Logistic Regression Model for Ordered Categorical
Response and Its Applications to Quantitative
Structure-Activity Relationship

Takenobu TASAKI and Masashi GOTO

(SHIONOGI Kaiseki Center)

SUMMARY

As a technique to evaluate the relationship between an ordered categorical bioactivity and physico-chemical descriptors in analysis of quantitative structure-activity relationship, ALS (Adaptive Least Square) method now draws much attention in drug designs. This procedure is a special technique mainly composed of algorithms which optimize the classification capability concerning given training sample. For the analysis of this kind of data, we propose an application of logistic regression model as an alternative approach. We investigate its behavior using several sets of data and find out its many advantages over ALS method. We also discuss a few graphical techniques which assist the model diagnosis.

*KEYWORDS: Quantitative structure-activity relationship;
Ordered categorical response; Logistic regression model;
model diagnosis*

1. Introduction

In this paper, we propose an application of the logistic regression model to analysis of data in which bioactivity is given as ordered categorical classification (for example, efficacy evaluation by ranking into three grades, $-$, $\pm$ and $+$), and discuss diagnostic procedures for the model along with the data which are important in the regression analysis. We also examine the applicability of the extended logistic regression analysis by using the data in two actual examples.

2. Description of a model and its estimation procedure

Let us denote by vector X p predictor variables (structure indexes), and by Y a criterion variable (a bioactivity) which takes a value in one of m ordered categories. Let us denote the ordered categories by integers of 1 to m . Now, we evaluate the relationship between X and Y by using the following probabilistic model:

$$\text{prob}(Y \leq s) = \frac{\exp(\theta_s - \beta'X)}{1 + \exp(\theta_s - \beta'X)} \quad (1)$$

$$(s=0, 1, \dots, m)$$

where $-\infty = \theta_0 < \theta_1 \leq \theta_2 \leq \dots \leq \theta_{m-1} < \theta_m = +\infty$, β is a "regression" coefficient vector and $\theta_1, \theta_2, \dots, \theta_{m-1}$ are

division points of categories on the transformed logistic scale. Estimates of these $(p+m-1)$ parameters are obtained by maximum likelihood method.

The main purposes of this extended logistic regression model is (1) to evaluate the relationship between ordered categorical variable and several prediction variables, (2) to apply to discrimination among ordered populations and (3) to predict a value of an underlying continuous variable for the response which is observed as an ordered categorical expression (Anderson & Philips, 1981).

Likelihood function

Let us denote n independent observations by $(y_1; x_1)$, $(y_2; x_2)$, $\dots, (y_n; x_n)$ of the response Y under the condition that $X=x$ are given. Then the likelihood of model (1) is given by

$$L(\beta, \Theta) = \prod \text{prob}(Y=y_i | X=x_i) \\ = \prod \left[\frac{\exp(\theta_{y_i} - \beta' x_i)}{1 + \exp(\theta_{y_i} - \beta' x_i)} - \frac{\exp(\theta_{y_{i-1}} - \beta' x_i)}{1 + \exp(\theta_{y_{i-1}} - \beta' x_i)} \right] \quad (2)$$

where $\Theta' = (\theta_1, \theta_2, \dots, \theta_{m-1})$. We denote by $l(\beta, \Theta)$, the log-likelihood of (β, Θ) , that is, $l(\beta, \Theta) = \ln \{ L(\beta, \Theta) \}$.

Iterative procedure for estimation of (β, Θ)

In order to maximize the log-likelihood function $l(\beta, \Theta)$ with respect to (β, Θ) , we use Newton-Raphson method. We denote by $\hat{\beta}$, $\hat{\Theta}$ the parameter values at which the log-likelihood

function attains to its maximum. As initial estimates of parameters for iterations, we choose

$$\beta_j=0; j=1, \dots, p,$$

and

$$\theta_s = \ln \left[\frac{\sum_{t=1}^s n_t + 1/2}{n - \sum_{t=1}^s n_t + 1/2} \right] ; s=1, \dots, m-1$$

where n_t is the number of observations fallen into the t -th category, or $n_t = \# \{ i; y_i = t \}$. In the course of iterations, if the estimated values $\{ \hat{\theta}_s \}$ do not satisfy the ordered relation : $\hat{\theta}_1 \leq \hat{\theta}_2 \leq \dots \leq \hat{\theta}_{m-1}$, then we adopt the initial values of Θ instead of the current $\hat{\Theta}$ at this iteration stage.

