World Academy of Science, Engineering and Technology
International Journal of Medical and Health Sciences
Vol:2, No:2, 2008

Methods for Data Selection in Medical Databases: The Binary Logistic Regression - Relations with the Calculated Risks

Cristina G. Dascalu, Elena Mihaela Carausu, and Daniela Manuc

***Abstract*—**The medical studies often require different methods for parameters selection, as a second step of processing, after the database's designing and filling with information. One common task is the selection of fields that act as risk factors using well-known methods, in order to find the most relevant risk factors and to establish a possible hierarchy between them. Different methods are available in this purpose, one of the most known being the binary logistic regression. We will present the mathematical principles of this method and a practical example of using it in the analysis of the influence of 10 different psychiatric diagnostics over 4 different types of offences (in a database made from 289 psychiatric patients involved in different types of offences). Finally, we will make some observations about the relation between the risk factors hierarchy established through binary logistic regression and the individual risks, as well as the results of Chi-squared test. We will show that the hierarchy built using the binary logistic regression doesn't agree with the direct order of risk factors, even if it was naturally to assume this hypothesis as being always true.

***Keywords*—**Databases, risk factors, binary logistic regression, hierarchy.

I. Introduction

IN medical statistical studies a very important task is to design and to create the database for data collecting, because it must offers the optimal frame for all the physician's demands concerning data storing and further processing. After the data are stored into a regular database, the physician begins usually the second step of data processing, by trying to select the most relevant parameters for further medical interpretations. Many statistical methods are available in this purpose, and they are strongly connected with the data nature and the researcher's projects. A classic method, for example, is the principal components analysis, which takes in consideration all the parameters stored in a database and selects, using some mathematical principles, the most important ones – by identifying and eliminating the parameters that don't change the global nature and behavior of data when they are missing; another method in the same area is the discriminant analysis, used eventually in connection with different algorithms for data clustering.

Another method for parameters selection, a bit more complicated, takes in consideration the internal links between parameters. This is the binary logistic regression, which can be viewed as a generalization of the linear regression models, and is useful when we want to investigate the connections between one or more categorical independent variables (ordinal or binary) and a dependant categorical binary variable. This method is very useful in the study of risk factors over a certain situation (diagnostics, behavior, a.s.o.), because it builds a model that establish an hierarchy between all the possible risk factors, by selecting the most relevant ones, which have prediction value over the presence / absence of the investigated situation.

In this way the database we are working with can be substantially simplified (and eventually divided in smaller data sets) - in cases when we want to find statistical results about a well-defined diagnosis. In such a case we will select from the main database only the records where the investigated diagnosis is found, and for those records, only the relevant fields for the diagnosis, identified through binary logistic regression, in order to redirect the further statistical analysis only over those data**.**

II. Material and Methods

The binary logistic regression is used, as we said before, when we want to make a prediction about the presence / absence of a certain parameter based on the values of a set of independent predictor variables [1] – which are categorical, ordinal or binary. The logistic regression curve coefficients can be used to estimate the relative risks (odds ratio) for each independent variable used in the model. A thing important to be noticed is that, in order to build a logistic regression model, is not necessary to check in advance the regular requirements about the nature of the values distribution for the predictor variables or about their variance – therefore the logistic regression can be viewed as an available alternative to be used when the compulsory requirements for the discriminant analysis, for example, are not fulfilled. Because of these characteristics, the logistic regression is used with good results especially in epidemiologic databases, because this model approximates the probability to find a certain result (the dependant variable) when a certain set of conditions is checked (the

Manuscript received February 12, 2008.

C. G. Dascălu, Ph.D., Lecturer, is with the University of Medicine and Pharmacy "Gr. T. Popa", Iaşi, Romania – The Medical Informatics and Biostatistics Department, Faculty of Dental Medicine (phone: 0040-232-206441, e-mail: cdascalu@umfiasi.ro).

