Kuráňová Pavlína
VŠB – Technical University of Ostrava, Ostrava Poruba, Czech Republic

Logistic regression as a relevant statistical tool for medical data investigation and evaluation

Keywords

logistic regression, medical data, atopy, Phadiatop test, surgery, morbidity.

Abstract

This paper presents the usage of logistic regression for predicting the classification of patients into one of the two groups. Our data come from patients who underwent Phadiatop test examinations and patients who underwent colectomy in the University Hospital of Ostrava. As the predictor variables were chosen personal and family anamneses for Phadiatop test and the physiological and operative scores for colectomy. For Phadiatop test, both of these anamneses were divided into four categories according to severity ranked by doctors. Scores for morbidity were based on the POSSUM system. The psychological score comprises 12 factors and the operative score comprises 6. The categorical dependent variable which we want to predict was Phadiatop test (respectively morbidity). The model for Phadiatop test was tested with the use of a medical database of 1027 clients and morbidity was tested upon a medical database of 364 clients. The developed models predict the right results with 75% probability for Phadiatop test and 70% probability for morbidity in surgery.

1. Introduction

The atopy rate of inhabitants of the Czech Republic is increasing. Atopy could be understood as a personal or family predisposition to become, mostly in childhood or adolescence, hyper sensible to normal exposure of allergens, usually proteins. These individuals are more sensitive to typical symptoms of asthma, eczema, etc. The Phadiatop test is used as a measure of atopy. Information obtained from personal and family anamneses were used for examining presence of asthma, allergic rhinitis, eczema or other forms of allergy (e.g. contact or food allergy). Family and personal anamneses of each patient were evaluated.

Unfortunately, the Phadiatop test is expensive, so we try to predict results of the test on the basis of a detailed family and personal anamneses. The knowledge of results of the Phadiatop test is very important especially for diagnosis of allergic dermatoses and also for the professional medical care for travellers [1]-[2].

Score system in surgery generally aims at quantification and consequently at objectification of risk of surgery patients. In particular, it means to determine the probability of occurrence of complications, morbidity for an individual patient or for a group of patients. Applied on the groups of patients they enable meaningful analysis of achieved records of morbidity and the stratification of patients. At the same time they provide a tool for objective assessment of newly implementing techniques and methods. In this case we have an operation and a physiological score. The operation score contains 6 hazard factors of the surgical intervention. The physiological score contains 12 factors related to the physiological state of the patient before operation. Scores for morbidity were based on the POSSUM system [6].

Information about the appearance of postoperative complications or atopy is given by results of Phadiatop test or morbidity, where there is the value 0 for none or low form atopy or morbidity and value 1 for medium or high form atopy or morbidity.

2. Logistic regression as a tool for discrimination

A common problem is to classify objects into one of the two given groups. Each object is described by

attributes. The aim of the task is to assign a new object into one of the groups. Assumed that the object belongs to one of the two groups (labeled as 0 and 1). The discriminatory problem will be solved on the basis of the logistic regression model. Generally, we have *n* objects with *p* measured attributes. But in case of some of objects, we do not know whether the object is a member of the group. The measured attributes are represented as *p*-dimensional random vectors $X_1, \ldots, X_n$.

The classification of the *i*-th object is expressed by random variable Y_i, which has the value 0 or 1 depending on their membership in the group.

The logistic regression was not originally created for the purpose of discrimination, but it can be successfully applied for this kind of analysis [3], [7]. A logistic regression model, which is modified for the purpose of discrimination, is defined as follows. Let $Y_1, \ldots, Y_n$ is a sequence of independent random variables with alternative distributions, whose parameters satisfy:

$$P(Y_i = 1 | X_i = x_i) = \frac{e^{\beta_0 + \beta' x}}{e^{\beta_0 + \beta' x} + 1},$$

