METHODS FOR MULTIVARIATE DATA ANALYSIS IN THE STUDY OF ORAL DISEASES: THE MULTIPLE LINEAR REGRESSION

Cristina Gena DASCALU
„Grigore T. Popa" University of Medicine and Pharmacy, Iasi, Romania
Faculty of Dental Medicine, Medical Informatics and Biostatistics Dept.

METHODS FOR MULTIVARIATE DATA ANALYSIS IN THE STUDY OF ORAL DISEASES: THE MULTIPLE LINEAR REGRESSION (**Abstract**): During the statistical analysis of medical data, in many situations it is necessary to identify the multiple correlations established between the studied parameters. In this purpose, one of the most useful methods is to build a model of multiple regression, which allows the modeling of a dependant variable values having at least the ordinal type, based on its linear relation with more than one independent variables satisfying the same restriction, called predictors. The multiple linear regression model is a generalization of the simple linear regression model, which identifies the parameters of an equation with n variables $y = b_0+b_1x_1+b_2x_2+\ldots+ b_nx_n+e$, based on which we can find the predictors that have a statistically significant influence over the dependant variable. We used this model to identify, on a set of 202 patients having different types of oral lesions, the biochemical analysis which can be eventually correlated with the oral diagnosis. We found that the values of leucocytes, hemoglobin and hematocrit are significant for the general oral diagnosis, the cholesterol and glucose values for the oral lesion type, and the hemoglobin for the periodontal disease. The identified predictors are useful for further data processing.

Keywords: multivariate analysis, multiple linear regression, oral diagnosis

INTRODUCTION

In many cases, in scientific researches, it is necessary to analyze the correlations between more than two parameters, in order to detect the internal influences between data. In such cases it is absolutely necessary to choose the right method to study the assumed correlations – because the parameters nature defines in fact the path we are going to follow. There are a few main possibilities to analyze the multiple correlations between data: the regressional models, the principal components analysis, the discriminant analysis, the clustering. The regressional analysis is the easiest available method between these, which tries to find a very clear pattern for data variation: one parameter is interpreted as being "dependant", so it will be influenced in its variation by all the other parameters. The only problem that remains in this case are to find an appropriate mathematic model for this influences schema; the regressional analysis works with a few mathematic models: the linear regression, the curve estimation, the binary and the multinomial logistic models, the ordinal regression, the probit regression and the non-linear regression. The difference between these models consists not only in their mathematical fundaments, but, more important, in the data types for which there are available.

When all the studied parameters are quantitative and we intend to study in which way one of these parameters, defined as "dependant", is influenced by all the other, the easiest way to

quantify this influence is to build a model of linear multiple regression and to check how well it fits with the real studied phenomenon.

MATERIAL AND METHODS

The linear regression method is used when we need to model the values of a dependant variable according with the values of at least two independent variables, also called "predictors", using the equation of a straight line. The main requirement that must be fulfilled by all the variables involved in the model is that these variables must have at least the scale type – but the model behaves best when all the variables are quantitative (Draper, 1981).

The linear regression model assumes that there is a linear relationship between the dependent variable and each predictor, described in the following formula:

$$y_i = b_0 + b_1x_{i1} + ... + b_px_{ip} + e_i ,$$

where:

- y_i is the value of the i-th case of the dependent scale variable
- p is the number of predictors
- b_j is the value of the j-th coefficient, $j \in \{0, 1, ..., p\}$
- x_{ij} is the value of the i-th case of the j-th predictor
- e_i is the error in the observed value for the i-th case

We can notice that we are dealing here with an equation of 1st degree, with p variables; b_0 is the intercept or the model-predicted value of the dependent variable when the value of every predictor is equal to 0 (the point where the line intersects the Oy axes, in a representation using a Cartesian coordinates system). The error term e_i must fulfill also the following conditions (Draper, 1981):

- Its distribution is normal, with a mean of 0;
- Its variance is constant across cases and independent of the variables in the model;
- Its value for a given case is independent of the values of the variables in the model and of its values term for other cases.

When we build the multiple linear regression model, we must follow a few steps.

First, it is necessary to check the model fit (Norusis, 2004). In this purpose, the ANOVA test is used, and the F statistic is calculated; if the F value is statistically significant ($p \leq 0.05$), it follows that the model fits well with the analyzed data, and using it is better than guessing the mean.

Secondly, the correlation coefficients R and R squared are calculated; these coefficients show also how well the model works: R squared, for example, shows the percentage from the dependant variable's variation which is explained by the model.

Then, for each predictor involved in the model, we calculate its unstandardized and standardized coefficients and its significance level (expressed using the t statistic); in this way we can separate, from all the predictors involved in the model, only the significant ones – in order to eliminate further from the model the non-significant predictors (variables which do not contribute too much to the model). At this step we can also find the relative importance of each significant predictor, which varies proportionally with its standardized coefficient Beta.