E. M. Carausu, Ph.D., Assoc. Professor, is with the University of Medicine and Pharmacy "Gr. T. Popa", Iaşi, Romania – The Public Health and Medical Management Department, Faculty of Dental Medicine (e-mail: cm72@email.ro).

D. Manuc is with Public Health Ministry, Bucharest, Romania (e-mail: cotrutz@yahoo.com).

independent variables).

Let's assume that we have n independent variables (predictors), x_1, x_2, … x_n, corresponding to the conditions we mentioned before, and we want to find the probability to appear the result (denoted by p_i) for an individual case i from all the cases in the studied database (for example a patient). In order to calculate this probability we add an auxiliary variable, denoted by Z, continuous, which can be interpreted as "the favorable tendency" to appear the desired result [2], [3] – in such a way so Z varies directly proportional with the p_i probability: as Z increases, the probability to appear the result increases too. Using the logistic regression model, the relation between Z and p_i is described by the function:

$$p_i = \frac{e^{z_i}}{1+e^{z_i}} = \frac{1}{1+e^{-z_i}}$$

$$\Rightarrow Z_i = \log\left(\frac{p_i}{1-p_i}\right).$$

Because Z is a continuous variable involved into a regressional model, it follows that the relation between Z and the predictor variables respects the general equation of the multiple linear regression:

$Z_i = b_0 + b_1 \cdot x_{i1} + b_2 \cdot x_{i2} + \ldots + b_n \cdot x_{in}$,

where b_j, $j \in \{1, 2, \ldots n\}$ - the coefficients of the predictors x_{i1}, x_{i2}, … x_{in}.

If the Z variable could be observed, in order to calculate the p_i probability it would be enough to fit a regression line to the values Z, x_{i1}, x_{i2}, … x_{in}. In practice, instead, the Z variable cannot be observed – therefore the p_i probability is calculated using the relation: $p_i = \frac{1}{1+e^{-(b_0+b_1 \cdot x_{i1}+b_2 \cdot x_{i2}+\ldots+b_n \cdot x_{in})}}$ - the equation of the logistic regression.

III. Results

We will use the binary logistic regression model in the analysis of a database made from 289 psychiatric patients from Iasi county, hospitalized in Grajduri hospital, and involved in 4 different types of offences. The set is made from 236 males (81.7%) and 53 females (18.3%), coming from the urban area (97 cases – 33.6%), but mostly from the rural area (192 cases – 66.4%). The patients suffer of 10 different types of psychiatric disorders, codified as it follows: DG1 – schizophrenia; DG2 – epilepsy; DG3 – organic psychic disorders; DG4 – ethylismus; DG5 – encephalopathy; DG6 – cerebral retard; DG7 – personality disorders; DG8 – bipolar affective disorder; DG9 – syndrome of depressive status; DG10 – psychotic disorders. The patients were involved in the following types of offences: OFF1 – murder; OFF2 – burglary; OFF3 – rape; OFF4 – serious bodily injury. The statistical processing was made using SPSS 15.0. All parameters were codified as binary variables (value 0 – absent / 1 – present) and our purpose was to identify from the 10 diagnostics, the most relevant ones (if there are) for each mentioned offence.

In this purpose we built every time a binary logistic regression model having the predictors DG1 – DG10 and the dependant variable OFF1 / respectively OFF2 … OFF4.

The first stage in defining the model consists in checking the hypothesis that the model describes adequately the observed data. In this purpose we used the Hosmer and Lemenshow test of "quality of fitting" [4], which gave results not statistically significant – which means that the model is adequate (Table I).

