Linear Regression Models for Bronchial Asthma Severity Prediction based on TSLP

Oleh Pihnastyi[a], Olha Kozhyna[b], Kostiantyn Voloshyn[c]

[a] *National Technical University "Kharkiv Polytechnic Institute", 2 Kyrpychova, Kharkiv, 61002, Ukraine*
[b] *Kharkiv National Medical University, 4 Nauky Avenue, Kharkiv, 61022, Ukraine*
[c] *V.N. Karazin Kharkiv National University, 4 Svobody Sq., Kharkiv, 61022, Ukraine*

Abstract

Bronchial asthma is a heterogeneous disease affecting more than 300 million people, and its occurrence increases every year. Severity of disease is characterized by interaction of genetic factors and numerous environmental factors. The technique of regression model construction is used to determine significance and dependence between factors. The data of 70 children with diagnosed bronchial asthma and 20 children of a control group were used in the study. 142 factors were studied and level of thymic stromal lymphopoietin (TSLP) in blood serum was taken as a value of a referable variable to construct regression models. Prediction model parameters were assessed. Errors distribution law was defined on the basis of a ten-factor regression model analysis. In the absence of large deviations from normal distribution law, validity coefficients are close to the errors under normal distribution law. A comparative study of histograms is provided to distribute model's regressors values.

Keywords

Bronchial asthma, child, regression model, residual plot, TSLP

1. Introduction

Bronchial asthma is a chronic pulmonary inflammatory disease [1]. Development of approaches in the disease treatment as well as new medications creation did not allow consolidating control over bronchial asthma. For this reason, the search of inflammation biomarkers is the most pressing challenge [2, 3].

Thymic stromal lymphopoietin can be used as one of inflammatory markers [4, 5]; that is hematopoetin cytokine isolated from mouse thymus epithelium culture [6, 7]. When studying the role of thymic stromal lymphopoietin its direct involvement in initiation and support of allergic inflammation in bronchial asthma is confirmed [8, 9]. Asthma phenotyping is based on diversity of disease clinical signs and heterogeneity [10, 11]. Special attention is paid to study of severe uncontrolled asthma in children [12, 13].

The phenotype of severe bronchial asthma in children is unique; it is different from the phenotype of mild and moderate bronchial asthma in children and from severe asthma in adults in terms of clinical and functional parameters, biological inflammatory markers, treatment response and prognostication [14, 15].

The use of machine learning in medicine makes it possible to analyze the interactions of multicomponent clinical data [16]. The apply of artificial intelligence to investigate the heterogeneity of asthma enhances the use and integration of rich asthma datasets and related clinical data [17]. Earlier detection of patients with a high risk of disease progression makes it possible to organize preventive measures, use individual treatment and observation methods as well as provide a lasting disease control [18].

ICST-2023: Information Control Systems & Technologies, September 21-23, 2023, Odesa, Ukraine.
EMAIL: pihnastyi@gmail.com (O. Pihnastyi); olga.kozhyna.s@gmail.com (O. Kozhyna); konstantin.voloshin@karazin.ua (K. Voloshyn)
ORCID: 0000-0002-5424-9843 (O. Pihnastyi); 0000-0002-4549-6105 (O. Kozhyna); 0000-0001-8262-5159 (K. Voloshyn)

Regression models are a common way to predict a course of bronchial asthma in children [19, 20]. Regression model defines a relationship between the observed value Y_i and factors X_K ($k = 1...K$) characterizing bronchial asthma. It can be represented as follows

$$Y_i = F(X_{1i}, X_{2i},..., X_{ki},..., X_{Ki}) + \varepsilon_i, \quad (1)$$

here Y_i is the value of a referable variable for the i-th test, $i = 1...n$; X_{ki} is the predictor's (regressor's) value for the i–th test; ε_i is the prediction error for the value of Y_i of the referable variable in the i–th test having the following properties:

$$E(\varepsilon_i) = 0, \qquad \sigma^2(\varepsilon_i, \varepsilon_i) = \sigma^2, \qquad \sigma^2(\varepsilon_i, \varepsilon_j) = 0,\ i \neq j. \quad (2)$$

The study purpose is research of the regression model features of thymic stromal lymphopoietin level in the blood serum.

The regression model (1) shows the trend of the referable variable change

$$\hat{Y} = F(X_1, X_2,..., X_K), \qquad E(Y) = \hat{Y}, \quad (3)$$

depending on the change of predictors' X_k values, which characterize bronchial asthma, and the range of Y_i values around the value of $F(X_1, X_2,..., X_K)$. The range of values for each level of predictors X_k follows probability distribution with mathematical expectation of $E(\varepsilon_i) = 0$ and mean square deviation of $\sigma(\varepsilon_i) = \sigma$.

Definition of a set of predictors X_k for the referable variable Y analyzing is the central task of regression model construction, which defines connection of factors characterizing bronchial asthma. Consistent increase of the number of predictors in the model is one of the methods (1). It is supposed that an addition of a new predictor to an existing set of predictors helps to decrease the $\sigma^2(\varepsilon_i)$ value. Selection of predictors X_k starts in the virtue of the regressor as a cause factor importance. The referable variable of Y depends on it. Selection of the type of functional dependence (3), as well as selection of the X_k factors, is one of the main problems of severe bronchial asthma study. In the most of the cases the dependence is defined as a result of a definite number of successive approximations. Model of linear regression is used as an initial approximation during functional dependence constructing [21]. Also special attention is to be paid to causal relationships of Y and X_k, given that statistical relationship of the referable variable Y and predicate X_k doesn't mean that the value of the referable variable Y causally depends on the value of predicate X_k. All the above questions will be studied in details in this paper.

2. Materials and methods

90 children aged 6 to 18 years were involved into the study of bronchial asthma in children in view of the phenotype associated with thymic stromal lymphopoietin. The main group amounted to 70 children with diagnosed bronchial asthma and the control group amounted to 20 children. The average age of children with bronchial asthma was 11 years. For every patient, information on 142 factors that could be the cause of bronchial asthma was gathered, processed and analyzed. 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 [22]. Parents were questioned about symptoms characteristic to bronchial asthma in patients as well as the patients'disease anamnesis. The results of the questioning were added to patient's materials. Clinical features of disease and results of laboratory tests were studied. Level of thymic stromal lymphopoietin in patients' blood serum was defined with enzyme immunoassay using a commercial test system manufactured by Bio-techne (ELISA, USA) on the immune-enzyme analyzer "Labline-90" (Austria). Enzyme immunoassay is based on sandwich-type technique which is characterized by cobinding of biotin-labeled antibodies to the analyzed analyte. Amount of thymic stromal lymphopoietin in a batch is defined on an analytical curve in accordance with routine practice for such experiments. Amount of thymic stromal lymphopoietin is measured in picograms per 1 ml of serum (pcg / ml) [23]. Indicators "Severe" and

"TSLP" were taken as a prediction parameter Y that defines severity of bronchial asthma in children. Negative effect of thymic stromal lymphopoietin cytokine on the course of disease was proved in [24, 25]. All the above questions will be studied in details in this paper.