$$P(Y_i = 0 | X_i = x_i) = \frac{1}{e^{\beta_0 + \beta' x} + 1},$$

for $i = 1, \ldots, n$, where $\beta' = (\beta_1, \ldots, \beta_p)'$, is unknown *p*-dimensional parameter and $X_1, \ldots, X_n$ are *(p+1)*-dimensional random vectors $(\beta_0, \ldots, \beta_p)$. This model can be called a learning phase, in which both values X_i and Y_i are known for each object (i.e. it is known to which group each object belongs to). Based on this knowledge, we try to predict parameters $\beta_0, \ldots, \beta_p$ and thus we try to estimate function $\pi(x)$, where

$$\pi(x) = P(Y = 1 | X = x) = \frac{e^{\beta_0 + \beta' x}}{e^{\beta_0 + \beta' x} + 1}.$$

Another object for which the classification is unknown is assigned to one of the two groups according to the value of decision function $\pi(x)$.

The object will be included in the first group if $\pi(x) > 0.5$. Otherwise, the object will be included in the second group. The main advantage of this model is that it does not require conditions for distributions of random vectors $X_1, \ldots, X_n$. However, the model assumes a very specific form of probability $P(Y = 1 | X = x)$ and we should verify the significance of the relationship

$$\pi(x) = P(Y = 1 | X = x) = \frac{e^{\beta_0 + \beta' x}}{e^{\beta_0 + \beta' x} + 1}$$

using appropriate statistical tests.

3 Application on biomedical data

The tested biomedical data are from the University Hospital of Ostrava, Department of Occupational and Preventive Medicine and Surgery, Ostrava, the Czech Republic. The logistic regression is used in order to predict the results of the Phadiatop test and morbidity. The medical database for predicted Phadiatop test contained information about 1027 patients, for operations it contained the morbidity information about 364 patients.

3.1. Phadiatop test

Patients in Group 0 have the results of Phadiatop test either 0 or I (no visible symptoms), so no treatment was necessary. The remaining patients with Phadiatop test II – VI are members of Group 1. For these patients a medical treatment is necessary.

We have one dependent variable *Y*, Phadiatop (*Ph*), which depends on two independent variable personal anamneses (*OA*) and family anamneses (*RA*).

Variable *Y* can be either 0 or 1, according to the membership of a patient to Group 0 or Group 1. Values of these independent factors were obtained from medical experts. The expert severity scores for personal and family anamneses are presented in Table 1. Here the category "Others" represents the score of various kind of allergies (e.g. contact or food allergies). The independent variable of the logistic regression was obtained as a sum of these scores for each patient from the database [4]-[5].

Table 1. Expert severity scores for personal and family anamneses.

Factor	Asthma	Allergic rhinitis	Eczema	Others
Score	10	8	6	4

3.2. Morbidity

The processed data come from the open and laparoscopic operations of the colon carrying out. Data file contains information about patients, such as their operational score *(OS)* and physiological score *(PS)* and the information about the appearance of postoperative complications. Operation score

contains 6 hazard factors of the surgical intervention (severity of the surgery, blood loss, contamination of peritoneum, etc.). Physiological score contains 12 factors related to the physiological state of the patient before surgery (age, cardiac stress, blood pressure, etc.). We have the information about the appearance of postoperative complications, where there is the value 0 for none or low form morbidity *(R)* and value 1 for medium or high morbidity.

3.3. Regression model

In accordance with the Chapter 2, the regression model has the form:

$$g(x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2,$$

where $\beta_i \in \mathbb{R}, i = 0, 1, 2$ are logistic regression coefficients and vector x has two components $x_1 = OA$, $x_2 = RA$ (respectively $x_1 = OS$, $x_2 = PS$). The dependence $\pi(x)$ on x has the form:

$$\pi(x) = \frac{e^{g(x)}}{1 + e^{g(x)}}.$$

Here $\pi(x)$ denotes the probability of occurrence of Phadiatop group, i.e. $\pi(x) = Ph$, (respectively $\pi(x) = R$). Unknown coefficients $\beta_i, i = 0, 1, 2$ are determined by the maximum likelihood method. In our case, the maximum likelihood for our logistic regression model can be expressed as

