged 6 to 18 years. The investigation contains data from the anamnesis of life and diseases of patients, laboratory and diagnostic indicators of the examination. The study was conducted with respect for human rights and in accordance with international ethical requirements; it doesn't violate any scientific ethical standards and standards of biomedical research. To analyze the dependence of the parameters on the quality of prediction, 142 factors were selected, which were encoded. As a result of the studies, for each examined patient, the values of 142 factors were recorded, on which, it is assumed, the severity of the course of the bronchial asthma disease may depend. As a result of preliminary analysis, invalid data were excluded from this set. The resulting dataset in the form of a 90x142 matrix [26] was used to build a regression model. As a result of the phased elimination, 11 factors were identified that match the criterion

$$r_{y\,x_m} \to \max , \qquad r_{x_m x_v} \to \min . \tag{2}$$

which are used in this work. Based on criterion (2), out of a total of 142 factors, those with a correlation $r_{y\,x_m}$ between the regressor and the observed value is the highest, and the correlation $r_{x_m x_v}$ between the regressors has the lowest value was selected. In other words, out of 142 factors, 11 factors were selected that have the largest correlation $r_{y\,x_m}$ values between the regressor and the observed value. Thus, it is assumed that the selected factors have the most important influence on the severity of bronchial asthma disease. The selected factors are tested for the condition $r_{x_m x_v} \to \min$, in order to exclude those factors that are highly correlated with each other. These factors are replaced by the following factors from the condition $r_{y\,x_m} \to \max$. As a result of several iterations, the factors were determined, the numerical characteristics are presented in Table 2. Each factor is characterized by mathematical expectation m_x, standard deviation σ_x and correlation coefficient with the observed value r_{yx}.

Table 2

Numerical characteristics of the factors selected to build models that determine the severity of the course of bronchial asthma disease

Code	Regressor name	m_x	σ_x	r_{yx}
X_1	Allergic rhinitis	0.4494	0.4974	0.3223
X_2	Atopic dermatitis	0.0562	0.2303	0.3767
X_3	Number of years from the first symptoms	5.5281	4.4396	0.3023
X_4	Bronchial asthma in father	0.0864	0.281	0.0309
X_5	Bronchial asthma in relatives of second generation	0.0658	0.2479	0.4157
X_6	Eosinophils %	3.913	3.4462	0.2646
X_7	Domestic dust	2.2319	1.1312	0.3116
X_8	Pillow feather	0.7536	0.8059	0.3681
X_9	Rabbit hair	0.5652	0.8925	0.2236
X_{10}	Sheep wool	0.5217	0.6507	0.3373
X_{11}	CD25 10*3 cells	0.6937	0.3087	0.2198

The selected factors are used in this work to construct linear regression models for predicting the severity of bronchial asthma disease. The type of model is determined by criteria (2). The criterion $r_{x_m x_v} \to \min$ indicates that the factors presented in Table 2 are weakly dependent on each other.

Due to the fact that to assess the severity of the course of bronchial asthma disease, a large number of weakly dependent factors with approximately the same resulting contribution of the predicted observable value, proportional r_{yx}, were selected, choice for prediction a linear model for prediction

suggests that the error ε has a normal distribution with distribution characteristics:

$$E(\varepsilon_i) = 0 , \quad \sigma^2(\varepsilon_i) = \sigma^2 , \quad \sigma^2(\varepsilon_i, \varepsilon_j) = 0 , \, j \neq i . \tag{3}$$

This feature, characteristic of models for predicting the severity of bronchial asthma disease, will be used to compare the prediction accuracy of linear models with different numbers of regressors.

It should also be added that the spread of error values ε_i for each range of predictor values X_m obeys a probability distribution with mathematical expectation $E(\varepsilon_i) = 0$ and standard deviation $\sigma(\varepsilon_i) = \sigma$.

3. Research methodology

The first step of the study after the choice of factors (Table 2) is to build a set of linear regression models for predicting the severity of bronchial asthma disease and comparing the quality of the prediction of the observed value for a different number of model regressors. As a criterion for comparing models, we will use the MSE value

$$MSE = \frac{1}{n}\sum_{i=1}^{n} \varepsilon_i^{\,2} . \tag{4}$$

For comparison, consider 1, 2, 3, 5 and 10 factor linear regression models, for the construction of which we used the factors from Table 2.

3.1. Construction and analysis of 10-factor linear regression models

For the eleven factors presented in Table 2, we construct 11 models with ten factors (Table 3).

Table 3
Coefficients for a ten-factor linear regression model.