3. Basic materials. Prediction model of bronchial asthma severity

3.1. Assessment of prediction model parameters

For linear model of multiple regression with model parameters of β_0, β_k :

$$Y_i = \beta_0 + \sum_{k=1}^{K} \beta_k X_{ki} + \varepsilon_i, \qquad \hat{Y}_i = \beta_0 + \sum_{k=1}^{K} \beta_k X_{ki}, \tag{4}$$

when regressor values X_{ki} are given, the value of the observed variable Y_i is defined by a sum of constant additive $\hat{Y}_i$ and random additive ε_i values. As defined for an error ε_i (2), the observed variable Y_i assessment follows:

$$E(Y_i) = E\left(\beta_0 + \sum_{k=1}^{K} \beta_k X_{ki} + \varepsilon_i\right) = \beta_0 + \sum_{k=1}^{K} \beta_k X_{ki} + E(\varepsilon_i) = \beta_0 + \sum_{k=1}^{K} \beta_k X_{ki}, \tag{5}$$

Model parameters of β_0, β_k have an assessment:

$$E(b_0) = \beta_0, \qquad E(b_k) = \beta_k, \qquad m_{x_k} = \frac{1}{n}\sum_{i=1}^{n} X_{ki}, \qquad m_y = \frac{1}{n}\sum_{i=1}^{n} Y_i, \tag{6}$$

$$b_0 = m_y - \sum_{k=1}^{K} m_{x_k} b_k, \qquad b_k = \frac{K_{x_k y}}{K_{x_k x_k}}, \tag{7}$$

$$K_{x_k x_k} = \frac{1}{n}\sum_{i=1}^{n} \left(X_{ki} - m_{x_k}\right)^2, \qquad K_{x_k y} = \frac{1}{n}\sum_{i=1}^{n} \left(X_{ki} - m_{x_k}\right)\left(Y_i - m_y\right).$$

Since, when predicting the value of the observed Y_i, the values of regressors X_{ki} are known, unknown values of β_0, β_k with assessment of their mathematical expectation (6) are defined by a linear combination of random values of Y_i of the observed variable. To estimate the mean square deviation of $\sigma(Y_i)$ let's use the definition of ε_i error (2)

$$\sigma^2\left(\beta_0 + \sum_{k=1}^{K} \beta_k X_{ki} + \varepsilon_i\right) = \sigma^2(\varepsilon_i), \ \sigma^2(\varepsilon_i, \varepsilon_j) = 0. \tag{8}$$

In the $\sigma = const$ assumption, the value of a random variable of Y is defined by probability distribution with mathematical expectation (5) and the mean square deviation of $\sigma = const$ which does not depend on predictor X value. Assessment of σ^2 value can be defined as follows:

$$\sigma^2 = E(MSE), \qquad MSE = \frac{SSE}{n-m}, \qquad SSE = \sum_{i=1}^{n} \left(Y_i - \hat{Y}_i\right)^2, \tag{9}$$

here $m = (K+1)$ is the number of constraints, it can be defined by a number of equations of (7) type. Retaining connections of $b_0 = B_0(Y_1, Y_2, ..., Y_n)$, $b_k = B_k(Y_1, Y_2, ..., Y_n)$ (7) correspond to the regression model (4). Regardless, $(n-m)$ of independent values Y_i can be defined. The rest m of Y_i values are defined from the system of equations (7). Expressions (6) and (9) give assessment for coefficients of the regression model and mean square deviation of σ. Defining of b_0, b_k coefficients according to the ordinary least squares technique, regardless the function of error rate distribution ε_i, makes it possible

to calculate the unbiased point estimations for b_0, b_k, that have minimal dispersion. However, to define the space of b_0, b_k parameters, assumption on the error rate ε_i distribution law is required.

3.2. Assessment of prediction error distribution law

When studying the factors characterizing bronchial asthma severity, let us introduce an assumption of a normal error rate distribution ε_i and define when the conditions of the assumption are met. For a normal law of error rate distribution ε_i with the conditions (2), designation $N(0,\sigma^2)$ is used. It is known that bronchial asthma severity is determined by a large number of factors that are weakly interconnected. When investigating severity of bronchial asthma in paper [26], more than 100 factors are being analyzed with correlation index $r_{x_k x_v}$:

$$r_{x_k x_v} = \frac{K_{x_k x_v}}{\sqrt{D_{x_k} D_{x_v}}}, \qquad \sigma_{x_k} = \sqrt{D_{x_k}}, \quad D_{x_k} = \frac{1}{n}\sum_{i=1}^{n}\left(X_{k\,i} - m_{x_k}\right)^2, \qquad k,v = 1...K, \tag{10}$$

between factors X_k, X_v that satisfies inequality $\left|r_{x_k x_v}\right| \le 0.15$. A weak relation is observed not only between factors that are used as the linear model regressors, but also between the referable Y value, characterizing bronchial asthma severity, and the regressor X_k. Since value of correlation index $r_{y\,x_k}$:

$$r_{y\,x_k} = \frac{K_{x_k y}}{\sqrt{D_y D_{x_k}}},, \qquad \sigma_y = \sqrt{D_y}, \qquad D_y = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - m_y\right)^2, \quad k,v = 1...K, \tag{11}$$

between the referable value Y and factor X_k is small, it is necessary for the model (1) to have enough number of regressors to provide required accuracy of referable value prediction. If the error rate ε of the referable value Y is formed under the influence of a large number of X_k factors, which are weakly dependent against each other and have similar rates of referable value Y formation, then the error rate distribution ε can be moved into proximity with normal distribution. If the number of factors which are weakly dependent on each other is more than ten, we can assume that the error rate distribution ε_i of the referable value for the linear regression model (4) is not different from the normal law of errors. Thus, error rate distribution ε_i follows the normal law of errors for the regression model (4). Indeed, if the model (4) has summands whose effect on the sum (16) dispersion is overbearingly large compared to other summands, then we can ignore the summands whose effect is vanishingly small. If the number of factors in resulting model is small, then the assumption of the normal law of errors ε_i should be discarded. Thus, due to the fact that the result of predicting of bronchial asthma severity depends on sufficiently large number of weakly dependent on each other X_k factors, having approximately similar rates of the referable variable Y value forming, prediction error ε_i distribution for a number of regressors $K \ge 10$ tends to the normal law of errors. In this study, let us use the model of linear regression (4) with ten regressors to justify our assumption about the normal law of errors. Additionally, let us mention that procedures of model parameters estimation based on the t-distribution are as a rule sensitive only to large deviations of errors from normal distribution. If there are no large deviations from normal law of errors, validity coefficients for any error are close to an error following the normal law of errors.

3.3. Assessment of factors using distribution histogram

Schematic prediction of regression factor of X_k is one the main criteria of model regressors selection. This analysis will allow us to obtain diagnostic information about a predictor variable by determining outlying values of X_k factor that has an effect on selection of regression function factors. Diagnostic information about a range and concentration of X levels in the study helps to adjust the range

of regression analysis certainty. For schematic prediction, when choosing model parameters, we use a distribution histogram of model regressors qualitative values. Distribution histograms of model regressors qualitative values are given in Fig.1.

