P05-26

"Development of a state-of-the-art multiple regression KY-method corresponding to the big data era and its application to fish toxicity"

Kohtaro Yuta

In Silico Data, Ltd., (http://www.insilicodata.com)

Introduction: Even when the number of samples to be analyzed is extremely large, we have developed a multiple regression method which has high analytical reliability. In the case of evaluating the toxicity of a compound using a computer, a multivariate analysis / pattern recognition method is applied. However the conventional multivariate analysis / pattern recognition method is not designed to handle a large amount of sample data such as big data. We developed the "multiple regression KY-method" as a data analysis method corresponding to the big data era and analyzed fish toxicity data.

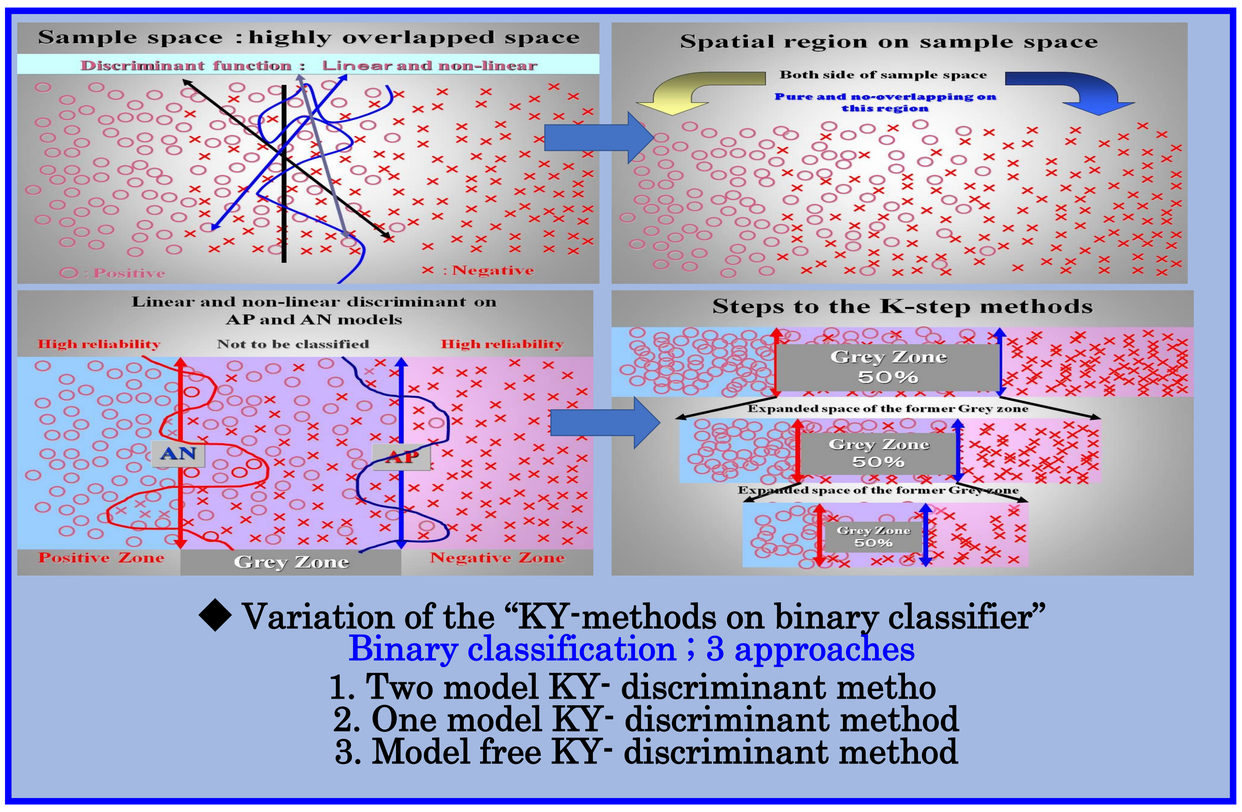
Method: In the KY-method, the sample group to be analyzed is divided into two groups. One is an in-lier group and the other is an out-lier sample group. The correlation coefficient (R) value of the in-lier sample group is greatly improved compared with the normal method.