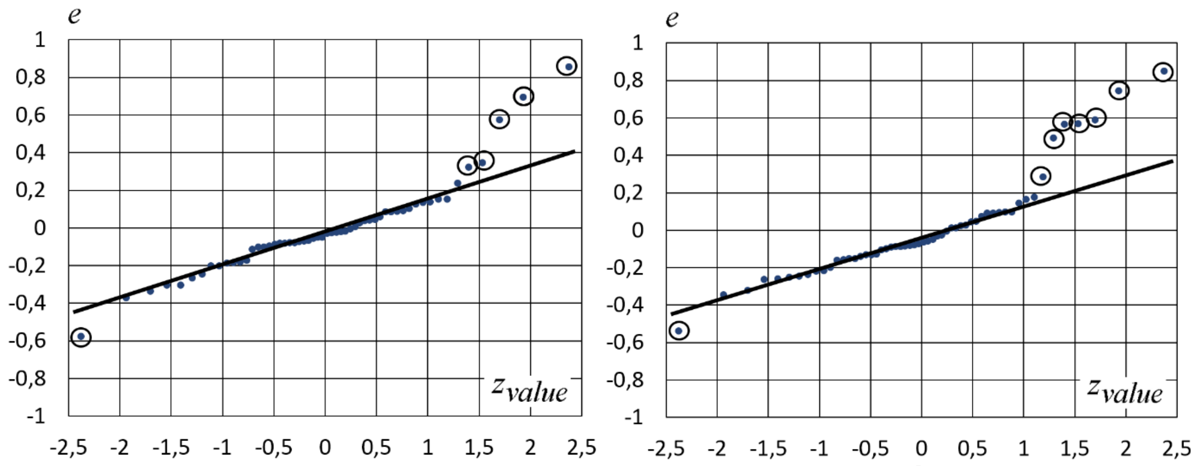
№ model	Number of examined	MSE	A	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	X_{11}
1	56	0.07	-0.21	0.12	0.45	0.00	-0.06	0.39	0.02	0.01	0.08	0.03	0.08	-
2	56	0.08	-0.21	0.11	0.47	0.00	-0.05	0.42	0.02	0.01	0.09	0.05	-	0.02
3	56	0.07	-0.22	0.11	0.45	0.01	-0.03	0.39	0.02	0.01	0.08	-	0.09	0.02
4	56	0.08	-0.2	0.14	0.46	0.00	-0.07	0.46	0.02	0.01	-	0.04	0.1	0.02
5	56	0.07	-	0.12	0.45	0.01	-0.06	0.4	0.02	-	0.08	0.03	0.08	0.02
6	56	0.07	-0.17	0.15	0.46	0.00	-0.1	0.4	-	0.02	0.07	0.04	0.08	0.01
7	61	0.08	-0.26	0.1	0.38	0.01	-0.09	-	0.02	0.02	0.11	0.03	0.09	-0.02
8	56	0.07	-0.23	0.11	0.45	0.01	-	0.4	0.02	0.01	0.08	0.03	0.08	0.02
9	56	0.07	-0.19	0.14	0.47	-	-0.06	0.41	0.02	0.01	0.07	0.04	0.08	0.00
10	56	0.09	-0.27	0.1	-	0.01	-0.04	0.32	0.02	0.02	0.08	0.04	0.1	0.01
11	56	0.07	-0.2	-	0.43	0.01	-0.01	0.38	0.02	0.01	0.09	0.03	0.08	-0.02

The columns labeled «X$_m$» show the values of the coefficients before the regressor code X$_m$, which can be identified using Table 2. The free term of an equation are presented in the table in column "A". Linear regression model №1 and № 10 are designated as Y_1, Y_{10} has the form:

$$Y_1 = -0.21 + 0.12X_1 + 0.45X_2 + 0X_3 - 0.06X_4 + 0.39X_5 + 0.02X_6 + 0.01X_7 + \\ + 0.08X_8 + 0.03X_9 + 0.08X_{10} \quad (5)$$

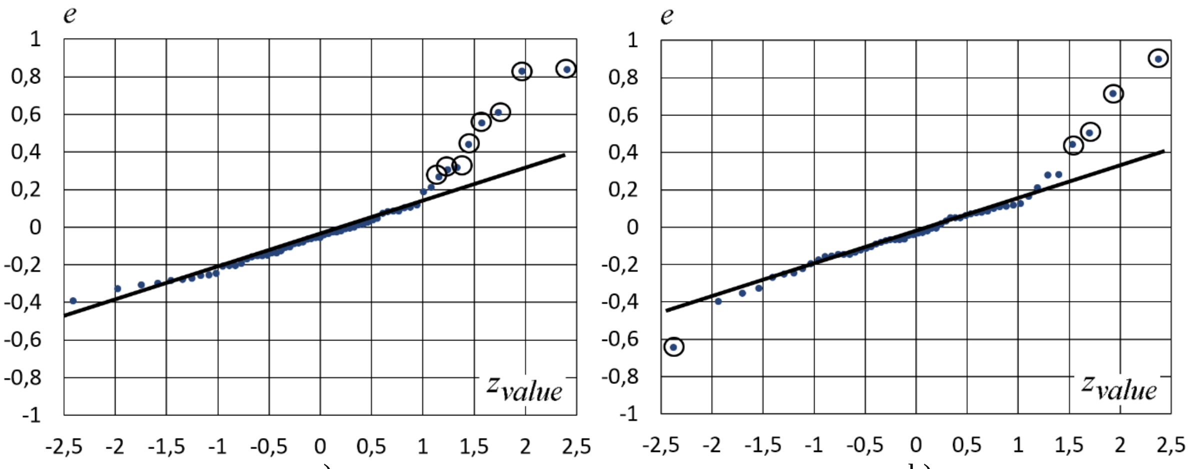
$$Y_{10} = -0.27 + 0.1X_1 + 0.01X_3 - 0.04X_4 + 0.32X_5 + 0.02X_6 + 0.02X_7 + 0.08X_8 + \\ + 0.04X_9 + 0.1X_{10} + 0.01X_{11}. \quad (6)$$

Residual plots presented in Figure 1 and Figure 2 match the linear regression models Y_1, Y_{10}, Y_2, Y_7. By virtue of the above-justified assumption, that error ε has a normal distribution, it follows that the points characterizing the value of residuals e_i for the value of prediction errors ε_i, should lie on a small neighborhood one straight line. Anomalous values are shown in circles on the graphs. The outliers were probably due to errors in the operation of the equipment used to change the values of quantitative factors or to the carelessness of the personnel in the preparation of raw survey data. Also, abnormal values can be associated with the presence of incorrect answers of patients in the personal survey sheet submitted for the study.