Prediction of Y

To predict a value of the response variable Y, the following two procedures are used :

(a) As the predicted value of Y, we choose such an integer s that $p_s^* = \max \{ p_i^* \}$, where

$$p_i^* = \frac{\exp(\theta_i - \beta'x)}{1 + \exp(\theta_i - \beta'x)} - \frac{\exp(\theta_{i-1} - \beta'x)}{1 + \exp(\theta_{i-1} - \beta'x)}$$

$$; i=1, \dots, m$$

where p_i^* is the estimated posterior probability of $Y=i$, given $X=x$ by substituting maximum likelihood estimates $\hat{\beta}, \hat{\Theta}$. Thus we denote the predicted value by $\tilde{y}$, that is, $\tilde{y}=s$.

(b) We calculate $z = \hat{\beta}'x$, and use such a integer s that satisfies $\hat{\theta}_{s-1} \leq z < \hat{\theta}_s$. Thus we write $\hat{y} = s$.

$\tilde{y}$ is more favorable for discrimination problem than $\hat{y}$, while $\hat{y}$ is useful to interpret because of its link with the idea of an underlying scale of measurement z .

3. Applications

The model described in section 2 was applied to two sets of data; one on β -adrenergic blocking effect of 1,2-Benzisoxazolyloxy-Propanolamine derivatives and another on inhibitory activity to photosynthesis of phenols and the analogues. Hereafter the former is called data of β -blocking activity and the latter data of photosynthesis inhibition activity. In practice, these bioactivities were observed as continuous variables and have been analyzed by usual linear regression analysis (Hansch-Fujita method). These two examples were chosen in order to evaluate the capability of the analytical method proposed in this paper by comparing with the results of the regression analysis.

Data of β -blocking activity: As to 37 compounds, pA_2 value was determined by van Rossum method for bioactivity and six values, π_1 , F_1 , R_1 , H_3 , L_1 and LM_1 were determined for structure descriptors. In linear regression analysis [π_1^2 ,

π_1 , R_1 , H_3 , LM_1 } was taken as an optimal subset of predictor variables (Yamakawa et al., 1981).

Data of photosynthesis inhibition activity: As to 40 compounds, pI_{50} value was determined for bioactivity and four values, $\log P$, $-E(HO)$, $-E(HOA)$ and pka , for structure descriptors. In 14 compounds pka was not determined. In four compounds pI_{50} was determined only to be "lower than 3.0".

Table 1 Estimates of regression coefficients

β -blocking activity n=37	Linear	Logistic
π_1^2	-0.380	-1.785
π_1	0.577	2.578
R_1	0.956	7.305
H_3	-0.384	-1.481
LM_1	-0.659	-3.284
correct agreement rate(%)		62.16
determination coefficient($\times 100$)	77.77	
Photosynthesis inhibition activity n=40	Linear*	Logistic
$(\log P)^2$	-0.075	-0.421
$\log P$	0.836	3.895
$-E(HOA)$	-0.488	-1.597
correct agreement rate(%)		57.50
determination coefficient($\times 100$)	47.59	

+ Excluding incomplete observations (n=36)

Table 2 Correlation coefficients among observation y , linear prediction $\hat{y}$ and logistic score z .