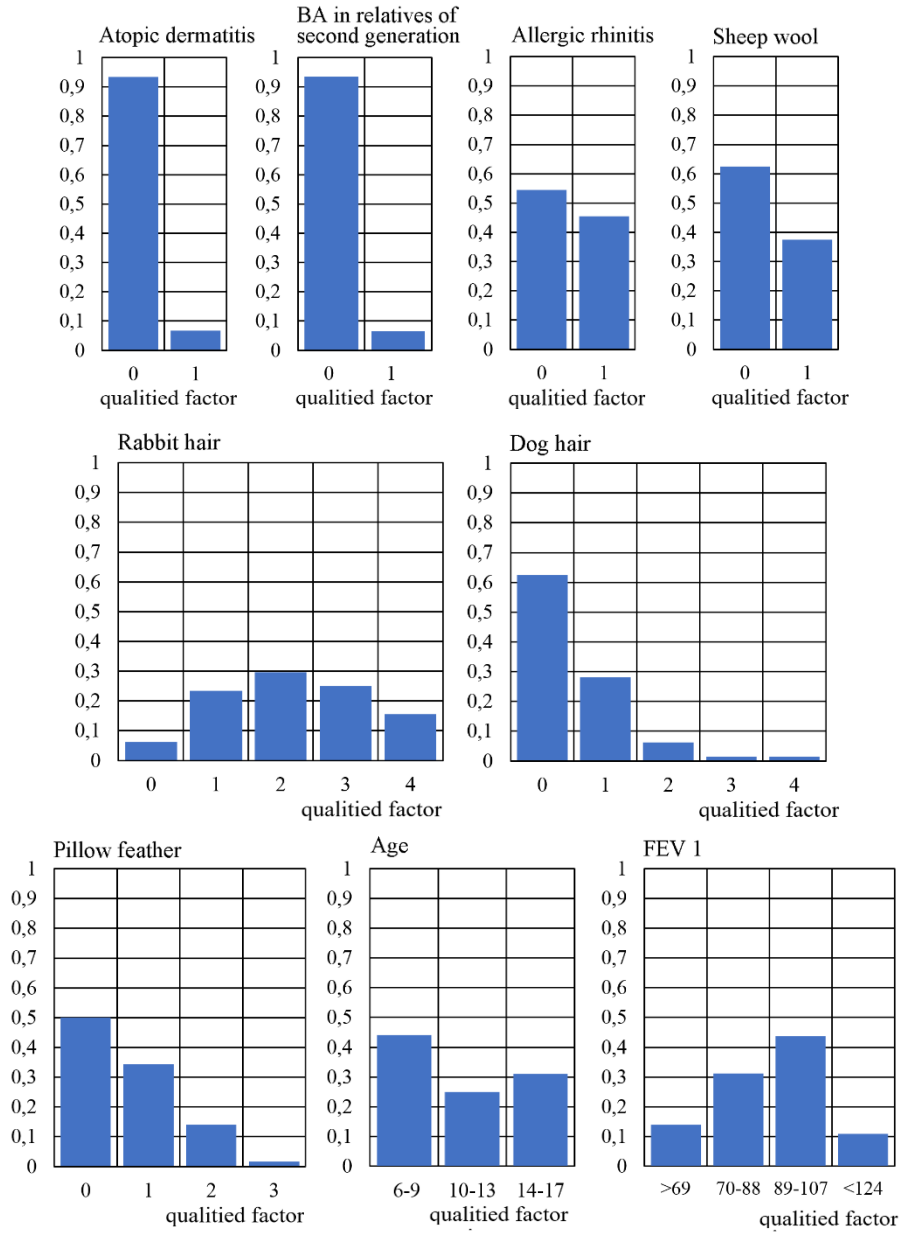

Figure 1: Distribution histogram of model regressors qualitative values

This analysis will allow us to correct the experimental data by excluding the data containing measurement errors. Comparative study of distribution histograms of model regressors values presented in various research papers is used to estimate the quality of experiment. The value for each regressor is defined by the value of a qualification factor, which characterizes the degree of regressor's effect on bronchial asthma severity (0 – no effect, and 1-,2-,3-,4- degree of regressor's effect on bronchial asthma manifestation in a child). The regressors "Atopic dermatitis", "Allergic rhinitis" and "BA in relatives of second generation" reproduces the results in papers [27]. The results of the regressors "Sheep wool", "Rabbit hair", "Dog hair", and "Pillow feather" are correspondent to the paper [28]. The rest of the factors correspond to statistical average data for the disease.

3.4. Model construction

Levels of thymic stromal lymphopoietin in 70 children were studied depending on clinical signs of bronchial asthma to determine the value of thymic stromal lymphopoietin in bronchial asthma pathogenesis, see table 1.

Table 1

Levels of thymic stromal lymphopoietin in children blood serum

Clinical sign of BA		Number of patients	Level of TSLP, pcg/ml Median (Q1;Q3)	P1	P2
Manifestation of asthma	-early (under 3 years)	20	17.93(6.13;40.63)	0.783	0.547
	-late (after 3 years)	50	12.44(5.95;28.01)		0.706
Duration of disease	-less than 3 years	16	7.84(5.95;19.93)	0.413	0.272
	-more than 3 years	54	13.17(5.04;32.64)		0.836
	-less than 7 years	31	7.84(5.22;20.47)	0.133	0.275
	-more than 7 years	39	13.94(5.95;36.45)		0.936
Atopy	-increased IgE level	59	11.76(4.50;27.01)	0.502	0.676
	-normal IgE level	11	13.21(9.58;20.47)		0.555
Clinical blood count	-eosinophilia	26	8.71(4.50;16.85)	0.099	0.277
	-level of eosinophils < 5 %	44	13.57(5.95;30.28)		0.971
Comorbid conditions	-atopic dermatitis	6	36.08(5.04;101.63)	0.195	0.324
	-atopic dermatitis	64	11.76(5/22;21.93)		0.472
	-allergic rhinitis	39	13.21(5.22;59.69)	0.131	0.731
	-no rhinitis	31	8.13(5.22;19.02)		0.159
Allergy heredity	-negative	30	15.03(7.37;55.33)	0.027	0.593
	-positive	40	7.99(4.50;21.2)		0.218
Asthma heredity	-negative	17	17.57(5.95;76.03)	0.148	0.437
	-positive	53	11.76(5.22;21.93)		0.340

P1 – comparison between the clinical sign presence and absence groups'
P2 – comparison with the control group

Analysis of difference of thymic stromal lymphopoietin levels depending on the presence or absence of asthma clinical signs sets only one possible difference – the value of negative heredity for allergy but not separately for asthma (table 2). The findings confirm the study of Gilda Varicchi, which shows that the gene of thymic stromal lymphopoietin is located on the 5q22.1 chromosome next to the "atopic cytokine" cluster on 5q31 and is responsible for allergy manifestations [29].

Table 2

Numerical characteristics of the factors severity of bronchial asthma

Code	Regressor name	m_x	σ_x
X_1	Atopic dermatitis	0.0562	0.2303
X_2	Bronchial asthma in relatives of second generation	0.0658	0.2479
X_3	Allergic rhinitis	0.4494	0.4974
X_4	Sheep wool	0.5217	0.6507
X_5	Domestic dust	2.2319	1.1312
X_6	Rabbit hair	0.5652	0.8925
X_7	Pillow feather	0.7536	0.8059
X_8	Dog hair	0.5362	0.8090
X_9	Bronchial asthma in father	0.0864	0.2810
X_{10}	Age	11.0674	3.6220

X_{11}	FEV1	100.739	18.0889
X_{12}	CD25 10*3 cells	0.6937	0.3087

A multifactor regression model can be constructed to make it possible to predict the disease severity. Let us use the following criteria during selection of m-factors of the regression model

$$r_{y\,x_m} \to \max, \quad r_{x_m x_v} \to \min, \tag{12}$$

here the calculation of $r_{y\,x_m}$, $r_{x_m x_v}$ is done in accordance with the formulas (10), (11). Raw experimental data for the factors selection are given in [30]. Numerical characteristics of the set of factors in accordance with the criterion (12) are given in Table 2. The selected factors are used to construct a linear regression model in this study (4).