a) b)

Figure 1: Residual plot for 10-factor linear regression model: a) model Y_1; b) model Y_{10}.



a) b)

Figure 2: Residual plot for 10-factor linear regression model: a) model Y_7; b) model Y_2

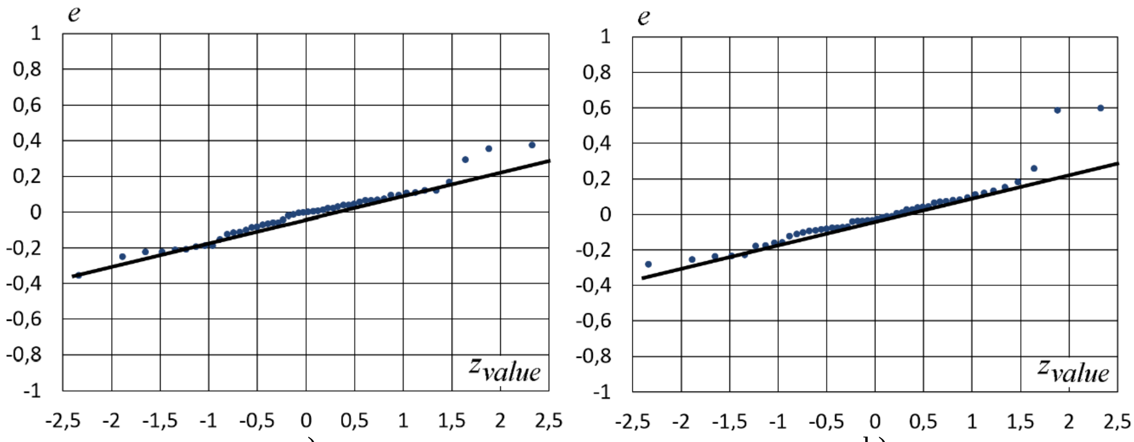
The presence of outliers can lead to a significant distortion of the form of the regression model, and, accordingly, to an increase in the error. In this regard, the anomalous values of the regressors in the prepared dataset should be changed or excluded from the set that will be used to build a linear regression model. Table 3 shows the MSE value for each of the models $Y_1 - Y_{10}$. Models Y_1, Y_{10} correspond to the lowest and highest MSE values. Each of the models in Table 3 has a significant number of anomalous values. The MSE value for each of the model $Y_1 - Y_{10}$ is approximately the same (Table 3).

To improve the prediction accuracy of the regression model, exclude outliers from the dataset and rebuild the models $Y_1 - Y_{10}$ based on the changed data. There are a number of methods for correcting the anomalous values that are presented in the dataset. We will take advantage of excluding rows from the dataset that correspond to patients with abnormal values of one or more regressors. After excluding six rows from the dataset, each of which corresponds to the outlier in Figure 1, the coefficients for the linear regression models were recalculated. Linear regression models Y_1, Y_{10} (5), (6) after recalculation of the coefficients have the form:

$$Y_1 = -0.19 + 0.11X_1 + 0.31X_2 - 0.01X_3 + 0.02X_4 + 0.31X_5 + 0.02X_6 + 0.01X_7 + \\ + 0.08X_8 + 0.03X_9 + 0.08X_{10}, \tag{7}$$

$$Y_{10} = -0.23 + 0.1X_1 - 0.002X_3 + 0.03X_4 + 0.29X_5 + 0.02X_6 \; + 0.02X_7 + 0.03X_8 + \\ + 0.05X_9 + 0.09X_{10} + 0.01X_{11}. \tag{8}$$

Linear regression models Y_1, Y_{10}, correspond to Residual plots, presented in Figure 1.



a) b)

Figure 3: Residual plot for 10--factor linear regression model after excluding outliers: a) model Y_1; b) model Y_{10}.

The final results of the analysis of the models after excluding outliers are presented in Table 4. For each model, the MSE value was determined before and after excluding anomalous values. There is a decrease in the MSE value by several times. For the model Y_1 the MSE value decreased from 0.071 to 0.027, and for the model Y_{10} the MSE value decreased from 0.088 to 0.037. As before excluding outliers, the model Y_1 has the best indicator according to the MSE criterion (4), and the model Y_{10} has the worst indicator. The characteristics of the models after excluding the anomalous values are presented in Table 4. As expected, the trend of decreasing MSE value for each of the models corresponds to the trend of decreasing MSE for the models Y_1, Y_{10}. Additionally, for each of the models presented in Table 4, a Residual plot is built and outliers are determined that are less significant than the initial ones and which can be used for a more in-depth analysis of the dataset. The headings of the table columns indicate the coded patient numbers to which the detected outliers correspond. The presence of a "+" symbol in the column indicates the presence of an outlier. The symbol "$\pm$" corresponds to a situation where the emission is negligible.

Analysis of the results presented in the table shows that the models Y_1, Y_5, Y_8 contain almost the same outliers with coded patient numbers 66, 39, 49, 35, 20, 26. Models Y_2, Y_4 also contain almost the same outliers with coded patients 66, 39, 35, 20, 26 and 66, 49, 35, 20, 26. Models Y_6, Y_9, Y_{11} contain outliers with coded patient numbers 66, 49, 39, 35, 20, 26.