$$L(\beta) = \prod_{i=1}^{n} \pi(x_i)^{y_i} \cdot (1 - \pi(x_i))^{1-y_i},$$

where n is the number of patients. The vector $\beta = (\beta_0, \beta_1, \beta_2)$ denotes the unknown regression coefficients. When the i-th patient is identified as a member of Phadiatop or morbidity Group 1, then $y_i = 1$, otherwise $y_i = 0$. The vector $x_i, i = 1, \ldots, n$ has two components: $x_{i,1}=0A$ and $x_{i,2}=RA$ which represent personal and family anamneses, respectively and $x_{i,1}=0S$ and $x_{i,2}=PS$ which represent operative and physiological score.
The estimation of regression parameters β is provided by maximization the logarithm of the maximum likelihood, which can be expressed as:

$$\ln L(\beta) = \sum_{i=1}^{n} [y_i \cdot \ln \pi(x_i) + (1 - y_i) \cdot \ln(1 - \pi(x_i))].$$

Now the regression model can be used for predictions, whether the patient with given personal and family anamneses (respectively physiological and operative score) is a member of the selected Phadiatop group (respectively morbidity).

4. Prediction results and logistic model

4.1. Phadiatop test

In the first step, we try to create a regression model of all supplied data. We used the data corresponding to all 1027 patients. We obtained the following logistic model:

$$\ln\left(\frac{Ph}{1 - Ph}\right) = -1.5435 + 0.2124 \cdot OA + 0.0146 \cdot RA.$$

Results of this model are summarized in *Table 2*, column Case A.
Prediction results of Phadiatop test were incorrect for 220 patients, which we could describe as error of prediction rate model: $\frac{220}{1027} = 0.2142$.
In the second step, we created a learning group as a random sample from 90% of database (926 patients). To verify the correctness of model assumptions, the logistic model was created using the learning group. We obtained the following updated model:

$$\ln\left(\frac{Ph}{1 - Ph}\right) = -1.5667 + 0.2112 \cdot OA + 0.0199 \cdot RA.$$

For testing this updated model, we analyzed remaining data set, i.e. 10% of the database (102 patients), which were not assumed for the training phase. In this case we calculated prediction error of model, too: $\frac{24}{124} = 0.1935$. The results are summarized in *Table 2*, column Case B.

Table 2. Prediction results of regression models.

	Case A	Case B	Case C	Case D
Number of correctly classified patients	807	78	240	26
Number of incorrectly classified patients	220	24	124	10
Number of patients predicted for	233	21	88	2

Group 1				
Number of real patients in Group 1	331	33	78	12
Prediction error	21.4%	19.4%	33.7%	30.5%

4.2. Morbidity

In total we had 364 patient data, who underwent surgery operations. In the first step, we created a model containing all data from years 2001-2006. We received the following logistic model:

$$\ln\left(\frac{R}{1-R}\right) = -2.1997 + 0.0726 \cdot PS + 0.0549 \cdot OS .$$

The model of morbidity was incorrectly predicted for 124 patients, which give us the following error prediction rate of the model: $\frac{124}{368} = 0.337$. The outcomes of this model are written in *Table 2*, column in Case C.

In the second step, we took the data from the group of 328 patients from years 2001-2005 and their records have been used for the creation of the new logistic regression model:

$$\ln\left(\frac{R}{1-R}\right) = -2.3339 + 0.0637 \cdot PS + 0.0702 \cdot OS .$$

We applied this model to the group of 36 patients, who underwent the surgery operation in year 2006. In summary, results of morbidity were incorrectly predicted for 36 patients, which we could calculate as a prediction error of the model: $\frac{10}{36} = 0.3056$. Obtained results are summarized in Table 2, column Case D.

4.3. Verifications of the models

We made the test for models for Phadiatop test and morbidity. For Phadiatop test we tested a model created from 90% of data. For morbidity we tested a model created from data form years 2001-2005.

We also provided the analysis of variance for the models, see *Table 3*. Since the p-value is less than 0.01, there is a statistically significant relationship between variables at 99% confidence level.

