Study of a logistic model with mutually correlated variables using a generation algorithm of dichotomous data with arbitrary sensitivity, specificity and correlation

Noriaki Ikeda[a,b], Leon Bax[b], Osamu Henmi[b], Noritaka Mamorita[b] , Masuo Shirataka[b], Yasuo Morohoshi[c] , Sadayasu Shibata[a] , Harukazu Tsuruta[a,b], Akihiro Takeuchi[a,b]

[a] *Department of Medical Informatics, School of Allied Health Sciences, Kitasato University, Japan*
[b] *Graduate School of Medical Sciences, Kitasato University, Japan*
[c] *School of Medicine, Kitasato University, Japan*

Abstract

Parameters of a logistic model were studied when the descriptive variables are not mutually independent. For this purpose a generation algorithm of multiple test data with arbitrary sensitivity, specificity and correlation was developed, and the effect of correlation on the parameters of the model was investigated. Based on the results, possibility of a new modeling framework for logistic analysis was proposed.

Keywords:
Logistic model, Correlation, Generation algorithm.

Introduction

A logistic model

$$y = \log(p/(1-p)) = b_0 + \Sigma\, b_i\, x_i \qquad (1)$$

with dichotomous descriptive variables x_i are considered, where p designates the probability of an event such as a particular disease. If the variables are mutually independent, we have the relations

$$b_i = \log(\text{Odds ratio of the i-th variable}) \qquad (2)$$

for the coefficients of the model. However, this relation does not hold when the variable x_i is correlated with some other variables. The purpose of this study is to examine how the correlation between the descriptive variables affects the values of coefficients b_i.

Methods

A generation algorithm of multiple test data with arbitrary sensitivity, specificity and correlation was developed [1], and was implemented using MATLAB 7.1 (The MathWorks, Inc.). Various numerical data were created, e.g. mutually independent tests with the same sensitivity and specificity, data with a positive correlation coefficient between two variables, data with a negative correlation coefficient, and data with correlations among many variables. Logistic analysis was performed on each case and the estimated model parameters, b_i, were examined. For the estimation of the model, the "logistic procedure" of SAS 9.1 (SAS Institute Inc.) and the "glmfit" function of MATLAB were employed.
The relationship between disease D and its clinical test x_i can be presented by a contingency table, Table 1. Variable D represents the state of the patient, having the disease by D=1 and not having the disease by D=0. Variable x_i represents the result of the i-th test, e.g., positive by x_i =1 and negative by x_i =0. The sensitivity and specificity of the test are represented by α_i and β_i, respectively. When D=1 the correlation coefficient between test i and test j is r_{ij}^+, and when D=0 the correlation coefficient between tests is r_{ij}^- .

Table 1: Contingency table of test x_i and the diagnosis D (=1: positive, =0: negative)

	D=1	D=0
$x_i = 1$	α_i	$1-\beta_i$
$x_i = 0$	$1-\alpha_i$	β_i

Results

Test data was created by the generation program [1] and precisely reconstructed the given conditions of sensitivity, specificity and correlation coefficients for arbitrary number of tests. Effect of correlation on the parameters of the logistic model was examined by using these data. In the following the number of variables in the model was set to 4, and the interaction model

$$y = \log(p/(1-p)) = b_0 + \Sigma\, b_i\, x_i + \Sigma\, b_{ij}\, x_i\, x_j \qquad (3)$$

was assumed for SAS procedure.

A All variables are mutually independent

We assumed the following conditions of sensitivities and specificities and no correlation among the variables:

$$\alpha = \beta = (0.80, 0.75, 0.70, 0.65).$$

By applying the logistic procedure we get the model parameters b_i's that satisfy the relation (2).

B Positive correlations between two variables

We assume that $\alpha = \beta = (0.7, 0.7, 0.7, 0.7)$ and $r_{ij}^{+} = r_{ij}^{-} = 0$ except for $r_{12}^{+} = r_{12}^{-} = 0.3$. The result is that the coefficients b_1 and b_2 were decreased due to the correlation between tests 1 and 2.