At this step it is also good to calculate the part and partial correlation coefficients

(Norusis, 2004); these coefficients help us to detect the possible multicollinearity problems. Such problems appear when the part and the partial correlations are very different by the zero-order correlation – which means that a large amount in the variance of the dependant variable that is explained by the analyzed predictor is also explained by the other predictors – so the predictors are "collinear" and their effects are overlapped. Another coefficient calculated at this step is the tolerance – or the percentage of the variance in a given predictor that cannot be explained by the other predictors; small tolerances show that large amounts in the variance in a given predictor are explained also by the other predictors, so again the multicollinearity is present, and large tolerances show that the multicollinearity is absent. Finally the multicollinearity is also measured using a Variance Inflation Factor (VIF) – that regards the standard error of the regression coefficients; a VIF factor greater than 2 is usually considered problematic, being a clue for predictors multicollinearity.

There are also a few diagnostics tests especially designed for collinearity (Weisberg, 1985):

- in the predictors matrix, the eigenvalues are calculated; if these values are close to 0, it means that the predictors are highly intercorrelated, and therefore, small changes in the data values may lead to large changes in the estimates of the coefficients;
- in the same matrix, the condition indices are also computed, as the squared roots of the ratios of the largest eigenvalue to each successive eigenvalue; values greater than 15 indicate a possible problem with collinearity; greater than 30, a serious problem.

The last step of the analysis regards the collinearity removing, as long as this is possible. In order to do this, the easiest way is to rerun the model using the z scores for all the variables involved (predictors, as well as the dependant variable) instead of their direct values. In this way, only the most useful predictors will be included in the model.

RESULTS

On a set of 202 patients having different types of oral lesions we will try to identify the biochemical analysis which can be eventually correlated with the oral diagnosis using the multiple linear regression model.

In the first step of our analysis, we will use in our model the following variables as predictors: glucose, hemoglobin, monocytes, thrombocytes, lymphocytes, eosinophils, cholesterol, creatinine, leukocytes, red blood cells, neutrophils, hematocrit. The dependant variable is the oral diagnosis, defined as a scale variable which represents the sum of the following binary variables regarding the oral health status: dental mobility, gingival hyper growing, gingival retraction, decay lesions, periodontal disease. Therefore, the 0 value for this variable means health, and its value grows proportionally with the number of new identified symptoms in the oral area.

TABLE I. The ANOVA test for model's fitting (1st study)

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	36.791	12	3.066	**1.648**	**.087**
	Residual	232.484	125	1.860		**NS**
	Total	269.275	137			

TABLE II. The global correlation coefficients of the model (1st study)

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.370	.137	.054	1.364

The ANOVA test (table I) shows that the multiple linear regression model doesn't fit significantly with the analyzed data, but its level of significance is quite close to the threshold value, so it is useful to continue the analysis. The correlation coefficients R and R squared (table II) come to strengthen the previous observation, through their low values: only 13.7% from the oral diagnosis variation is covered by the regressional model.

Anyway, we will still try to identify which are the significant predictors of this model. Checking the significance levels of all the predictors (table III), we notice that the only significant predictors for the model are Leukocytes, Hemoglobin and Hematocrit. According to the Beta coefficient values, it follows that the most important predictor is Hematocrit (β = -0.661), followed by Hemoglobin (β = 0.584). The less important predictor is Leukocytes (β = 0.254). Checking the tolerance values, we can see that most of them are very different of 0; similarly, the VIF factor is bigger than 2 in only 5 cases from all 12 – so, there are not important problems concerning the predictors collinearity.

TABLE III. The predictors coefficients and significance levels (1st study)

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Correlations			Collinearity Statistics	
		B	Std. Error	Beta			Zero-order	Partial	Part	Tolerance	VIF
1	(Constant)	3.723	2.606		1.429	.156					
	Leukocytes	.182	.077	.254	2.366	**.020**	.188	.207	.197	.600	1.667
	Red blood cells	-.446	.433	-.141	-1.029	.306	-.106	-.092	-.086	.366	2.730
	Hemoglobin	.633	.259	.584	2.443	**.016**	-.075	.213	.203	.121	8.277
	Hematocrit	-.212	.083	-.661	-2.559	**.012**	-.130	-.223	-.213	.104	9.652
	Thrombocytes	.000	.001	-.020	-.209	.835	.057	-.019	-.017	.726	1.377
	Neutrophils	-.008	.021	-.065	-.383	.703	.041	-.034	-.032	.242	4.127
	Lymphocytes	-.014	.023	-.091	-.597	.552	-.062	-.053	-.050	.297	3.362
	Monocytes	-.068	.063	-.105	-1.079	.283	-.126	-.096	-.090	.733	1.364
	Eosinophile	-.011	.046	-.022	-.229	.819	-.003	-.020	-.019	.783	1.276
	Cholesterol	.001	.003	.024	.271	.787	.039	.024	.023	.895	1.117
	Creatinine	.508	.690	.084	.736	.463	.002	.066	.061	.536	1.866
	Glucose	.005	.004	.110	1.236	.219	.144	.110	.103	.877	1.140