We evaluated coefficients of *OA* and *RA* (respectively *OS* and *PS*) using the Pearson Chi-Square significance test, see *Table 4*. Variable *OA* (personal anamneses) is statistically significant at the 95% confidence level. On the other hand, p-value for *RA* variable (family anamneses) is larger than 0.05. Thus, *RA* variable is not statistically significant and may be excluded from the model. This result is maybe due to insufficient information on family anamneses in the database.

For coefficients of logistic regression *PS* and *OS* we can see (*Table 4*., lines Morbidity) that they are both statistically significant at the 95% confidence level. This result has shown, that both variables are important and we cannot exclude any of them.

Table 3. Analysis of variance.

	Source	Deviance	Df	P-Value
Phadiatop test	Model	197.312	2	0.0000
	Residual	966.168	923	0.1575
	Total(corr.)	1163.168	925	
Morbidity	Model	11.2321	2	0.0036
	Residual	432.099	324	0.0001
	Total (corr.)	443.332	326	

Table 4. Test of statistical significance.

	Factor	Chi-Square	Df	P-Value
Phadiatop test	OA	169.768	1	0.0000
	RA	2.12564	1	0.1448
Morbidity	OS	6.42149	1	0. 0113
	PS	5.37696	1	0.0204

5. Conclusion

The multidimensional logistic regression analysis has been applied to the actual medical data which describe the results of Phadiatop test and of the open surgeries of colon.

Phadiatop test is a cost-expensive medical procedure. For this reason it would be very interesting to predict patient diagnosis by assuming personal and family anamneses, which can be easily obtained. The data of patients include the results of the Phadiatop test with detailed description of personal illnesses, allergies and family anamneses. It was statistically proved that the family anamnesis is not statistically significant, probably due to insufficient information from patients.

Model for morbidity was designed and its good prediction qualities were demonstrated on the group of the patients from year 2006. Also the results of the analysis of variance and the likelihood ratio test of the significance of the regression coefficients were positive for this model. New prediction model for Phadiatop test which were developed using 90% of data and the updated model were successfully tested using remaining 10% of data.

Models which were created are suitable for predictions of the Phadiatop test with 75% probability of success and for morbidity with 70% probability of success.
In the future research we would like to predict the results of atopy or morbidity in more groups according to the seriousness of illnesses. Detailed analysis of results will also be important for future search of biomedical relations, which are hidden in the given biomedical database.

Acknowledgement:

The research is supported by The Ministry of Education, Youth and Sports of the Czech Republic under grant code 1M06047.

References

[1] Hajduková, Z., Pólová, J. & Kosek, V. (2005). The importance of Atopy Investigation in the Department of Travel Medicine. *Allergies: Magazine for continuous education in allergy and clinical immunology,* No.2, 106-109.
[2] Hajduková, Z., Vantuchová, Y., Klimková, P., Makhoul, M. & Hromádka, R. (2009). Atopy in patiens with allergic contact dermatitis. *Jurnal of Czech Physicians: Occupational therapy,* No.2, 69-73.
[3] Hosmer, D.W. & Lemeshow, S. (2000). *Applied Logistic Regression.* NewYork: Wiley-Interscience
[4] Kuráňová, P., Praks, P. & Hajduková, Z. (2010). Logistic regression as a tool for athopy investigation. *Mendel 2010,* 187-190.
[5] Kuráňová, P., Praks, P. & Hajduková, Z. (2010). *Statitical modelling of the Phadiatop test.* WOFEX 2010. 174-178.
[6] Martínek, L. (2006). *Aplikace specializovaných skórovacích systémů pro objektivizaci rizik laparoskopických operací kolorekta.* The doctoral thesis.
[7] Rabasová, M., Briš, R. & Kuráňová, P. (2010). Modified logistic regression as a tool for discrimination. *Reliability, Risk, and Safety: Theory and Applications. R. Briš, C. Guedes Soares, S. Martorell (eds.);* Vol 3, 1967-1972.