β -blocking activity; $\text{corr}(y, \hat{y})=0.882$

categorization										corr.	
numbers	division points			sizes			(z,y)	($\hat{y}$,z)			
4	6.61	7.49	8.45	9	9	10	9	0.859	0.974		
4	6.97	7.82	8.90	11	12	11	3	0.880	0.998		
3	6.97	7.82		11	12	14		0.877	0.995		
3	6.97	8.90		11	23	3		0.852	0.966		
2	7.49			18	19			0.804	0.912		
2	8.90			34	3			0.606	0.687		

Photosynthesis inhibition activity; $\text{corr}(y, \hat{y})=0.690$

numbers	categorization								corr.	
	division points			sizes					(z,y)	($\hat{y}$,z)
4	3.000	3.854	4.523	12	10	8	10	0.684	0.992	
4	3.000	3.854	4.886	12	10	14	4	0.688	0.997	
3	3.000	4.174		12	13	15		0.672	0.974	
3	3.000	4.886		12	24	4		0.670	0.971	
2	3.854			22	18			0.684	0.991	
2	4.886			36	4			0.678	0.983	

In linear regression analysis, $\{ (\log P)^2, \log P, -E(\text{HOA}) \}$ was taken as an optimal subset (Ezumi et al., 1983).

In fitting of logistic regression model, the continuous

(a)

5.	45
	93 96
6.	26 35
	52 52 59 61 78 97
7.	29 31 32 35 39 45 49
	66 71 73 81 82
8.	00 10 19 36 45
	53 61 69 79 84 90
9.	01 02 27

(b)

1.	886
2.	260 260
	509 523 721 959 959
	< 3 < 3 < 3 < 3
3.	131 208 215 292
	523 553 585 770 824 854
4.	018 167 174 292 319 495
	523 523 638 698 721 824 886
5.	000 149 149 301

Figure 1 Stem-and-leaf display of observations

(a) β -blocking activity (pA_2)

(b) Photosynthesis inhibition activity (pI_{50})

scale of bioactivity given was first divided into four portions to make each portion include approximately equal numbers of categorial samples. The distributions of observation of bioactivity are shown in Figure 1(a) and (b). The chosen division points were as follows:

Data of β -blocking activity (pA_2): 6.61, 7.49, 8.45

Data of photosynthesis inhibition activity (pI_{50}):

3, 3.854, 4.523

Table 1 compares estimates of regression coefficients by logistic regression analysis using these ordered categorical classifications with those by linear regression analysis using the original continuous observations. Correct agreement rates in logistic regression analysis and determination coefficients in multiple regression analysis are also shown. These results show that the regression coefficients in two analyses are approximately proportional for each data. The ± 1 difference agreement rate was 97.30% in data of β -blocking activity and 87.50% in data of photosynthesis inhibition activity. Table 2 shows correlation coefficients between logistic score (z) in logistic regression analysis and linear regression prediction when we used various categorizations. Figure 2(a), 2(b) and 2(c) show scatter diagrams of observations, linear regression predictions, and logistic scores in data of β -blocking activity. It is thought from these results that the lose of information resulting from categorizing continuous observations and applying logistic regression analysis to them is not essential, and that usefulness of the method we proposed in this

study is suggested. The logistic regression analysis will be a effective tool particularly for analyzing data which include incomplete observations, such as "lower than" as in the data of photosynthesis inhibition activity, or "higher than".

Figure 2 Scatter plots

- (a) observations vs.
linear predictions
- (b) observations vs.
logistic scores
- (c) logistic scores vs.
linear predictions

4. Discussion

The major problems in applying usual discriminant analysis or canonical discriminant analysis to a prediction of the ordered categorical response are that ordered relation can not be taken up explicitly as analytical restrictions and that plural axes are needed (however, the former problem can be treated also by manipulation to depress the cost resulting from erroneous classification to nearer category lower than that to farther category.)