TABLE I
HOSMER AND LEMESHOW TEST – QUALITY OF DATA FITTING

	Step	Chi-square	df	Sig.		Step	Chi-square	df	Sig.
OFF1	1	.000	0	.	OFF2	1	.000	0	.
	2	.000	0	.		2	.007	1	.934
	3	.000	0	.		3	.070	1	.791
	4	.000	1	1.000		4	1.276	3	.735
OFF3	1	.000	0	.	OFF4	1	.000	0	.
						2	.000	1	1.000

In the following stage the logistic regression model analyzes all the predictors and makes their selection, by calculating for each of them a statistical score. The final purpose is to identify the predictors that influence significantly the values of the dependant variable, and therefore can be used to make hypothesis about those values. The calculation method was "Forward stepwise" [1], [4]: we begin at step 0 with no predictors and at each step we add a new predictor into the model - the one with the highest statistically significant score. The procedure stops when there are not new predictors to be added into the model (all the recorded scores are not statistically significant). Finally, the obtained predictors are checked again, in order to reconfirm that the probability changes added by these predictors are indeed significantly. We obtained the following results:

1) For the variable OFF1: The predictors were found in 4 steps, by adding them in the following order: (Table IIA). All the predictors were confirmed as adding statistically significant changes over the model's probability (Table IIIA).

TABLE IIA
THE SCORES OF THE PREDICTORS ADDED INTO THE MODEL – OFF1

OFF1		Score	df	Sig.
Step 1	DG9	6.263	1	.012
Step 2	DG1	7.758	1	.005
Step 3	DG5	11.201	1	.001
Step 4	DG10	10.243	1	.001

TABLE IIIA
CHANGES IN PROBABILITY BROUGHT BY THE PREDICTORS – OFF1

Variable		Model Log Likelihood	Change in -2 Log Likelihood	df	Sig. of the Change
Step 1	DG9	-155.281	4.968	1	.026
Step 2	DG1	-152.797	7.740	1	.005
	DG9	-152.237	6.620	1	.010
Step 3	DG1	-150.337	10.887	1	.001
	DG5	-148.927	8.067	1	.005
	DG9	-148.629	7.471	1	.006
Step 4	DG1	-149.436	17.874	1	.000
	DG5	-146.078	11.158	1	.001
	DG9	-145.321	9.644	1	.002
	DG10	-144.894	8.790	1	.003

2) For the variable OFF2: The predictors were also found in 4 steps, by adding them in the following order: DG6, DG2, DG4, DG10 (Table IIB). They influence statistically significant the model's probability (Table IIIB).

TABLE IIB
THE SCORES OF THE PREDICTORS ADDED INTO THE MODEL – OFF2

OFF2		Score	df	Sig.
Step 1	DG6	20.319	1	.000
Step 2	DG2	9.268	1	.002
Step 3	DG4	6.206	1	.013
Step 4	DG10	4.323	1	.038

TABLE IIIB
CHANGES IN PROBABILITY BROUGHT BY THE PREDICTORS – OFF2

Variable		Model Log Likelihood	Change in -2 Log Likelihood	df	Sig. of the Change
Step 1	DG6	-157.677	18.731	1	.000
Step 2	DG2	-148.311	7.919	1	.005
	DG6	-155.193	21.682	1	.000
Step 3	DG2	-145.278	7.027	1	.008
	DG4	-144.352	5.175	1	.023
	DG6	-153.321	23.113	1	.000
Step 4	DG2	-142.638	7.057	1	.008
	DG4	-141.332	4.446	1	.035
	DG6	-149.554	20.889	1	.000
	DG10	-141.764	5.310	1	.021

3) For the variable OFF3: The predictor was found in a single step: DG6 (Tables IIC, IIIC).

TABLE IIC
THE SCORES OF THE PREDICTORS ADDED INTO THE MODEL – OFF3

OFF3		Score	df	Sig.
Step 1	DG6	3.898	1	.048

TABLE IIIC
CHANGES IN PROBABILITY BROUGHT BY THE PREDICTORS – OFF3

Variable		Model Log Likelihood	Change in -2 Log Likelihood	df	Sig. of the Change
Step 1	DG6	-75.278	3.497	1	.061

4) For the variable OFF4: The predictors were found in 2 steps (Tables IID, IIID).