The values of correlation coefficients between the factors of the $r_{x_m x_v}$ model, as well as between the model factors and the observed value $r_{y\,x_m}$ are given in Table 3.

Table 3

Correlation coefficients' values

	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	X_{11}	X_{12}
X_1	-	0,08	0,08	0,03	0,26	0,11	0,6	0,11	0,08	0.03	0,16	0.17
X_2	-0,08	-	0,10	0,21	0,25	0,05	0,29	0,00	-0,09	0,16	-0,19	-0,22
X_3	0,08	0,10	-	-0,01	0,27	0,05	0,09	-0,19	0,24	0,16	-0,20	-0,03
X_4	-0,03	0,21	-0,01	-	0,16	0,15	0,34	0,20	0,02	0,00	0,02	-0.17
X_5	0,26	0,25	0,27	0,16	-	0,08	0,19	0,12	-0,05	0,17	0,03	-0,07
X_6	-0,11	0,05	0,05	0,15	0,08	-	0,06	0,11	0,16	0,29	-0,20	-0,23
X_7	0,06	0,29	0,09	0,34	0,19	0,06	-	0,11	-0,05	0,04	0,04	0,08
X_8	0,11	0,00	-0,19	0,20	0,12	0,11	0,11	-	-0,17	-0,02	0,10	-0,07
X_9	0,08	0,09	0,24	0,02	-0,05	0,16	-0,05	-0,17	-	0,12	0,14	-0,05
X_{10}	-0,03	0,16	0,16	0,00	0,17	0,29	0,04	-0,02	-0,12	-	-0,15	-0,22
X_{11}	0,16	0,19	0,20	0,02	0,03	0,20	0,04	0,10	-0,14	-0,15	-	0,02
X_{12}	-0,17	-0,22	-0,03	-0,17	0,07	-0,23	0,08	-0,07	-0,05	-0,22	0,02	-
Severe	0,23	0,44	0,31	0,31	0,32	0,11	0,38	0,20	-0,10	0,18	0,19	-0,20
TSLP	0,25	0,22	0,25	0,48	0,22	0,50	0,30	0,16	0,27	0,23	0,26	-0,22

In Table 3, there are twelve factors for construction of a linear regression model, containing ten factors. The number of techniques which are used to select 10 regressors for the model (16) out of twelve factors is defined by the following expression

$$\frac{12!}{2!(12-2)!} = 66. \tag{13}$$

Thus, 132 models are studied in this paper: 66 models with the above mentioned regressors for the observed value of Severe and 66 models with the above mentioned regressors for the observed value of TSLP. In Tables 4 and 5 four models with ten regressors for the Severe and TSLP values prediction are given. The value of $\sqrt{MSE}$ (9) was the selection criterion of models for TSLP prediction. The models are placed in the descending order of the $\sqrt{MSE}$ value. The models to predict the Severe value contain the same regressors as the model to predict the TSLP. The value of $\sqrt{MSE}$ for the 66 TSLP prediction models corresponds to the inequality $22{,}97 < \sqrt{MSE} < 26{,}44$. It follows the assumtion that the regressors given in Table 2 have approximately similar rates of the referable value Y formation. If there is a predominant factor in a set of regressors, the value of $\sqrt{MSE}$ in the TSLP prediction models under

the absence of this factor would be significantly larger than in the model where this predominant factor is present.

Table 4
Coefficients for a ten-factor linear regression model for the observed Severe value.

№ model	Number of examined	$\sqrt{MSE}$	β_0	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8	β_9	β_{10}	β_{11}	β_{12}
53	56	0,310	0,1405	0,1465		-0,0901	0,2867		0,0362	0,0415	0,0773	0,0956	-0,002	-0,1372	-0,0008
49	56	0,307	0,203	0,151		-0,0596	0,3138	0,0632	0,0365	0,0451	0,096		-0,0021	-0,1815	-0,0029
---	---	---													
21	56	0,268	0,1676	0,1474	0,5055	-0,0933	0,3737	0,035			0,0783	0,0864	-0,0031	0,0031	0,002
17	56	0,267	0,1617	0,1493	0,4965	-0,1139	0,3788	0,0314		0,0259	0,0775	0,0791	-0,0029		0,0004

Table 5
Coefficients for a ten-factor linear regression model for the observed TSLP value.

№ model	Number of examined	$\sqrt{MSE}$	β_0	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8	β_9	β_{10}	β_{11}	β_{12}
115	56	26,44	33,40	6,77		9,85	6,09	4,69	3,62	8,03	8,85		-0,22	-16,62	0,09
76	56	25,74	22,92	7,65	31,57	6,98	19,34	3,68	3,24	8,58	-		-0,21	1,30	0,14
---	---	---													
96	56	23,13	24,93	6,61	29,76	4,68		1,29	2,61	6,02	6,75	11,16	-0,31		0,49
101	56	22,97	34,01	7,41	30,65	3,77	-	1,58		5,76	7,44	11,3	-0,32	-4,82	0,52

4. Analysis of results

4.1. Diagnostic assessment of the model using Residual Plot

Let us consider the residual

$$e_i = Y_i - \hat{Y}_i \tag{14}$$

as an observed value during diagnostic assessment of linear regression model. The observed residuals e_i show the properties of the ε_i error (4) that is fundamental for linear regression model assessment

$$m_e = \frac{1}{n}\sum_{i=1}^{n} e_i = 0, \qquad s^2 = \frac{1}{n-m}\sum_{i=1}^{n}\left(e_i - m_e\right)^2 = \frac{1}{n-m}\sum_{i=1}^{n} e_i^2 = MSE. \tag{15}$$

The number of freedom degrees for the linear regression model with ten regressors is $m = 11$. The value of $\sqrt{MSE}$ acts as an assessment of mean square deviation σ. We will use a residual plot showing the dependences of residuals $e = f\left(z_{value}\right)$ and $e = f\left(\hat{Y}\right)$ (Fig. 2) for the model diagnostic assessment. The value of $z_{value\,j}$ is defined by equation

$$P_j = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{z_j} \exp\left(-\frac{x^2}{2}\right)dx, \tag{16}$$

if $z_j = z_{value\,j}$. To calculate cumulative probability P_j for the sorted residual of e_j, $\left(j = 1..n_p\right)$ let us use the formula

$$P_j = (2j-1)/2n_p. \tag{17}$$

The emphasis should be on assessment of experimental data outliers. In Fig. 2, two outliers in experimental data used for prediction of TSLP in the linear model No 101 (Table 2) with ten regressors are circled. This model contains the factors given in Table 2, excluding X_4, X_6. Model No 101 with the specified ten factors has the minimum value of $s = 22.9728$ among analyzed 66 models of linear regression for the observed TSLP.

