IOP Conf. Series: Materials Science and Engineering **124** (2016) 012047 doi:10.1088/1757-899X/124/1/012047

Regression analysis for solving diagnosis problem of children's health

Yu A Cherkashina and O M Gerget
Tomsk Polytechnic University, 30, Lenina ave., Tomsk, 634050, Russia

E-mail:cherr999y@mail.ru

Abstract. The paper includes results of scientific researches. These researches are devoted to the application of statistical techniques, namely, regression analysis, to assess the health status of children in the neonatal period based on medical data (hemostatic parameters, parameters of blood tests, the gestational age, vascular-endothelial growth factor) measured at 3-5 days of children's life. In this paper a detailed description of the studied medical data is given. A binary logistic regression procedure is discussed in the paper. Basic results of the research are presented. A classification table of predicted values and factual observed values is shown, the overall percentage of correct recognition is determined. Regression equation coefficients are calculated, the general regression equation is written based on them. Based on the results of logistic regression, ROC analysis was performed, sensitivity and specificity of the model are calculated and ROC curves are constructed. These mathematical techniques allow carrying out diagnostics of health of children providing a high quality of recognition. The results make a significant contribution to the development of evidence-based medicine and have a high practical importance in the professional activity of the author.

1. Introduction

The problem of diagnosing health of children attracts attention of a growing number of researchers. This issue is especially relevant in pediatrics, where the main objective is to identify pathologies and chronic diseases at the early stages of organism development.

It is well known that the human predisposition to various diseases is laid mainly in the first year of life, which is why the prediction of a health status in this period is up-to-date [1].

Therefore, it is necessary to evaluate the state of children's health during the neonatal period to give the necessary recommendations to avoid prepathological and pathological changes at the early stages of organism development.

The aim is to use statistical methods to predict the state of children's health.

2. Characteristics of investigated data

For the research, the data provided by the doctors of Siberian State Medical University (SSMU) and a health center 'Healthy mother – strong baby' were used.

3 hemostatic parameters, 9 parameters of general and biochemical blood tests, the gestational age and a vascular endothelial growth factor, measured at 3-5 days of children's life are available for the diagnosis.

A hemostatic system is a biological system in the organism whose function is to save the liquid state of blood, to stop bleeding in case of damaged vessel walls and to dissolve blood clots which have fulfilled their functions. The investigations of the hemostatic system are easy to perform and have

highly informative results. The need to use hemostasis is confirmed by the fact that the hemostatic system disorders are the most dangerous in the neonatal period and reach 10% of the diagnoses of dead newborns.

The gestational age needs to account age-related morphological and functional features of brain vessels of newborns. The vascular endothelial growth factor influences the formation of post-hypoxic structural changes in the brain of the newborn, which are important for central nervous system damages.

A blood test is one of the most important methods of health monitoring. The necessary information for diagnosis of some diseases, disease severity and evaluation of dynamics of treatment can be obtained by the analysis of blood. The composition and concentration of cellular elements in blood vary under different physiological and pathological conditions.

3. Binary logistic regression

A regression analysis is the research of influence of one or more independent variables X_1, X_2, ..., X_p on dependent variable Y [1].

The task of the regression analysis is to build a mathematical model that allows evaluating the dependent variable by the values of independent variables [1].

The main purposes of the regression analysis are

- to predict values of the dependent variable using independent variables;
- to determine the contribution of individual independent variables in the change of the dependent variable.

General regression equation is determined by the equation [4]:

$$Y = F(X_1, X_2, ..., X_p), \tag{1}$$

where F - unknown function to be determined;
Y – dependent variable;
$X_1, X_2, \ldots, X_p$ – set of independent variables;
p – total number of independent variables.

There are two main types of a regression [4]:

- linear;
- nonlinear (hyperbolic, exponential, polynomial, logarithmic, logistics).

Logistic regression was used in the work for the regression analysis.

Logistic regression is a kind of a multiple regression. It is used when the dependent variable can have only two mutually exclusive values, i.e. it is binary (dichotomous). The binary logistic regression method allows studying the dependence of the dichotomous variable on independent variables that can be measured in different scales. In most cases of using dichotomous variables, it is about some events that may or may not happen (0 - Event has not happened, 1 - event happened). Binary logistic regression allows us to calculate the probability of the event, depending on the values of the independent variables. It uses the following regression equation

$$P(Y = 1 \mid X_1, ..., X_p) = \frac{1}{1 + \exp(-(\beta_0 + X_1\beta_1 + ... + X_p\beta_p))} \tag{2}$$

where Y – dependent variable which has values between 0 and 1;
$X_1, X_2, \ldots, X_p$ – set of independent variables;
$\beta_0, \ldots, \beta_p$ – logistic regression coefficients.

Therefore, the equation (2) describes the probability that the event occurs, that is, independent variable Y takes a value of 1, depending on the independent variables of X_1, ..., X_p.

If the value of P is less than 0.5, it can be assumed that the event will not occur, otherwise the occurrence of the event can be assumed.