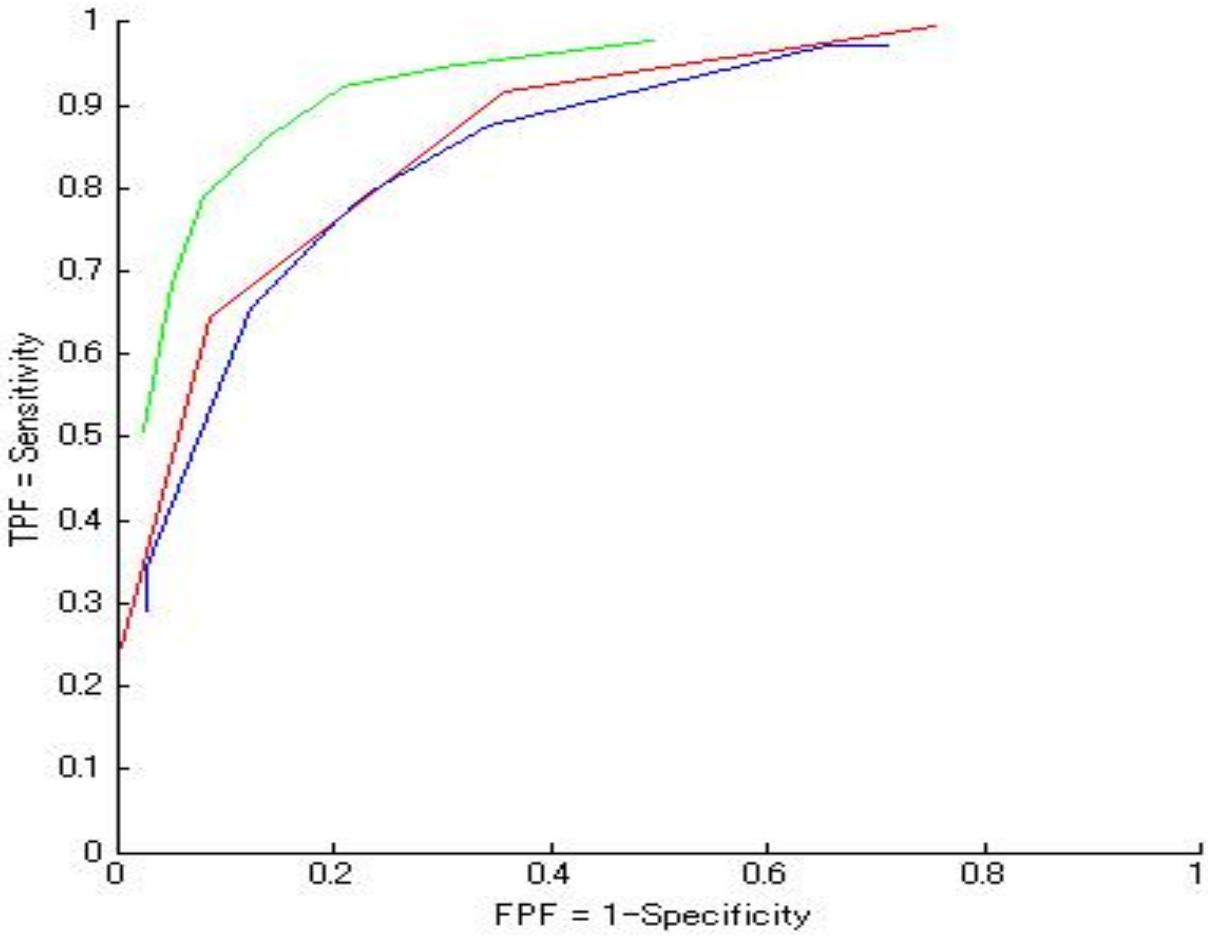
Parameter	Estimated value	SE	p
b_0	-2.991	0.142	<0.0001
b_1	1.288	0.123	<0.0001
b_2	1.288	0.123	<0.0001
b_3	1.702	0.120	<0.0001
b_4	1.702	0.120	<0.0001
b_{ij}	0.000	0.000	1.0000

C Negative correlations between two variables

We assume that $\alpha = \beta = (0.7, 0.7, 0.7, 0.7)$ and $r_{ij}^{+} = r_{ij}^{-} = 0$ except for $r_{12}^{+} = r_{12}^{-} = -0.3$. The result is that the coefficients b_1 and b_2 increased due to the correlation between tests 1 and 2.

Parameter	Estimated value	SE	p
b_0	-4.444	0.207	<0.0001
b_1	2.760	0.167	<0.0001
b_2	2.760	0.167	<0.0001
b_3	1.685	0.131	<0.0001
b_4	1.685	0.131	<0.0001
b_{ij}	0.000	0.000	1.0000

Fig.1 ROC curve of diagnosis for cases A: independent (red), B: positive correlations (blue), and C: negative correlations (green). In all cases it was assumed that $\alpha = \beta = (0.7, 0.7, 0.7, 0.7)$.



D More complex correlations

We assume that $\alpha = \beta = (0.7, 0.7, 0.7, 0.7)$ and $r_{ij}^{+} = r_{ij}^{-} = 0$ except for $r_{12}^{+} = r_{13}^{+} = r_{12}^{-} = 0.3$. The result is that the coefficients b_1, b_2 and b_3 decrease due to the correlation between tests 1 and 2, and tests 1 and 3. Parameter b_1 is not statistically significant and the interaction terms are quite complicated.

Parameter	Estimated value	SE	p
b_0	-2.487	0.190	<0.0001
b_1	0.012	0.282	0.9664
b_2	1.508	0.242	<0.0001
b_3	0.936	0.236	<0.0001
b_4	1.612	0.221	<0.0001
b_{12}	0.227	0.256	0.3769
b_{13}	1.719	0.258	<0.0001
b_{14}	0.086	0.252	0.7336
b_{23}	-0.535	0.255	0.0360
b_{24}	0.037	0.251	0.8829
b_{34}	0.098	0.244	0.6868

Discussion

Our results indicate that there is a need for a new modeling framework of logistic analysis, if we wish to separate the effect of each variable from their interactions, and to use the model parameters b_i for the diagnostic or risk evaluation purposes. Based on the theory of joint probability of dichotomous tests with correlations [2], we could show how the bi's are affected by the correlation coefficients between xi's.

Conclusion

The effects of correlation on a logistic model with dichotomous variables were examined. Correlation between explanatory variables made the coefficients of logistic model difficult to interpret. Based on the results, we propose a possible new modeling framework for logistic analyses.

References

[1] Ikeda N et al, Diagnostic performance of combined tests using a generation algorithm of multiple tests with arbitrary sensitivity, specificity and correlation. 3rd Int Conf on Medical Signal Processing (MEDSIP 2006).

[2] Bahadur RR, A representation of the joint distribution of responses to n dichotomous items. In H. Solomon, editor, Studies in Item Analysis and Prediction, pp.158-176, Stanford University Press, 1966.

Address for correspondence

ikeda@kitasato-u.ac.jp, Kitasato, Sagamihara, 228-8555, Japan