TABLE IID
THE SCORES OF THE PREDICTORS ADDED INTO THE MODEL – OFF4

OFF4		Score	df	Sig.
Step 1	DG7	5.224	1	.022
Step 2	DG1	4.732	1	.030

TABLE IIID
CHANGES IN PROBABILITY BROUGHT BY THE PREDICTORS – OFF4

Variable		Model Log Likelihood	Change in -2 Log Likelihood	df	Sig. of the Change
Step 1	DG7	-143.500	4.500	1	.034
Step 2	DG1	-141.251	4.762	1	.029
	DG7	-142.505	7.271	1	.007

The next step was to generate the classification tables, which show the practical results of using the identified regression model (Table IV). These tables show the number of correct and incorrect predictions made by the model at each step, as a consequence of using the predictors previously identified, and calculate the percentage of right predictions (at each step and globally). We have to check also here if the percentage is improved from a step to another or not.

TABLE IV
THE CLASSIFICATION TABLES

Observed			Predicted		
			OFF1		Percentage Correct
			0	1	
Step 1	OFF1	0	222	1	99.6
Overall Percentage: 77.9		1	63	3	4.5
Step 2	INFR1	0	222	1	99.6
Overall Percentage: 77.9		1	63	3	4.5
Step 3	INFR1	0	218	5	97.8
Overall Percentage: 78.2		1	58	8	12.1
Step 4	INFR1	0	218	5	97.8
Overall Percentage: 78.2		1	58	8	12.1
Observed			OFF2		Percentage Correct
			0	2	
Step 1	OFF2	0	221	0	100.0
Overall Percentage: 76.5		2	68	0	.0
Step 2	OFF2	0	220	1	99.5
Overall Percentage: 76.8		2	66	2	2.9
Step 3	OFF2	0	219	2	99.1
Overall Percentage: 77.2		2	64	4	5.9
Step 4	OFF2	0	219	2	99.1
Overall Percentage: 77.2		2	64	4	5.9
Observed			OFF3		Percentage Correct
			0	3	
Step 1	OFF3	0	268	0	100.0
Overall Percentage: 92.7		3	21	0	.0
Observed			OFF4		Percentage Correct
			0	4	
Step 1	OFF4	0	232	0	100.0
Overall Percentage: 80.3		4	57	0	.0
Step 2	OFF4	0	232	0	100.0
Overall Percentage: 80.3		4	57	0	.0

The synthesis of the obtained results is then presented in the tables of parameters estimation (Table V), which show the effect of each predictor over the values of the dependant variable. The coefficient of logistic regression, B, is the measure of each predictor's importance for the model – because it represents the probability for the dependant variable to have the "present" value when the predictor is present. The B sign shows also the nature of the relation between predictor and the dependant variable. The Wald statistics (= (B / S.E.)^2) is used to check if the predictor is significant for the model or not. The value Exp(B) has the significance of a relative risk (odds ratio), being calculated with the formula:

$\text{Exp(B)} = e^B = \frac{P(\text{var}.dep = 1)}{P(\text{var}.depend. = 0)}$ in the presence of the specified predictor.

TABLE V
THE PARAMETERS ESTIMATION

OFF1		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1(a)	DG9	.262	.129	4.108	1	.043	1.300
Step 2(b)	DG1	.802	.292	7.561	1	.006	2.229
	DG9	.307	.131	5.527	1	.019	1.359
Step 3(c)	DG1	.994	.309	10.376	1	.001	2.702
	DG5	.416	.143	8.515	1	.004	1.516
	DG9	.328	.131	6.276	1	.012	1.389
Step 4(d)	DG1	1.489	.383	15.091	1	.000	4.433
	DG5	.515	.150	11.846	1	.001	1.673
	DG9	.383	.133	8.249	1	.004	1.467
	DG10	.152	.050	9.114	1	.003	1.164
OFF2							
Step 1(a)	DG6	.217	.050	19.011	1	.000	1.242
Step 2(b)	DG2	.649	.223	8.462	1	.004	1.915
	DG6	.240	.052	21.647	1	.000	1.271
Step 3(c)	DG2	.622	.227	7.487	1	.006	1.863
	DG4	.418	.181	5.348	1	.021	1.519
	DG6	.251	.052	22.932	1	.000	1.285
Step 4(d)	DG2	.632	.231	7.456	1	.006	1.881
	DG4	.387	.181	4.561	1	.033	1.472
	DG6	.240	.053	20.698	1	.000	1.272
	DG10	-.146	.076	3.726	1	.054	.864
OFF3							
Step 1(a)	DG6	.149	.077	3.699	1	.054	1.160
OFF4							
Step 1(a)	DG7	.143	.064	4.900	1	.027	1.153
Step 2(b)	DG1	.705	.328	4.623	1	.032	2.024
	DG7	.197	.070	7.854	1	.005	1.218