Table 4
Characteristics for a 10-way linear regression model after eliminating outliers

№	№ model	MSE before	Number of patients											MSE after
			19	20	25	26	35	36	39	48	49	66	69	
1	1	0.071		+		+	+		+		+	+		0,027
2	5	0.071		+		+	+		+		+	+		0,028
3	8	0.071		+		+	+		+		+	+		0,028
4	3	0.072		+		+	+	±	±		±	+	±	0,03
5	11	0.072		+		+	+		+		+	+		0,027
6	9	0.073		+		+	+		+		+	+		0,026
7	6	0.074		+		+	+		+		+	+		0,033
8	4	0.075		+		+	+				+	+		0,028
9	2	0.076		+		+	+		+			+		0,031
10	7	0.079		+	+	+	+	+	+	+	+		+	0,034
11	10	0.088	+	+		+	+		+		+	+	+	0,037

Models Y_7, Y_{10} also contain almost the same outliers with coded patient numbers 69, 48 ,49, 36, 20, 39, 25, 26 and 66, 39, 19, 69, 25, 20, 49, 26. The sequence is indicated in ascending order of the residual error value. Analysis of the results of Table 4 allows us to form a set of rows to which outliers correspond and which are candidates for exclusion of the dataset.

3.2. Construction and analysis of 7-factor linear regression models

The next step of the study is the construction of 7-factor linear regression models. To build the models, we use the factors from Table 2. From the eleven factors, 330 seven-factor models were built and analyzed

$$C_{11}^{7} = \frac{11!}{7!(11-7)!} = 330, \tag{9}$$

coefficients for some of them are presented in table 5.

Models are selected from Table 5 according to the smallest and largest values of MSE, Y_{78} (MSE=0.025) и Y_{182}=(MSE=0.061):

$$Y_{78} = -0.18 + 0.09X_1 + 0.27X_2 + 0.02X_4 + 0.35X_5 + 0.03X_6 + 0.04X_9 + 0.1X_{10}, \tag{10}$$

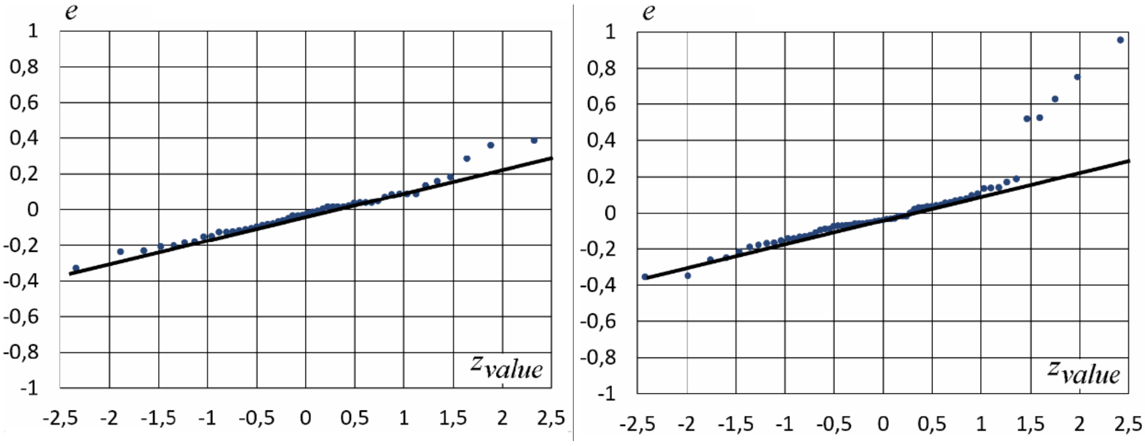
$$Y_{182} = -0.15 + 0.13X_1 - 0.01X_3 + 0.04X_7 + 0.04X_8 + 0.07X_9 + 0.08X_{10} - 0.03X_{11}. \tag{11}$$

The values of these indicators will be used to analyze the quality of prediction for models with a different number of regressors.

The residual plots shown in Figure 4 correspond to the linear regression models Y_{78}, Y_{182}. We see that the residual plot for the model Y_{182} (Figure 4) (Figure 4) contains a sufficiently large number of anomalous values for the residuals e_i, which led to a significant increase in the MSE value of the model Y_{182}.

Table 5
Coefficients for a seven-factor linear regression mode.

№	№ model	Number of examined	MSE	A	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	X_{11}
1	78	50	0.0245	-0.18	0.09	0.27	-	0.02	0.35	0.03	-	-	0.04	0.1	-
2	109	50	0.0246	-0.2	0.1	0.26	-	-	0.35	0.02	0.01	-	0.05	0.1	-
3	112	50	0.0246	-0.2	0.09	0.3	-	-	0.31	0.02	-	0.05	0.05	0.07	-
....	....	.	.	.		.	.	.	…	…					
328	178	63	0.0576	-0.2	0.09	-	-0.00	-	-	0.02	0.02	0.07	0.08	-	-0.02
329	179	63	0.0582	-0.21	0.1	-	0.01	-	-	0,02	0.02	0.04	-	0.12	-0.02
330	182	63	0.0608	-0.15	0.13	-	-0.01	-	-	-	0.04	0.04	0.07	0.09	-0.03



a) b)

Figure 4: Residual plot for 7-factor linear regression model: a) model Y_{78} ; b) model Y_{182}

Note that for the seven-factor linear regression model, anomalous values for the residuals are shown, but not excluded e_i , which correspond to certain outliers for the ten-factor model. The models have not been improved to allow comparison of the prediction accuracy of models built on the same dataset.