A logistic curve is a dependence linking the likelihood of events and the value of Y. The logistic curve is shown in Figure 1.

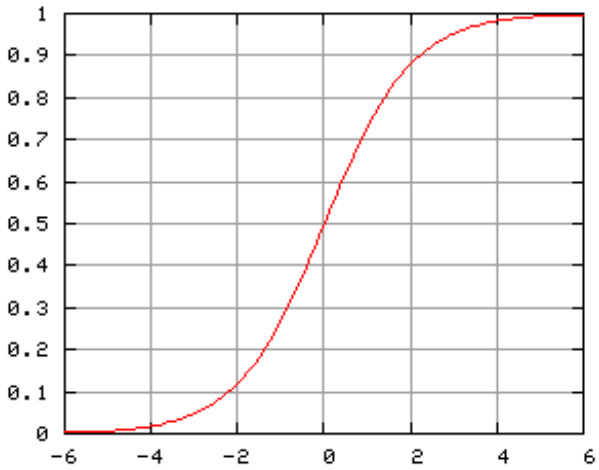

Figure 1. Logistic curve

From formula (2) and figure 1 we can see that regardless of regression coefficients β_0, ..., β_p or variables X_1, ..., X_p, predicted value Y will always be between 0 and 1 in this model.

The maximum likelihood method is used to find the coefficients of logistic regression. The essence of this method is the process of evaluating the regression coefficients coming down to maximization of the occurrence probability of a specific sample.

4. ROC-analysis

The ROC curve is a fundamental tool for the diagnostic test evaluation. ROC curves or Receiver Operating Characteristic curves are taken from the methodology of Signal Detection Analysis. Although the Theory of Signal Detectability (TSD) originally comes from the electronics and electrical engineering, it can also be used in medicine for the analysis of the interaction of sensitivity and specificity of the diagnostic test.

When one considers the results of a particular method in two populations, one population with a disease, the other population without the disease, a perfect separation between the two groups are rarely observed. Indeed, the distribution of the results will overlap, as shown in Figure 2.

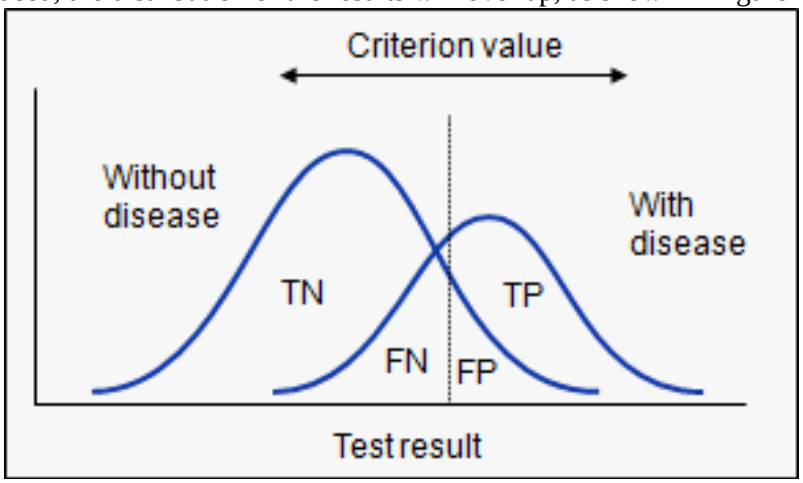

Figure 2. Distribution of the results of a particular method

For every possible cut-off point or criterion value one selects to discriminate between the two populations, there can be some cases with the disease correctly classified as positive (TP = True

Positive fraction), but some cases with the disease will be classified negative (FN = False Negative fraction). On the other hand, some cases without the disease can be correctly classified as negative (TN = True Negative fraction), but some cases without the disease will be classified as positive (FP = False Positive fraction).

Test sensitivity is the proportion of true positive predictions in the total number of patients:

$$Sensitivity = \frac{TP}{TP + FN} \tag{3}$$

This value characterizes the ability of the test to filter out patients with dubious diseases as accurately as possible.

A specificity of the method is the proportion of true negative in healthy patients:

$$Specificity = \frac{TN}{TP + FP} \tag{4}$$

The ability of the method is to detect only the patients with the disease characterized by this index.

Sensitivity and representativeness are reduced to 1 by using the ROC curve. The diagnosed value with a zero degree of prediction is represented by a line inclined at an angle of 45 degrees (diagonal). The more concave the curve of ROC, the more accurate the prediction of test results. The area under the ROC curve is an indicator of this property. The area is equal to 0.5 for the zero-degree prediction, and it is equal to 1 for the maximum degree of prediction.

5. Summary of the results

Experimental investigations were carried out to get a result (whether the child is ill or healthy) and to compare the result with the diagnosis made by a doctor.

The sample is formed by the results of the inspection of 102 children. It is divided into 2 classes (diagnosis): health (31 children) and patients (71 children).

The calculations were performed using the Statistical Package SRSS Statistics [5].

