Application of the Logistic Regression Model to an Inquiry of Respiratory Health

Carla Mónica Santos Dias Pereira
Cempa, Universidade Portucalense
4200 Porto, Portugal

Keywords: Logit, Odds-ratio, Goodness-of-fit, Missing values, Software package

The aim of this study is to identify the risk factors that could lead the doctor to make a diagnostic of asthma. We will try to describe a relationship between a dependent variable which we call response, reporting or not a diagnosis of asthma, and a set of independent variables, or covariates, of different kinds. Due to the fact that our response variable has only two levels, it will be used the logit transformation and the logistic regression model which is very useful in epidemiological studies and available in many software packages. However, beyond this technique, a set of problems connected with the inquiry will emerge, the most complex of which is the treatment of the missing values.

We are dealing with a sample of 363 subjects and 23 variables, collected by postal questionnaires. Due to redundancy problems, these variables were grouped in a set of 16 variables. We will begin with a careful univariate analysis of the independent variables in order to get the maximum information that can help us to identify the most important features and lead us to the "best model". This step involves the computation of some test statistics like Likelihood Ratio and Wald tests . The criteria used to choose the variables to be included in the model, were their biological importance and the use of the value 0,25 as a limit for the significance. So, conserning these tests we have decided to exclude three dichotomous variables. Moreover, biologically speaking, these variables didn't make much sense.

The multivariate analysis begins with the estimation of the coefficients of the model with all the variables, not excluded in the first step, and the assessment of its significance. In this step we found out that three dichotomous weren't significant, namely, ´sex´, ´went to doctor because of breathing difficulties´ and ´be unable to do daily activities´. Regarding nominal variables we have rejected four more, connected with wheezing, by the fact that all design variables associated with them showed p-values higher than 0,25 in both tests. The variable that describes ´smoking habits´ had only two of their four categories that exceeded 0,25. So we can not be sure about their contribution and we have decided not to exclude it. Another approach to select the variables was the use of Stepwise method. At this point the problem of missing values emerged. If we perform this procedure by selecting this method like it appears in most software packages, like SPSS, then by default the cases with at least one missing value, shall be deleted.

However, the way to get more information is to determine the p-values from the Likelihood ratio test, step by step. We have bear in mind this kind of problems and study imputation procedures to get better results that may lead us to a more realistic model. As the stepwise procedure revealed itself compatible with the previous analysis, we can say that the essential variables to the model are 'social class (x_1)', 'smoking habits (x_2)', 'wheezing with cold (x_5)', 'home peak expiratory flow meter', 'current medication (x_3)' and the continuos variable 'number of years with breathing difficulties (x_4)'. After we had obtained the variables that were considered important to the model the next step was to confirm the assumption of the linearity in the logit for the only continuous variable, but we came to the conclusion that there was no need to perform any kind of transformation. Afterwards we considered the need for inclusion of second order interaction terms for the variables selected in the previous step, using again the Stepwise method, and we found out that none was statistically significant.

Before concluding that we have reached the "best model", we need to compute diagnostic measures that may help us to assess if the model is the appropriate or if it exists some "bad" observations. At this stage we have to group the data into covariate patterns. We have calculated summary measures used to identify subjects poorly fitted, and others which provide our attention to subjects with great deal of influence on the values of the estimated parameters, like Chi-square and Deviance residuals, leverage and Cook's distance. Plots of these measures versus the estimated logistic probabilities showed that there were nine of the 248 covariate patterns, that were poorly fitted but that none of them was influent.

Before coming to some conclusions, based on the values of the odds ratio, we have decided to exclude the variable 'home expiratory flow meter' (not biologically important) because the results of the diagnostic without it were better (identical residuals but less influent measures). The final model was:

$$y = -1{,}134d1x_1 - 1{,}057d2x_1 - 1{,}039d3x_1 - 0{,}496d1x_2 + 0{,}092d2x_2 - 0{,}415d3x_2 - 1{,}161d4x_2 + 0{,}806d1x_3 + 1{,}999d2x_3 + 0{,}0557x_4 - 1{,}1855x_5$$

So, we can say that, for example, asthma diagnosis is 0,32 as frequent among those subjects who belong to the medium class than among those who belong to the lower one. For the continuous variable the odds ratio computed, for an increase of ten years, showed that the diagnosis of asthma occurs 1,7 as more often in subjects with breathing difficulties than among without them.

References

Collett, D. (1991) - Modelling Binary Data. Chapman & Hall.

Hosmer , D.W. and Lemeshow, S. (1989) - Applied Logistic Regression. John Wiley & Sons.

Hosmer D.W, Taber S. and Lemeshow, S. (1991) - The Importance of Assessing the Fit of Logistic Regression Models. *A.J.P.H.* Vol.81, N°13, pp.1630-1635.

Vach, W. (1994) - Logistic Regression With Missing Values in the Covariates. Springer-Verlag.