In conclusion, from this analysis follows that, if we want to make further studies about the patients involved in OFF1 – murder, it is enough to take in considerations only the diagnostics DG9 – syndrome of depressive status; DG1 – schizophrenia; DG5 – encephalopathy and DG10 – psychotic disorders – eventually in this order; to study the patients involved in OFF2 – burglary, it is enough to take in consideration only the diagnostics DG6 – cerebral retard; DG2 – epilepsy; DG4 – ethylismus and again DG10 – psychotic disorders; to study the patients involved in OFF3 – rape, it is enough to take in considerations only the diagnostics DG6 – cerebral retard, and to study the patients involved in OFF4 – serious bodily injury, it is enough to take in considerations only the diagnostics DG7 – personality disorders and DG1– schizophrenia.

IV. DISCUSSIONS

Now we are going to compare these results with the direct calculated risks for each diagnosis, taken in consideration independently. As we said before, the value Exp(B) from Table 5 has the significance of a relative risk – being exactly equal with the odds ratio when we build the logistic regression model using only a single predictor and the dependant variable.

In this context, our assumption was that the order of predictors selected by the logistic regression model must be the same with the predictors descending order according to their independent calculated odds ratio or positive risks. This order can be eventually influenced by the results of the Chi-squared significance test (only the predictors that have a statistically significant influence over the dependant variable being taken into consideration). The number of predictors found within the regression model must also be equal with the number of predictors having positive risks (risks > 1.00) and statistically significant results of the Chi-squared test.

In order to check this assumption, we calculated these statistics for each dependant variable. The results are showed in Tables VIA – 6D (the cells colored in grey shows the positive risks).

TABLE VIA
THE DIRECT RISKS FOR EACH DIAGNOSIS – OFF1

Diagnosis	The risk for the cohort OFF1 = 1	Odds Ratio (1/0)	Pearson Chi-squared	Asimpt. sig. p	
DG1	1.707	2.006	6.118	**.013**	**SS**
DG2	.316	.259	3.719	**.054**	**SS**
DG3	.917	.895	.037	.848	NS
DG4	-	-	2.749	.097	NS
DG5	2.550	4.488	5.643	**.018**	**SS**
DG6	.603	.531	3.110	.078	NS
DG7	.712	.655	.566	.452	NS
DG8	.571	.505	.811	.368	NS
DG9	3.393	10.571	6.263	**.012**	**SS**
DG10	1.385	1.553	1.178	.278	NS

The diagnosis selected from this table, taking in consideration the direct risks (positive and arranged in descending order) and the results of the Chi-squared test are: DG9, DG5, DG1, DG10. We can see that the same diagnosis were selected also using the regressional model – only their order differs (at steps 2 and 3).