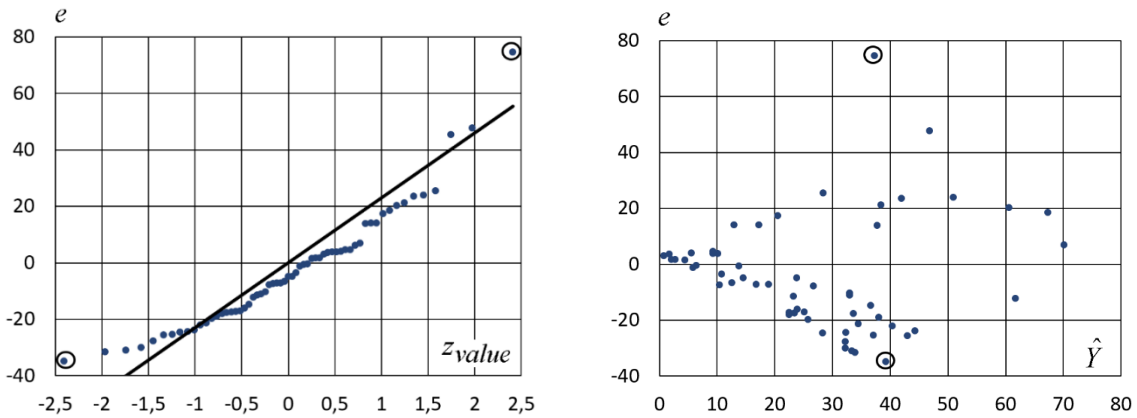

Figure 2: Residual plot with outliers 22–26

An outlier for the residual e_j can be connected with the measurement error of both the factor being predicted and one of its regressors. Let us construct M models with (M-1) regressors to define the source of the outlier in the M-factor model. If the outlier of e_j remains in each of M (M-1)-factor models constructed, hence the Y_i value is the source of the outlier. If for one of M models the outlier e_j disappears, then the factor which is absent in the model is the source of the outlier. This technique will be called M-regressors technique. On each step of the M-regressors technique application we remove several outliers specifying the source of their appearance (Table 6). The values of data for definite sources of outliers will be replaced with empty values. The criterion for outlier definition is as follows: experimental selections with minimum and maximum deviation of the actual value of e_j from the expected value of e_j (deviation of the actual value of e_j from the straight line the expected values are placed on) are taken. If absolute value of one of the values is significantly larger than of the other one, the iteration should be made for the maximum absolute value of e_j. Only the experiments with existing model regressors values were used to define coefficients of the linear regression model. In Table 6 iterative approximants to define the outliers for the selection with number of elements $n_p = 61$ with the total number of experiments of 90 are given. In 29 cases of experimental data, the model regressors

contain missed values. As the first iteration, residuals e_j for patients number 22 and 26 (Fig. 3-6) were assessed. The sources of outliers (Y_i, X_{mi}) were defined based on the outliers M in the models (M-1) study.

Table 6
Iterative approximants for definition of outliers.

No of approximation	Number of examined	$\sqrt{MSE}$	Counting number of e_i	e_i actual	e_i expected	z_{value}	Y_i	regressor, responsible for outlier
start		26,97						
1, low	22		1	-34,71	-55,31	-2,41	4,5	X_2
1, high	26	19,85	61	74,72	55,34	2,40	111,9	TSLP
2, low	81		1	-37,21	-51,33	-2,39	2,32	X_6, X_{10}, TSLP
2, high	39	17,89	59	70,70	51,35	2,49	94,5	X_1, X_2, X_3, X_5, X_7, X_8, TSLP
3, low	73		1	-32,31	-42,65	-2,38	8,13	TSLP
3, high	69	16,76	57	35,94	42,66	2,38	149.0	X_3
4, low	46		1	-28,32	-39,70	-2,37	2,32	X_5, X_7, X_8, TSLP
4, high	20	15,89	55	28,63	39,72	2,37	65.5	X_5, X_7, X_8, X_{10}