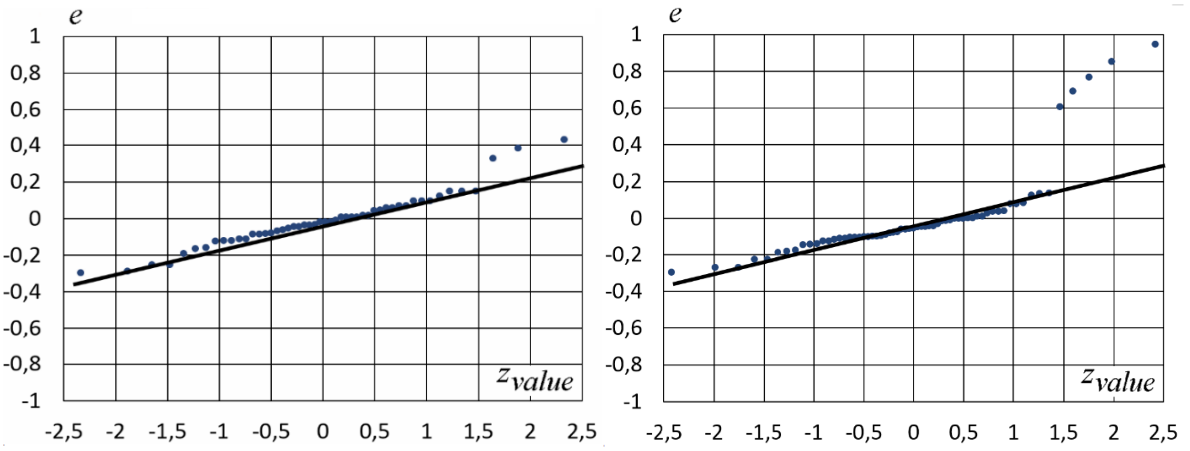
3.3. Construction and analysis of 5-factor linear regression models

When we construct five-factor models, we use the same approach that was used to build seven-factor models. To build the models, we use the factors from Table 2. From the eleven factors, 462 five-factor models were built and analyzed

$$C_{11}^{5} = \frac{11!}{5!(11-5)!} = 462, \tag{12}$$

coefficients for some of them are presented in table 6.

To graphically represent the analysis results, residual plots were selected for the model Y_{309} , with the lowest MSE value and for the model Y_{133} with the lowest MSE value and for the model (Figure 5).



a) b

Figure 5: Residual plot for 5-factor linear regression model: a) model Y_{309} ; b) model Y_{133}

Table 6

Coefficients for a five-factor linear regression model

№	№ model	Number of examined	MSE	A	X₁	X₂	X₃	X₄	X₅	X₆	X₇	X₈	X₉	X₁₀	X₁₁
1	309	50	0.0245	-0.15	-	0.29	-	-	0.36	0.03	-	-	0.05	0.08	-
2	34	70	0.0254	-0.00	0.11	0.36	-	0.03	0.48	-	-	-	-	-	-0.03
3	1	70	0.0264	-0.01	0.14	0.38	-0.01	0.04	0.49	-	-	-	-	-	-
....	....	.	.	.		.	.	.	...	...					
460	99	55	0.0665	-0.06	0.11	-	-0.00	0.05	-	-	0.05	-	-	-	-0.05
461	404	63	0.0666	-0.11	-	-	-0.01	-	-	-	0.05	0.06	-	0.1	-0.07
462	133	63	0.0673	-0.11	0.1	-	0.0	-	-	-	0.04	0.08	-	-	-0.04

Model Y_{309} with the lowest MSE and the model Y_{133} with the highest MSE have an analytic view:

$$Y_{309} = -0.15 + 0.29X_2 + 0.36X_5 + 0.03X_6 + 0.05X_9 + 0.09X_{10}, \qquad (13)$$

$$Y_{133} = -0.11 + 0.1X_1 + 0.001X_3 + 0.04X_7 + 0.08X_8 - 0.04X_{11}, \qquad (14)$$

The model Y_{133} with the highest MSE contain a sufficient number of anomalous values for the residuals e_i , which, as in the case of constructing seven-factor models, explains the high MSE value. It should be noted that the points characterizing the residual values e_i for the models Y_{133}, Y_{309} , with the exception of a few outliers, practically lie on one straight line. This allows us to assume that for the considered five-factor models, the error ε is distributed according to the normal law.

3.4. Construction and analysis of 3-factor, 2-factor, and paired linear regression models

Three-factor models are not widely used in the analysis of the severity of bronchial asthma disease. These models are used as an indicator for superficially determining the severity. However, we believe that three-factor models are worth considering for a general understanding of the issue of how much the forecast accuracy increases when moving from a three-factor linear regression model to a five-factor or seven-factor linear regression model. To build three-factor models, the factors from Table 2 were used. Of the eleven factors, 165 three-factor models were constructed and analyzed

$$C_{11}^{3} = \frac{11!}{3!(11-3)!} = 165, \qquad (15)$$

coefficients for some of them are presented in table 7. To demonstrate the analysis results, residual plots were selected for the model Y_3 and Y_{103} (Figure 6).

Table 7
Coefficients for a three-factor linear regression model.