TABLE VIB
THE DIRECT RISKS FOR EACH DIAGNOSIS – OFF2

Diagnosis	The risk for the cohort OFF2 = 1	Odds Ratio (1/0)	Pearson Chi-squared	Asimpt. sig. p	
DG1	.419	.330	13.411	.000	SS
DG2	1.952	2.650	5.599	**.018**	**SS**
DG3	.888	.858	.069	.792	NS
DG4	2.469	4.306	5.295	**.021**	**SS**
DG5	.464	.397	.796	.372	NS
DG6	2.525	3.678	20.319	**.000**	**SS**
DG7	1.267	1.377	.462	.497	NS
DG8	1.450	1.675	.845	.358	NS
DG9			1.248	.264	NS
DG10	.235	.186	6.318	.012	SS

The diagnosis selected from this table, using the same criteria, are DG6, DG4, DG2 – in this order. DG10 was not selected, because its risk is not positive (being much smaller than 1); instead, DG10 has a statistically significant influence over the dependant variable OFF2 – and this probably is the reason for which the regressional model selected this diagnosis.

TABLE VIC
THE DIRECT RISKS FOR EACH DIAGNOSIS – OFF3

Diagnosis	The risk for the cohort OFF3 = 1	Odds Ratio (1/0)	Pearson Chi-squared	Asimpt. sig. p	
DG1	.774	.759	.352	.553	NS
DG2	1.065	1.070	.008	.930	NS
DG3	1.496	1.554	.321	.571	NS
DG4	1.556	1.625	.204	.652	NS
DG5	1.556	1.625	.204	.652	NS
DG6	2.260	2.440	3.898	**.048**	**SS**
DG7			2.051	.152	NS
DG8	.913	.907	.008	.927	NS
DG9			.318	.573	NS
DG10	.388	.369	.992	.319	NS

There are a few diagnosis with positive risks, but only one has also a significant result at the Chi-squared test; this is DG6, which is identically with the choice of the regressional model.

TABLE VID
THE DIRECT RISKS FOR EACH DIAGNOSIS – OFF4

Diagnosis	The risk for the cohort OFF4 = 1	Odds Ratio (1/0)	Pearson Chi-squared	Asimpt. sig. p	
DG1	1.398	1.519	2.002	.157	NS
DG2	.973	.966	.004	.947	NS
DG3	1.072	1.092	.023	.880	NS
DG4	1.131	1.169	.037	.848	NS
DG5			2.282	.131	NS
DG6	.492	.429	4.492	**.034**	**SS**
DG7	2.070	2.713	5.224	**.022**	**SS**
DG8	.664	.613	.408	.523	NS
DG9			.997	.318	NS
DG10	.913	.893	.056	.813	NS

There are also a few diagnosis with positive risks, and DG7 has a significant result at the Chi-squared test; the other diagnosis with positive risks failed at the Chi-squared test, and, even if DG6 passed the Chi-squared test, its risk is smaller than 1. This order is different again from the selection made by the regressional model, which uses also the diagnosis DG1 (found here with positive risk, but with not significant Chi-squared result).

V. CONCLUSION

It follows from these tables that our assumption is not entirely checked, because the binary logistic regressional model can find more significant predictors than the direct risks calculation. The reason for this situation is the way in which the predictor scores are calculated, which takes in consideration not only the predictors, but also the connections between them. We also must notice that, every time when a new predictor is added into the model, all scores are updated, and the analysis of the new possible predictors takes in consideration these new results.

Therefore, the binary logistic regressional model is a sensitive method to identify *sets of predictors*, which are related one to the other – being the best solution to study the internal links between parameters. Instead, every time when we want to analyze independent parameters, it is not necessary to use this model, being more accurate and easier to calculate only the relative risks and the corresponding Chi-squared statistics.

REFERENCES

[1] B.H.Munro, *Statistical Methods for Health Care Research.* Philadelphia, New York: Lippincott Press, 1997.
[2] D.W. Hosmer, S.Lemeshow, *Applied Logistic Regression.* New York: John Wiley & Sons, 2000.
[3] D.G. Kleinbaum, *Logistic Regression: A Self-Learning Text*. New York: Springer-Verlag, 1994.
[4] M. Norusis, SPSS *13.0 Statistical Procedures Companion.* Upper Saddle-River, N.J.: Prentice Hall, Inc. 2004.