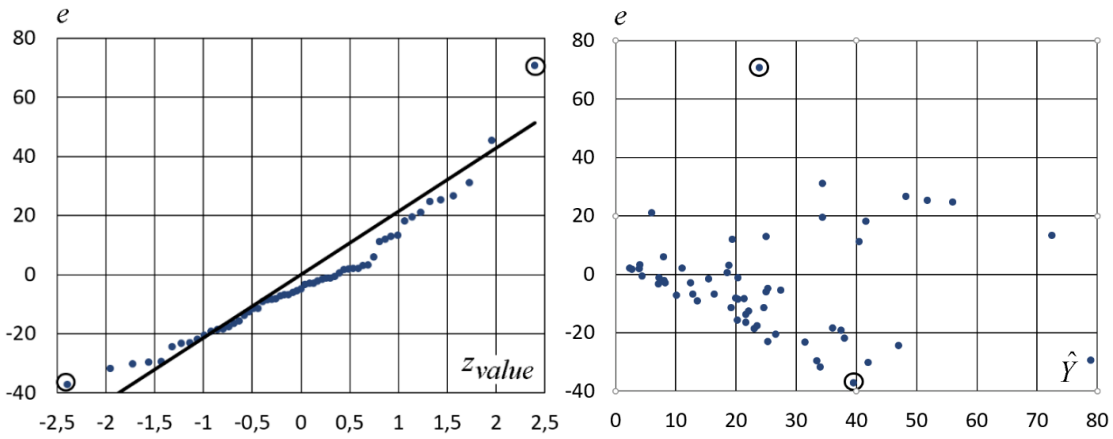

Figure 3: Residual plot with outliers 81–39

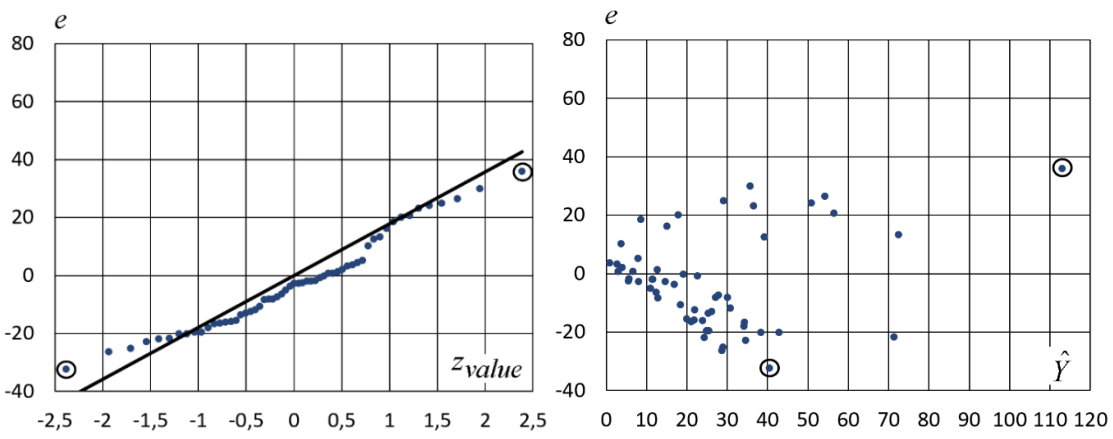

Figure 4: Residual plot with outliers 73–69

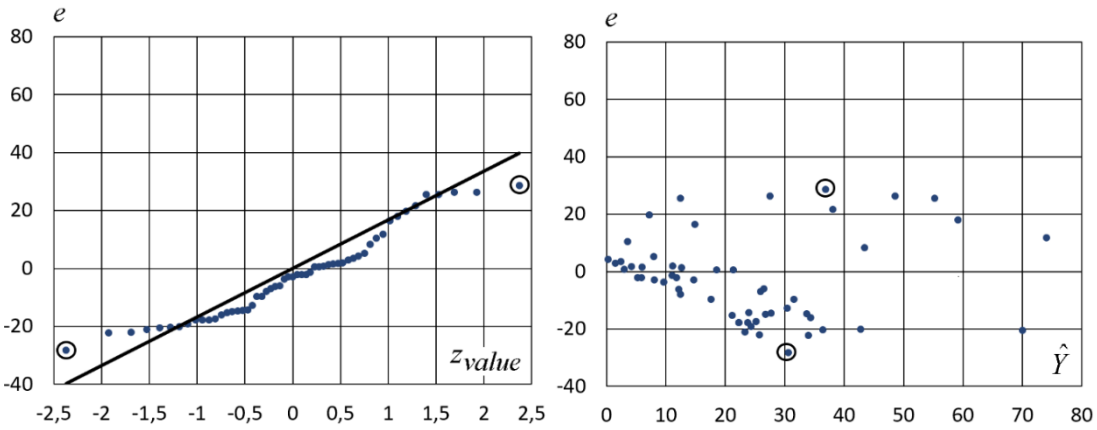

Figure 5: Residual plot with outliers 46–20

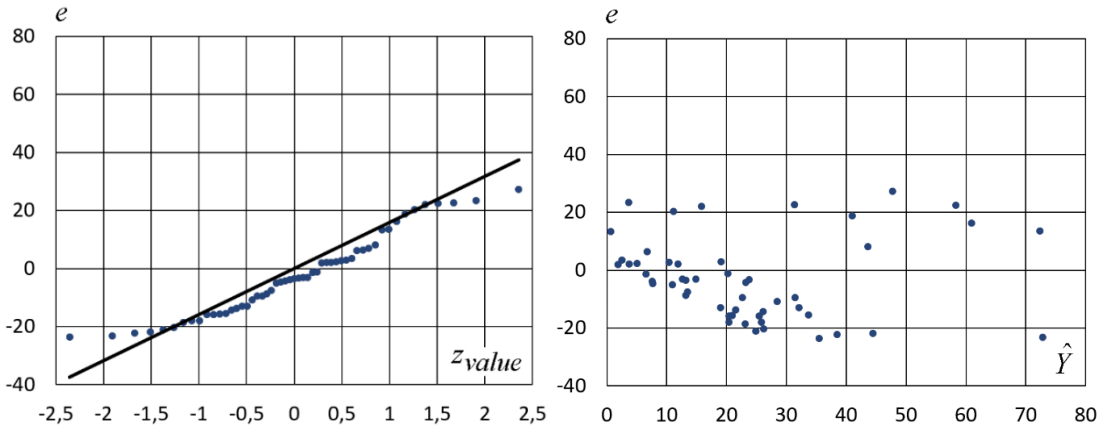

Figure 6: Residual plot after outliers removal

Fig. 2 shows the outliers for the first iteration. Using of the M-regressors technique helped to reveal the sources of outliers, one per each experiment. In the next iteration experiments 22 and 26 are excluded. Outliers for the second step of iteration are shown in Fig. 3. The three factors X_6, X_{10}, TSLP are the source for the outlier 81, and the seven factors X_1, X_2, X_3, X_5, X_7, X_8, TSLP are the source for the outlier 39. If there is no factor X_7 in a model, then there is no outlier as well. Along with this, factor X_7 is the source for both outlier 81 and 39.

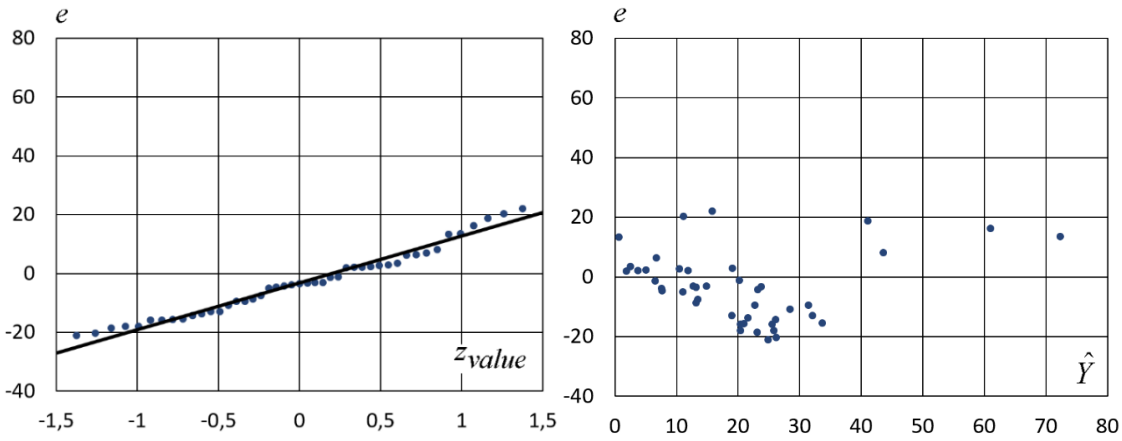

Figure 7: Residual plot for the selection values of $|z_{value}| > 1{,}5$

Since several factors are the source for each outlier, then it is supposed that the TSLP factor also takes an effect on the e_j value of the outlier. The consecutive removal of outliers (Table 6, Fig. 4, Fig. 5) resulted to the change of $\sqrt{MSE}$ value for the ten-factor model.

Table 4 shows the consecutive change of $\sqrt{MSE}$ value on each step of the outliers removal. After the outliers removal, the $\sqrt{MSE}$ value decreased by 30%: from $\sqrt{MSE}$ =22.97 to $\sqrt{MSE}$ =15.89 under 11 restrictions applied by linear regression model's coefficients. For the four values of the error e_j in Fig. 6 $z_{value} < -1{,}5$ and for the four values of the error e_j when $z_{value} > 1{,}5$. Error e_j probability can be defined from expression (16), where $z_j = e_j / \sqrt{MSE}$. Calculations show that $|z_j| \approx 1{,}5$. This corresponds to the error value probability of $P(|z| > 1{,}5) = P_j < 0{,}1$. In Fig. 7 residual plot of $e = f(\hat{Y})$ depending on the predicted value of $\hat{Y}$ for the selection values of $|z_{value}| > 1{,}5$ up to the value of $\sqrt{MSE}$ =10.51 for 45 patients in the selection and under 11 restrictions applied by linear regression model's coefficients is shown.