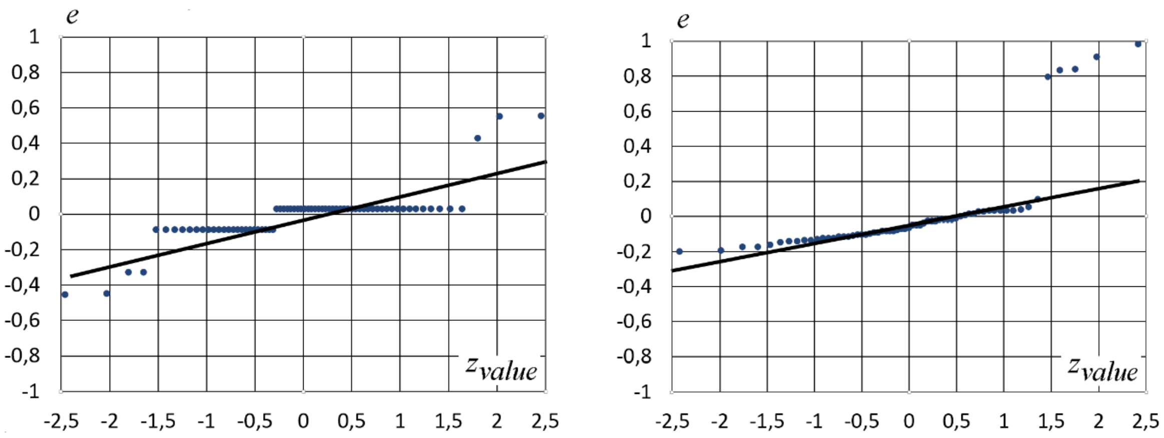
№	№ model	Number of examined	MSE	A	X₁	X₂	X₃	X₄	X₅	X₆	X₇	X₈	X₉	X₁₀	X₁₁
1	3	70	0.0248	-0.03	0.12	0.36	-	-	0.48	-	-	-	-	-	-
2	54	70	0.0254	0.02	-	0.37	-	0.090	0.51	-	-	-	-	-	-
3	47	70	0.0258	-0.01	-	0.37	0.00	-	0.5	-	-	-	-	-	-
….	….	.	.	.		.	.	.	…	…					
163	108	63	0.0695	0.07	-	-	-0.00	-	-	-	-	-	0.09	-	-0.08
164	106	63	0.0709	0.01	-	-	0.01	-	-	-	-	0.1	-	-	-0.08
165	103	63	0.0723	-0.02	-	-	0.00	-	-	-	0.06	-	-		-0.09

As in previous multivariate model analyzes, the first model Y_3 corresponds to the lowest MSE, and the second model corresponds to the highest MSE of the analyzed number from three-factor models. The analytical presentation of the models has the form:

$$Y_3 = -0.11 + 0.12X_1 + 0.36X_2 + 0.48X_5, \quad Y_{103} = -0.02 + 0.004X_3 + 0.06X_7 - 0.09X_{11}, \qquad (16)$$

The minimum and maximum MSE for three-factor models does not differ much from the minimum and maximum MSE for five-factor models.

The jump-like dependence of the values for the residuals e_i from z_{value} is explained by the fact that for the forecast regressors are used, which are represented by qualitative values (for example, there is the presence of a feature or there is no presence of a feature).



a) b)

Figure 6: Residual plot for a 3-factor linear regression model: a) model Y_3; b) model Y_{103}

Indeed, [Allergic rhinitis], [Atopic dermatitis] and [Bronchial asthma in relatives of second generation] were chosen as regressors for predicting the severity of bronchial asthma disease for the model corresponding to the best result in terms of the quality of fit (4).

Table 8

Coefficients for a two-factor linear regression model

№	№ model	Number of examined	MSE	A	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	X_{11}
1	13	70	0.026	0.05	-	0.38	-	-	0.5	-	-	-	-	-	-
2	4	70	0.034	-0.01	0.13	-	-	-	0.5	-	-	-	-	-	-
3	21	70	0.035	0.01	-	-	-	0.01	0.5	-	-	-	-	-	-
....	....	.	.	.		.	.	.	...	...					
53	24	63	0.07	-0.06	-	-	0.01	-	-	-	-	0.1	-	-	-
54	49	63	0.071	0.01	-	-	-	-	-	-	0.6	-	-	-	-0.11
55	23	63	0.072	-0.1	-	-	0.01	-	-	-	0.06	-	-	-	-

These factors are decisive in the superficial diagnosis of the observed value. In contrast to the model Y_3 (Figure 6.a), for the model Y_{103} the points characterizing the values of the residuals lie on one straight line, with the exception of several outliers e_i which contains each of the models considered above. Model Y_{103} is presented by regressors [Number of years from the first symptoms], [Domestic dust], [CD25 10*3 cells], among which the values of the two regressors are given by quantitative continuous variables. Thus, in three-factor models, the use of bronchial disease to predict the severity of the course of the disease is not appropriate.

Table 9
Model Coefficients for paired regression

№	№ model	Number of examined	MSE	A	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	X_{11}
1	5	70	0.036	0.05	-	-	-	-	0.47	-	-	-	-	-	-
2	2	83	0.05	0.04	-	0.36	-	-	-	-	-	-	-	-	-
3	4	75	0.051	0.05	-	-	-	0.1	-	-	-	-	-	-	-
....	....	.	.	.		.	.	.	...	...					
9	10	63	0.068	-0.01	-	-	-	-	-	-	-	-	-	0.13	-
10	8	63	0.07	-0.01	-	-	-	-	-	-	0.6	-	-	-	-
11	7	63	0.071	-0.07	-	-	0.01	-	-	-	0.06	-	-	-	-

In conclusion, we will consider two-factor (Table 8) and one-factor (Table 9) linear regression models. As a result of the analysis, 55 two-factor models and 11 paired regression models were considered:

$$C_{11}^{2}=\frac{11!}{2!(11-2)!}=55,\ \ C_{11}^{1}=\frac{11!}{1!(11-1)!}=11, \qquad (17)$$

coefficients for some of them are presented in tables 8 and 9.