We will repeat the analysis, in order to find the influence of the same set of predictors (glucose, hemoglobin, monocytes, thrombocytes, lymphocytes, eosinophils, cholesterol, creatinine, leukocytes, red blood cells, neutrophils, hematocrit) over another dependant

variable - oral lesion type. This variable also has the scale type, with 5 values corresponding to increased gravity diagnosis: 0 – no lesions; 1 – epithelial-conjunctive hyperplasia; 2 – chronic ulceration; 3 – prosthesis stomatitis; 4 – precancerous lesions.

TABLE IV. The ANOVA test for model's fitting (2nd study)

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	21.215	12	1.768	2.083	**.022**
	Residual	106.089	125	.849		**SS**
	Total	127.304	137			

TABLE V. The global correlation coefficients of the model (2nd study)

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.408	.167	.087	.921

This time, the ANOVA test (tab. IV) shows that the multiple linear regression model fits significantly with the analyzed data (p = 0.022 – SS), even if the correlation coefficients R and R squared (tab. V) still have low values: only 16.7% from the oral lesion type variation is covered by the regressional model.

TABLE VI. The predictors coefficients and significance levels (2nd study)

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Correlations			Collinearity Statistics	
		B	Std. Error	Beta			Zero-order	Partial	Part	Tolerance	VIF
1	(Constant)	2.372	1.760		1.347	.180					
	Leukocytes	-.069	.052	-.139	-1.318	.190	-.087	-.117	-.108	.600	1.667
	Red blood cells	.160	.293	.074	.546	.586	.005	.049	.045	.366	2.730
	Hemoglobin	-.103	.175	-.138	-.588	.558	.018	-.053	-.048	.121	8.277
	Hematocrit	.004	.056	.020	.080	.936	.020	.007	.007	.104	9.652
	Thrombocytes	.001	.001	.123	1.284	.202	.020	.114	.105	.726	1.377
	Neutrophils	-.016	.014	-.195	-1.175	.242	-.102	-.105	-.096	.242	4.127
	Lymphocytes	-.018	.016	-.173	-1.156	.250	-.016	-.103	-.094	.297	3.362
	Monocytes	.001	.042	.001	.014	.989	.086	.001	.001	.733	1.364
	Eosinophile	.050	.031	.149	1.618	.108	.235	.143	.132	.783	1.276
	Cholesterol	.006	.002	.214	2.478	**.015**	.153	.216	.202	.895	1.117
	Creatinine	.873	.466	.209	1.872	.064	.109	.165	.153	.536	1.866
	Glucose	-.007	.003	-.215	-2.462	**.015**	-.189	-.215	-.201	.877	1.140

The VIth table shows the significant predictors of the model: this time, only Cholesterol and Glucose have this property – followed by the Creatinine predictor, whose significance level is very close by the threshold (p = 0.064). The most important predictor is Glucose (β =

-0.215), but the other two are also very close: Cholesterol - $\beta = 0.214$ and Creatinine - $\beta = 0.209$. The tolerance values are again high, and the VIF factor is again bigger than 2 in 5 cases from all 12 (but not for the predictors identified as significant); therefore, we don't have again problems concerning the predictors collinearity.

CONCLUSIONS

The multiple linear regression model is useful to identify the correlations between more than two parameters; it allows to select from large set of predictors only the significant ones, and also to classify them according to their importance in the variance of the dependant variable. This method can be successfully used as a preliminary step in data analysis, because the model usually must be rebuilt – including only the significant predictors or, even better, studying only the simple regression line – with a single predictor and the dependant variable – which allows to identify exactly the corresponding correlation coefficients.

REFERENCES

1. Draper, N. R., H. Smith, Applied regression analysis, 2nd ed. John Wiley and Sons, New York,1981.
2. Norusis, M., SPSS 13.0 Guide to Data Analysis. Prentice Hall, Inc., Upper Saddle-River, N.J., 2004.
3. Norusis, M., SPSS 13.0 Statistical Procedures Companion. Prentice Hall, Inc., Upper Saddle-River, N.J., 2004.
4. Weisberg, S., Applied Linear Regression, 2nd ed. John Wiley and Sons, New York,1985.