4.2. Asessment of regression model coefficients

Let us consider the test for β_m coefficients

$$H_0 : \beta_m = 0, \qquad H_\alpha : \beta_m \neq 0. \tag{23}$$

The case of $\beta_m = 0$ corresponds to assessment for Y

$$E(Y) = \beta_0 + \beta_1 X_1 + \ldots + \beta_{m-1} X_{m-1} + 0 \cdot X_m + \beta_{m+1} X_{m+1} + \ldots + \beta_M X_M .$$

For the error ε_i normal distribution this condition means not only the fact that there is no linear relationship between Y and X_m. In fact, it means that there is no relationship between Y and X_m at all, because for the model under studying only linear relationship between Y and X_m is supposed. The value of the b_m coefficient can be represented as

$$b_m = r_{yx_m} \frac{\sigma_y}{\sigma_{x_m}} + 0\left(r_{x_v x_m}\right) \approx \frac{K_{x_m y}}{K_{x_m x_m}} = \frac{\sum_{i=1}^{n} \left(X_{mi} - m_{x_m}\right)\left(Y_i - m_y\right)}{\sum_{i=1}^{n} \left(X_{mi} - m_{x_m}\right)^2}, \quad m_{x_m} = \frac{1}{n} \sum_{i=1}^{n} X_{mi}.$$

For regression model with normal error ε_i distribution

$$b_m \approx \frac{\sum_{i=1}^{n} \left(X_{mi} - m_{x_m}\right) Y_i - \sum_{i=1}^{n} \left(X_{mi} - m_{x_m}\right) m_y}{\sum_{i=1}^{n} \left(X_{mi} - m_{x_m}\right)^2} = \frac{\sum_{i=1}^{n} \left(X_{mi} - m_{x_m}\right) Y_i}{\sum_{i=1}^{n} \left(X_{mi} - m_{x_m}\right)^2} = \sum_{i=1}^{n} k_{mi} Y_i . \tag{24}$$

Coefficient k_{mi} is a determinate function of X_{mi} regressor having the following properties

$$\sum_{i=1}^{n} k_{mi} = \frac{\sum_{i=1}^{n} \left(X_{mi} - m_{x_m}\right)}{\sum_{i=1}^{n} \left(X_{mi} - m_{x_m}\right)^2} = 0, \qquad k_{mi} = \frac{X_{mi} - m_{x_m}}{\sum_{i=1}^{n} \left(X_{mi} - m_{x_m}\right)^2}, \tag{25}$$

$$\sum_{i=1}^{n} k_{m\,i} X_{mi} = \frac{\sum_{i=1}^{n} \left(X_{mi} - m_{x_m}\right) X_{mi}}{\sum_{i=1}^{n} \left(X_{mi} - m_{x_m}\right)^2} = 1, \quad \sum_{i=1}^{n} k_{mi}^2 = \frac{\sum_{i=1}^{n} \left(X_{mi} - m_{x_m}\right)}{\left(\sum_{i=1}^{n} \left(X_{mi} - m_{x_m}\right)^2\right)^2} = \frac{1}{\sum_{i=1}^{n} \left(X_{mi} - m_{x_m}\right)^2}.$$

Coefficient b_m normal distribution law results from the fact that coefficient b_m is a linear combination of random value of Y_i. To assess the regression model coefficient b_m, let us consider the following expressions:

$$\begin{aligned} E(b_m) &\approx E\left(\sum_{i=1}^{n} k_{m\,i} Y_i\right) = \sum_{i=1}^{n} k_{m\,i} E(Y_i) = \sum_{i=1}^{n} k_{m\,i} \left(\beta_0 + \sum_{v=1}^{M} \beta_v X_{v\,i}\right) = \beta_0 \sum_{i=1}^{n} k_{m\,i} + \sum_{i=1}^{n} k_{m\,i} \sum_{v=1}^{M} \beta_v X_{v\,i} = \\ &= \sum_{i=1}^{n} k_{m\,i} \sum_{v=1}^{M} \beta_v X_{v\,i} = \sum_{v=1}^{M} \beta_v \sum_{i=1}^{n} k_{m\,i} X_{v\,i} = \beta_m + \sum_{\substack{v=1,\\ v \neq m}}^{M} \beta_v \sum_{i=1}^{n} k_{m\,i} X_{v\,i} = \beta_m + 0\left(r_{x_v x_m}\right), \end{aligned} \tag{26}$$

$$\sigma^2(b_m) \approx \sigma^2\left(\sum_{i=1}^{n} k_{m\,i} Y_i\right) = \sum_{i=1}^{n} k_{m\,i}^2 \sigma^2(Y_i) = \sigma^2 \sum_{i=1}^{n} k_{m\,i}^2 + 0\left(r_{x_v x_m}\right) = \frac{\sigma^2}{\sum_{i=1}^{n} \left(X_{mi} - m_{x_m}\right)^2} + 0\left(r_{x_v x_m}\right).$$

We will use them to calculate the value of Student's t-test [31]

$$t_{m\,value} = \frac{b_m}{s(b_m)}, \qquad s(b_m) = \sigma(b_m) \approx \frac{\sigma}{\sqrt{\sum_{i=1}^{n} \left(X_{mi} - m_{x_m}\right)^2}}, \tag{27}$$

for coefficient b_m of the model when the number of items in selection is n=45 (Fig. 7) and the number of restrictions is $M = 11$. Critical value of $t_{\alpha=0.05, r=34} \approx 2.032$ corresponds to the statistics (27) when signification factor is $\alpha = 0.05$ and the number of freedom degrees is $r = (45 - 11) = 34$. Since distribution of $b_m / s(b_m)$ is the t-distribution, then confidence interval for b_m coefficients can be defined by relationship

$$P\left(b_m - s(b_m) \cdot t_{\alpha=0.05, r=34} < \beta_m < b_m + s(b_m) \cdot t_{\alpha=0.05, r=34}\right) = 1 - \alpha, \tag{28}$$

hence

$$\beta_m = b_m \pm \Delta b_m, \qquad \Delta b_m = s(b_m) \cdot t_{\alpha=0.05, r=34}. \tag{29}$$

Results of the confidence interval calculations for several factors are given in Table 7.

Table 7

Confidence intervals for coefficients β_m when $t_{\alpha=0.05, r=34} \approx 2.032$.

Factor	b_m	$b_m / s(b_m)$	p_{value}	Δb_m	$\beta_m = b_m \pm \Delta b_m$
Atopic dermatitis	62,59	5,29	0,0000	24,02	38.57< β_m <86,61
Rabbit hair	7,85	2,75	0,0093	5,80	2.06< β_m <13,65
Pillow feather	8,12	3,49	0,0014	4,73	3.39< β_m <12,85
FEV1	-0,46	4,23	0,0002	28,63	-0.69< β_m <-0,24

Table 7 shows that the confidence interval for coefficients β_m is limited rather widely. It proves the necessity to further improve the model. Improvement of the model includes data processing with the aim of model factors values decrease.

5. Conclusions

The technique of multifactor regression model construction is developed to predict bronchial asthma severity in children, when thymic stromal lymphopoietin level in blood serum is taken as a referable variable. Whereas disease severity is defined by a large number of factors that are weakly dependent on each other, assumption on normal error distribution was suggested. In preparation of experimental data comparative analysis of distribution histograms for model regressors values was used. The comparative study of experimental data and the data given in research papers on bronchial asthma severity was done to assess the quality of raw data. The results of histograms analysis were used to correct initial values of raw experimental data. During the analysis the data containing measurement error were excluded. The technique of multifactor linear regression model construction was developed to analyze bronchial asthma severity in children. The model of linear regression with ten regressors was studied in details. More than 100 factors effecting on bronchial asthma severity in children were analyzed to form a set of model's regressors. The criteria which allow forming an optimal set of model's regressors were developed. As a result of statistical analysis it was revealed that a confidence interval for coefficients of linear regression model is limited rather widely. It proves the necessity of the model further improvement. Further research perspectives involve model construction techniques based on nonlinear regression equations or neural network application.

6. References

[1] WHO. Asthma/ WHO, 2020). URL: https://www.who.int/news-room/fact-sheets/detail/asthma

[2] S. Sánchez-García, A. Habernau Mena, S. Quirce, Biomarkers in inflammometry pediatric asthma: utility in daily clinical practice, Eur Clin Respir, 4 1 (2017). doi:10.1080/20018525.2017.1356160.