Analytical representation of the two-factor model Y_{21}, Y_{23} is:

$$Y_{21}=0.01+0.01X_4+0.5X_5,\ Y_{23}=-0.1+0.01X_3+0.06X_7\,. \qquad (18)$$

The model Y_{21} corresponds to the lowest MSE and the model Y_{23} corresponds to the highest MSE. Note to MSE that the paired regression model contains the same factor that is present in the two-factor model:

$$Y_5=0.05+0.47X_5, Y_7=-0.07+0.06X_7\,. \qquad (19)$$

Thus, a two-factor linear regression model is a refinement of a paired regression model. An important note is that the paired regression model with the minimum MSE contains the [Bronchial asthma in relatives of second generation] regressor, which in the two-parameter model is supplemented by the [Bronchial asthma in father] factor, and in the three-parameter model [Allergic rhinitis], [Atopic dermatitis]. The prediction accuracy of the two-parameter regression model is quite close to the prediction accuracy of the three-parameter regression model.

4. Analysis of results

In the previous section, a detailed analysis of multivariate linear regression models, consisting of ten, seven, five, three, two factors, as well as an analysis of the paired regression model was carried out. For each category of models, the model with the lowest and highest MSE values was found. The obtained MSE values are used to compare the quality of predicting the observed value, presented in Figure 7. The dotted line in the graph shows the average value MSE, which is half the sum of the smallest and largest values.

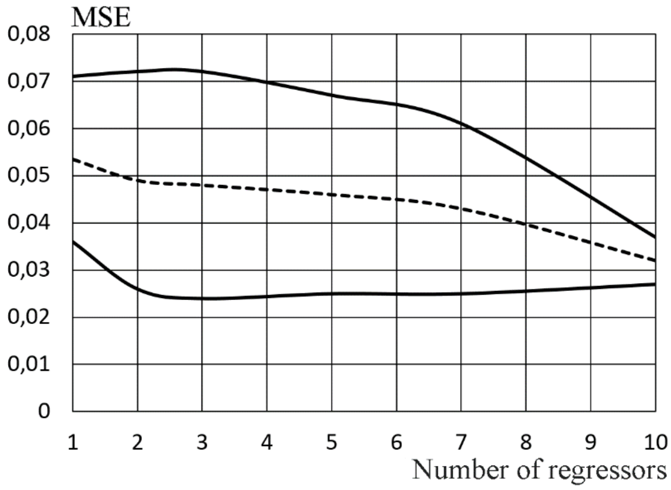


Figure 7: Criterion for the quality of predicting the severity of the course of bronchial asthma

The results obtained clearly show that for linear regression models dependent on two to five factors, the value MSE has almost the same value. Improving the forecasting quality is achieved by increasing the number of regressors. With an increase in the number of regressors, the range of variation of the value is significantly narrowed $MSE \in [MSE_{\min}; MSE_{\max}]$. In this case $MSE_{\min}$ the value changes slowly with an increase in the number of regressors. When moving from a five-factor linear regression model to a ten-factor linear regression model, the the average value MSE (the dotted line) decreased by an amount not exceeding 20%. Approximately the same value $MSE_{\min}$ is explained by the fact that when the number of factors decreases, outliers are excluded.

Indeed, on the one hand, a decrease in the number of factors should lead to an increase in the error. On the other hand, a model with fewer factors contains only outliers that correspond to the model factors, which accordingly improves the model's accuracy. In this regard, an important conclusion should be made about the need for preliminary data processing. The presence of outliers can lead to a decrease in accuracy with an increase in the number of regressors.

As shown in this work, due to the fact that for models with ten or more factors, the error ε has a normal distribution with distribution characteristics (3), and therefore, the linear regression model is the most successful for predicting the severity of bronchial asthma. However, the construction of regression models for predicting an observed value with a number of factors significantly greater than ten factors is associated with significant computational difficulties. A slow decrease in the value $MSE_{\min}$ with an increase in the number of model regressors practically makes it impossible to significantly increase the accuracy of the forecasting model by increasing the number of factors. The performed numerical experiments showed that the computational time required to calculate the coefficients of the linear regression model quadratically depends on the number of regressors in the model.

5. Conclusion

In this work, we performed a comparative analysis of the quality of predicting the severity of the course of bronchial asthma depending on the number of regressors in the model. For comparative analysis, a multivariate linear regression model was used. The substantiation of the distribution law for the forecasting error ε is given. The comparative analysis of values MSE for multivariate linear regression models using the example of the considered dataset shows that the use of models with less than six factors is inappropriate. The results obtained indicate that linear regression models with a small number of factors have approximately the same value MSE. As a result of performing this study, an important conclusion was obtained that the value MSE slowly decreases with an increase in the number of model regressors. This raises the relevance of the search for new methods for predicting the severity of bronchial asthma disease, including the use of Machine learning. A prospect for further research is to analyze the quality of fit the observed value depending on the number of regressors for different types of nonlinear regression models.

6. References

[1] Global Initiative for Asthma: Global Strategy for Asthma Management and Prevention. 2020. https://ginasthma.org/wpcontent/uploads/2019/01/2014-GINA.pdf.