The regression analysis found that only three quantitative variables had an influence on the diagnosis: the period of reaction, gestational age, and total bilirubin. The other variables were not included in the regression equation.

The predicted values of the dependent variable are calculated from the regression equation, compared with the actual values observed in the classification table. The classification table is presented below (Table 1).

Table 1. Classification table

	Predicted		
Observed	health	ill	Percentage correct
health	30	1	96,8
ill	2	69	97,2
Overall percentage	97,1		

If the probability is less than 0.5, it means that the diagnosis is ‘healthy’ (the value of the variable ‘diagnosis’ is 0), in opposite case it is ‘sick’ (the value of the variable ‘diagnosis’ is 1). Table 1 shows that 97.1% of research subjects predicted the results were correct. This is a good outcome, the percentage of correct recognition of sick children is high.

Taking into account the found coefficients, the regression equation can be written as a formula:

$$P(Y = 1 \mid X_1, X_2, X_3) = \frac{1}{1 + e^{-(65{,}98 + 0 - 1{,}27 X_1 - 1{,}59 X_2 - 0{,}02 X_3)}}$$

where Y – dependent variable ‘diagnosis’;

X_1, X_2, X_3 – set of independent variables: reaction period, gestational age and total bilirubin.

To calculate the probability that a child is sick, regression equation variables X_1, X_2, X_3 corresponding to each object (the child) are substituted.

For conducting the ROC analysis it is necessary to calculate the following indicators relying on the logistic regression:

- TP=69;
- FN=2;
- TN=30;
- FP=1.

Let us calculate the sensitivity and specificity of the formulas (3) and (4).

$$Sensitivity = \frac{TP}{TP + FN} = \frac{69}{71} = 97.2$$

$$Specificity = \frac{TN}{TP + FP} = \frac{30}{31} = 96.8$$

The sensitivity characterizes the ability of the test as accurately as possible to filter out patients with questionable presence of the disease. The sensitivity of more than 90% means the ability of the test to have a high ability to recognize sick children.

The specificity characterizes the ability of a test to detect patients with extremely questionable presence of the disease. The value of representativeness is 96%, which means that the test has a high recognition ability of healthy children.

The ROC curve is shown in Figure 3.

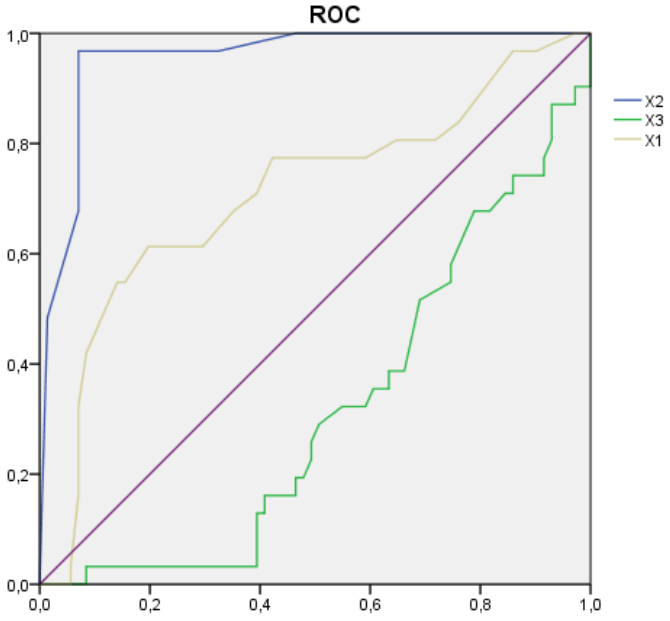

Figure 3. ROC curve

The more curved the curve of ROC, the more accurate the prediction of the test results. An indicator of this property is the area under the ROC, which test for a zero degree of prediction is equal to 0.5, and for the case with a maximum degree of prediction - 1. From Figure 3 it is clear that the largest area is for the index of X2.

6. Conclusion

Logistic regression is of great importance and has a practical application in medicine. The method was tested on the basis of real medical data provided by health professionals. The recognition quality of

97,1% of children was correctly classified to the appropriate class, therefore it can be considered acceptable.

References

[1] Gubler E V 1978 *Medicine* 296
[2] Gerget O M and Kochegurov V A 2002 *Tomsk, Publishing house TPU* 145
[3] Gerget O M and Devyatykh D V 2013 *Computer science and information technologies CSIT'2013: proceedings of the 15th International Workshop, Viena, 15 September-21 October 2013* 126-131
[4] Aguinis H 2004 *The Guilford press* 195
[5] 2015 SPSS software. Predictive analytics software and solutions [electronic resource] Available at: http://www-01.ibm.com/software/analytics/spss/
[6] Gerget, O M, Kochegurov V A, Titarenko E Yu 2015 Dynamic Susceptibility of a System *Applied Mechanics and Materials : Scientific Journal* **756** pp 378-381
[7] Tyurin Y N and Makarov A N 2008 Data analysis on the computer *Forum*