[3] A. Licari, R. Castagnoli, I. Brambilla, A. Marseglia, M. Tosca, G. Marseglia et al, Asthma endotyping and bomarkers in childhood asthma, Pediatr Allergy Immunol Pulmonol, 31 2 (2018) 44–55. doi: 10.1089/ped.2018.0886.

[4] L. Bjerkan, A. Sonesson, K Schenck, Multiple functions of the new cytokine-based antimicrobial peptide Thymic stromal lymphopoietin (TSLP), Pharmaceuticals 31 2 (2016). doi: 10.3390/ph9030041.

[5] A. Chauchan, M. Singh, A. Agarwal et al, Correlation of TLSP, IL-33, and CD4+, CD5+, Fox P3+, T regulatory (Treg) in pediatric asthma, J Asthma, 52 (2015) 868–872.

[6] C. Kuo, S. Pavlidis, M. Loza, F. Baribaud, A. Rowe, I. Pandis et al, A Transcriptome-driven Analysis of Epithelial Brushings and Bronchial Biopsies to Define Asthma Phenotypes in U-BIOPRED, American Journal of Respiratory and Critical Care Medicine, 195 4 (2017) 443–455. doi:10.1164/rccm.201512-2452OC.

[7] M. Kuruvilla, F. Lee, G. Lee, Understanding Asthma Phenotypes, Endotypes, and Mechanisms of Disease, Clinical Reviews in Allergy & Immunology, 56 2 (2018) 219–233. doi: 10.1007/s12016-018-8712-1.

[8] M. Elmaraghy, M. Hodie, R. Khattab, M. Abdelgalel, Association between TSLP gene polymorphism and bronchial asthma in children in Beni Suef Governorate in Egypt, Comparative Clinical Pathology, 27 (2018) 565–570. doi:10.1007/s00580-017-2626-9.

[9] S. Liu, M. Verma, L. Michalec, W. Liu, A. Sripada, D. Rollins et al, Steroid resistance of airway type 2 innate lymphoid cells from patients with severe asthma: the role of thymic stromal lymphopoietin, J Allergy Clin Immunol, 141 1 (2018) 257-268. doi: 10.1016/j.jaci.2017.03.032.

[10] A. Koczulla, C. Vogelmeier, H. Garn, H. Renz, New concepts in asthma: clinical phenotypes and pathophysiological mechanisms, Drug Discov Today, 22 2 (2017) 388–396. doi: 10.1016/j.drudis.2016.11.008.

[11] M. Loza, R. Djukanovic, K. Chung, D. Horowitz, K. Ma, P. Branigan et al, Validated and longitudinally stable asthma phenotypes based on cluster analysis of the ADEPT study, Respiratory Research, 17 1 (2016). doi: 10.1186/s12931-016-0482-9.

[12] I. Pavord, N. Hanania, Controversies in Allergy: Should Severe Asthma with Eosinophilic Phenotype Always Be Treated with Anti-IL-5 Therapies, The Journal of Allergy and Clinical Immunology: In Practice, 7 5 2019, 1430–1436. doi: 10.1016/j.jaip.2019.03.010
[13] T. Bittar, S. Yousem, S. Wenzel, Pathobiology of severe asthma, Annual Review of Pathology: Mechanisms of Disease, 10 (2015) 511–545.
[14] S. Ramratnam, L. Bacharier, T. Guilbert, Severe asthma in children, J Allergy Clin Immunol Pract, 5 (2017) 889-898. doi: 10.1016/j.jaip.2017.04.031.
[15] A. Fitzpatrick, W. Moore, Severe Asthma Phenotypes – how should they guide evaluation and treatment? J Allergy Clin Immunol Pract, 5 4 (2017) 901-908. doi: 10.1016/j.jaip.2017.05.015.
[16] G. Collins, K. Moons, Reporting of artificial intelligence prediction models, Lancet, 393 (2019)1577-1579. doi:10.1016/S0140-6736(19)30037-6.
[17] A. Ray, J. Das, SE. Wenzel, Determining asthma endotypes and outcomes: Complementing existing clinical practice with modern machine learning, Cell reports. Medicine, 3 12 (2022) 10085. doi:10.1016/j.xcrm.2022.100857.
[18] L. Guilleminault, H. Ouksel, C.Belleguic, Y.Le Guen, P. Germaud, E. Desfleurs, Personalised medicine in asthma: from curative to preventive medicine, European Respiratory Review, 26 143 (2017) 160010. doi:10.1183/16000617.0010-2016.
[19] D. M Kothalawala, L. Kadalayil, V. Weiss, M. Aref Kyyaly, S. Hasan Arshad, John W Holloway et al, Prediction models for childhood asthma: A systematic review, Pediatr Allergy Immunol, 31 6 (2020) 616-627. doi: 10.1111/pai.13247.
[20] Y. Zhang, C. Zhou, J. Liu, H. Yang, S. Zhao, A new index to identify risk of multi-trigger wheezing in infants with first episode of wheezing, J Asthma, 51 10 (2014) 1043–1048. doi:10.3109/02770903.2014.936449.
[21] O. Pihnastyi, O. Kozhyna, Methods for constructing estimated two-factor linear regression models for diagnosing the severity of bronchial asthma in children, Innovare Journal of Medical Sciences, 9 1 (2021) 23–30. doi:10.22159/ijms.2021.v9i1.40408.
[22] UNESCO.ORG, Universal Declaration on Bioethics and Human Rights, 2005. URL: http://portal.unesco.org/en/ev.php. 2005.
[23] ELISA-Kit. URL: https://www.cusabio.com/ELISA-Kit/Human-thymic-stromal lymphopoietinTSLP-ELISA-Kit-110403.html
[24] B. Watson, G. Gauvreau, Thymic stromal lymphopoietin: a central regulator of allergic asthma, Journal Expert Opinion on Therapeutic Targets, 18 7 (2014) 771-785. doi: 10.1517/14728222.2014.915314.
[25] S. Colicino, D. Munblit, C. Minelli, A. Custovic, P. Cullinan, Validation of childhood asthma predictive tools: a systematic review, Clin Exp Allergy, 49 4 (2019) 410- 418.
[26] O. Kozhyna, O. Pihnastyi, Covariance coefficients factors from a clinical study of the severity of bronchial asthma in children of the Kharkov region, 2017, Mendeley Data, 1 (2019).
[27] G. Luo, FL Nkoy, BL Stone, D. Schmick, MD. Johnson, A systematic review of predictive models for asthma development in children, BMC Med Inform Decis Mak, 15 99 (2015). doi: 10.1186/s12911-015-0224-9.
[28] H. Mohammad, D. Belgrave, K. Kopec Harding, C. Murray, A. Simpson, A. Custovic, Age, sex, and the association between skin test responses and IgE titres with asthma, Pediatr Allergy Immunol, 27 (2016) 313–319. doi: 10.1111/pai.12534.
[29] G. Varricchi, A. Pecoraro , G. Marone, G. Criscuolo, G. Spadaro, A. Genovese et al, Thymic Stromal Lymphopoietin Isoforms, Inflammatory Disorders, and Cancer, Frontiers in Immunology, 9 (2018). doi:10.3389/fimmu.2018.01595.
[30] O. Kozhyna, O. Pihnastyi, Data Structure of Clinical Research, Human Health & Disease, 3 9 (2019) 71-79.
[31] Student, The probable error of a mean, Biometrika, 6 1 (1908)1-25.