[2] Wang, Xin, Tapani Ahonen, and Jari Nurmi. "Applying CDMA technique to network-on-chip." IEEE transactions on very large scale integration (VLSI) systems 15.10 (2007): 1091-1100.

[3] P. S. Abril, R. Plant, The patent holder's dilemma: Buy, sell, or troll?, Communications of the ACM 50 (2007) 36–44. doi:10.1145/1188913.1188915.

[4] S. Cohen, W. Nutt, Y. Sagic, Deciding equivalances among conjunctive aggregate queries, J. ACM 54 (2007). doi:10.1145/1219092.1219093.

[5] Fuchs O, Bahmer T, Rabe KF, von Mutius E. Asthma transition from childhood into adulthood. Lancet Respir Med 2017; 5:224–234.

[6] de Vries R, Dagelet YWF, Spoor P, Snoey E, Jak PMC, Brinkman P, et al. . Clinical and inflammatory phenotyping by breathomics in chronic airway diseases irrespective of the diagnostic label. Eur Respir J 2018; 51:1701817.

[7] Konradsen JR, Skantz E, Nordlund B, Lidegran M, James A, Ono J, et al. . Predicting asthma morbidity in children using proposed markers of Th2-type inflammation. Pediatr Allergy Immunol 2015; 26:772–779.

[8] Fitzpatrick AM, Jackson DJ, Mauger DT, Boehmer SJ, Phipatanakul W, Sheehan WJ, et al. . Individualized therapy for persistent asthma in young children. J Allergy Clin Immunol 2016; 138:1608–1618.e12.

[9] Fitzpatrick AM, Moore WC. Severe asthma phenotypes—how should they guide evaluation and treatment? J Allergy Clin Immunol Pract 2017; 5:901–908.

[10] Carr TF, Bleecker E. Asthma heterogeneity and severity. World Allergy Organ J. 2016; 9(1): 41

[11] Bush A, Fleming L, Saglani S. Severe asthma in children. Respirology. 2017; 22(5): 886- 897.

[12] Smit HA, Pinart M, Antó JM, et al. Childhood asthma prediction models: a systematic review. Lancet Respir Med. 2015; 3(12): 973-984.

[13] . Colicino S, Munblit D, Minelli C, Custovic A, Cullinan P. Validation of childhood asthma predictive tools: a systematic review. Clin Exp Allergy. 2019; 49(4): 410- 418.

[14] Amin P, Levin L, Epstein T, et al. Optimum predictors of childhood asthma: persistent wheeze or the Asthma Predictive Index? J Allergy Clin Immunol Pract. 2014; 2(6): 709- 715.

[15] Luo G, Nkoy FL, Stone BL, Schmick D, Johnson MD. A systematic review of predictive models for asthma development in children. BMC Med Inform Decis Mak. 2015; 15(99).

[16] Grabenhenrich LB, Reich A, Fischer F, et al. The novel 10-item asthma prediction tool: external validation in the German MAS birth cohort. PLoS ONE. 2014; 9(12):e115852.

[17] van der Mark LB, van Wonderen KE, Mohrs J, van Aalderen WM, ter Riet G, Bindels PJ. Predicting asthma in preschool children at high risk presenting in primary care: development of a clinical asthma prediction score. Prim Care Respir J. 2014;23(1):52–9.

[18] Chatzimichail E, Paraskakis E, Sitzimi M, Rigas A. An intelligent system approach for asthma prediction in symptomatic preschool children. Comput Math Methods Med. 2013;2013:240182.

[19] Caudri D, Wijga A, Schipper CM A, Hoekstra M, Postma DS, Koppelman GH, et al. Predicting the long-term prognosis of children with symptoms suggestive of asthma at preschool age. J Allergy Clin Immunol. 2009;124(5):903–10.

[20] Pescatore AM, Dogaru CM, Duembgen L, Silverman M, Gaillard EA, Spycher BD, et al. A simple asthma prediction tool for preschool children with wheeze or cough. J Allergy Clin Immunol. 2014;133(1):111–8.

[21] Mikalsen IB, Halvorsen T, Eide GE, Øymar K. Severe bronchiolitis in infancy: can asthma in adolescence be predicted? Pediatr Pulmonol. 2013;48(6):538–44.

[22] Vial Dupuy A, Amat F, Pereira B, Labbe A, Just J. A simple tool to identify infants at high risk of mild to severe childhood asthma: the persistent asthma predictive score. J Asthma. 2011;48(10):1015–21.

[23] Smolinska A, Klaassen EM, Dallinga JW, van de Kant KD, Jobsis Q, Moonen EJ, et al. Profiling of volatile organic compounds in exhaled breath as a strategy to find early predictive signatures of asthma in children. PLoS One. 2014;9(4), e95668.

[24] Marenholz I, Kerscher T, Bauerfeind A, Esparza-Gordillo J, Nickel R, Keil T, et al. An interaction between filaggrin mutations and early food sensitization improves the prediction of childhood asthma. J Allergy Clin Immunol. 2009;123(4):911–6.

[25] Bose S, Kenyon CC, Masino AJ Personalized prediction of early childhood asthma persistence: A machine learning approach. (2021) Personalized prediction of early childhood asthma persistence: A machine learning approach. PLOS ONE 16(3): e0247784. https://doi.org/10.1371/journal.pone.0247784.

[26] O. Kozhyna, O. Pihnastyi, Covariance coefficients factors from a clinical study of the severity of bronchial asthma in children of the Kh beforearkov region, 2017, Mendeley Data, 1, 2019.