The adaptive least square (ALS) method which was proposed by Moriguchi(1980) has been widely used to solve these difficulties. The logistic regression analysis proposed in this paper is an alternative to the ALS method and superior to it in its clear formulation of the relationship between an ordered categorical response and predictor variables. One of the advantages of the formulation of a model structure is that evaluation of fitness and identification of outlier and influential observation become easy (Tsaski and Goto, 1984). For example, Figure 3 shows how maximum log-likelihood changes when one sample is excluded in fitting of three variables regression model to data of photosynthesis inhibition activity. The figure shows that the fitness is greatly improved by excluding the 22nd compound. The appropriateness

of the logistic transformation can be evaluated by the following class of models which includes a model given by the equation (1) (Inoue and Goto, 1983).

$$P \equiv \text{prob}(Y \leq s)$$

$$T(P|\lambda) = \theta_s - \beta' X$$

$$T(P|\lambda) \equiv \ln \left[\frac{1}{\lambda} \left\{ \left(\frac{1}{1-P} \right) - 1 \right\} \right], \lambda > 0$$

$$\ln \left\{ \ln \left(\frac{1}{1-P} \right) \right\}, \lambda = 0.$$

Figure 3 log-likelihood sensitivity plot for data of photosynthesis inhibition activity.

This regression model based on asymmetric family of power transformations is reduced to the logistic regression model of this paper when $\lambda=1$. The optimum power exponent λ is estimated to be 2.0 in fitting of three variables model to the above-mentioned data of photosynthesis inhibition activity excluding one compound. The maximum log-likelihood at this time is -32.274, being almost equal to the corresponding value in logistic regression model, -32.510. Same values were obtained for correct and ± 1 difference agreement rates (64.10% and 89.74%, respectively). These results indicate that logistic scale is as favorable the optimum power transformed scale for the present data.

Quadratic terms such as π_1^2 and $(\log P)^2$ are included in the model which was fitted to the two sets of data analyzed in this paper. Although the necessity of terms of high orders can be also evaluated regular tests (for example, likelihood ratio test), it is more important to explore visually. The extended types of Copas's p vs. x plot (Copas, 1983) are shown in Figure 4(a) and (b). In the figures, the curves with indicator 1 represent smoothed estimates of prob $(Y>1)$, the curves with indicator 2 that of prob $(Y>2)$ and the curves with indicator 3 prob $(Y>3)$. The broken lines represent the corresponding fitted logistic curves of quadratic

(a)

(b)

Figure 4 Plotting p against x for ordered categorical response

(a) β -blocking activity (width=0.75)

(b) Photosynthesis inhibition activity (width=1.00)

Figure 5 Dendrogram by AID for data of β -blocking activity

model. These data suggest that a quadratic model is fitted well in the first set, and that observations at larger log P values are needed to establish the effect of $(\log P)^2$ in the second set.

As to both linear and logistic models, detection of interaction term and stratification factor is important in application of the regression models. Automatic interaction detector (AID) is one of the helpful means of approaching to this problem. Figure 5 shows an example of its application using original observations. It is necessary to classify predictor variables into categories in order to apply AID. How to treat data of ordinal scale is another unanswered question. We are now studying these problems.

REFERENCES

- Anderson, J.A. and Philips, P.R. (1981). Regression, discrimination and measurement model for ordered categorical variables. *Appl. Statist.*, 30, 22-31.
- Copas, J.B. (1983). Plotting p against x. *Appl. Statist.*, 32, 25-31.
- Ezumi, K. et al. (1983). QSAR for inhibitory activity to photosynthesis of phenols and analogues. *Proc. of The 10th QSAR Symposium*, 220-225.
- Inoue, T. and Goto, M. (1983). Power transformation of counting data and its applications. *Proc. of The 5th Applied Statistics Symposium*.
- Moriguchi, I. (1980). Adaptive least squares method. *IGAKU-*

NO-AYUMI, 113, 909-914.

Tasaki,T. and Goto,M. (1984). Diagnoses of logistic models.
Proc. of The 52th JSA Annual Meeting, 71-72.

Yamakawa,M. et al. (1981). QSAR for β -adrenergic blocking
activity of 1,2-Benzisoxazolyloxy-Propanolamine deriva-
tives. Proc. of The 3rd Medicinal Chemistry Symposium,
50-53.