What is the KY (K-step Yard sampling) - methods

◆ Binary classification ◆

1. Constantly achieve perfect (100%) classification under any conditions
   - Highly overlapped class sample data set
   - Quite large number of sample data set (tens and several thousands of)
2. Starting sample set was divided into
   - small and clean sample set
   - small and hierarchical samples

→ Repeat these operation, until all samples are correctly classified



◆ Variation of the "KY-methods on binary classifier"

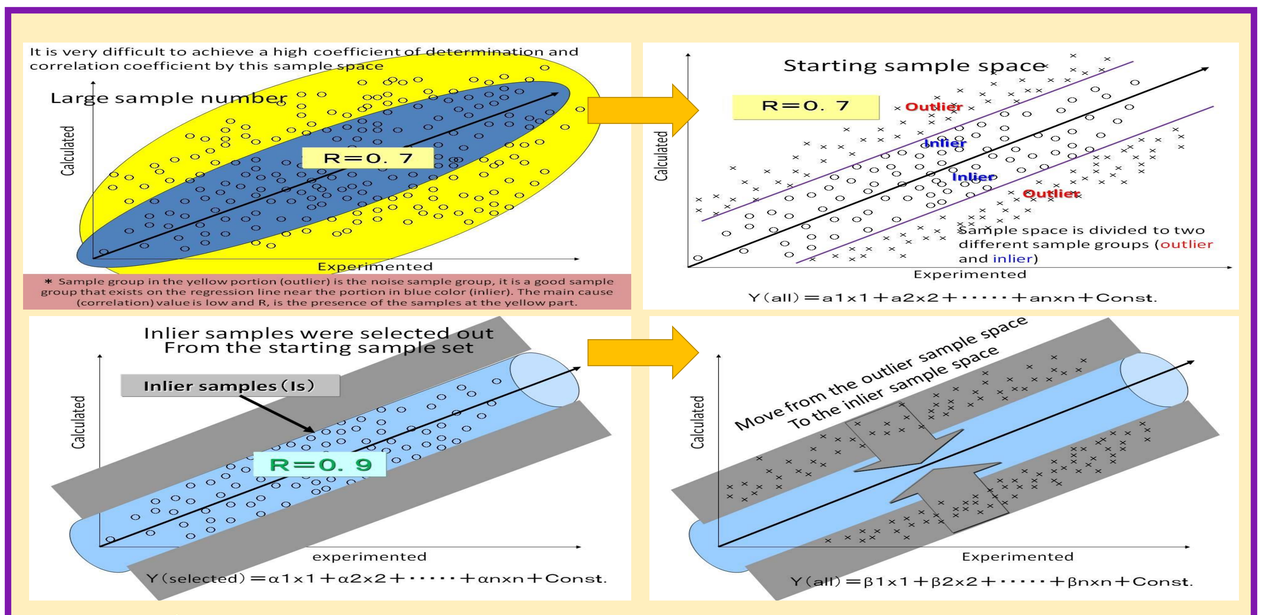
Binary classification ; 3 approaches

1. Two model KY- discriminant metho
2. One model KY- discriminant method
3. Model free KY- discriminant method

◆ Fitting : Regression analysis ◆

1. Constantly achieve high correlation and high decision coefficient under any conditions
   - Widely distributed sample data space
   - Quite large number of sample data set (tens and several thousands of)
2. Starting sample set was divided into
   - 'inlier' and 'outlier' sample set
   - small and hierarchical sample set

→ Repeat these calculation, until no more can this operation



◆ Variation of the "KY-methods on regression methods"

Fitting (multi regression); 3 approaches

1. KY-fitting with discriminant method
2. Three zone KY-fitting method
3. Model free KY-fitting method

Application to fish toxicity

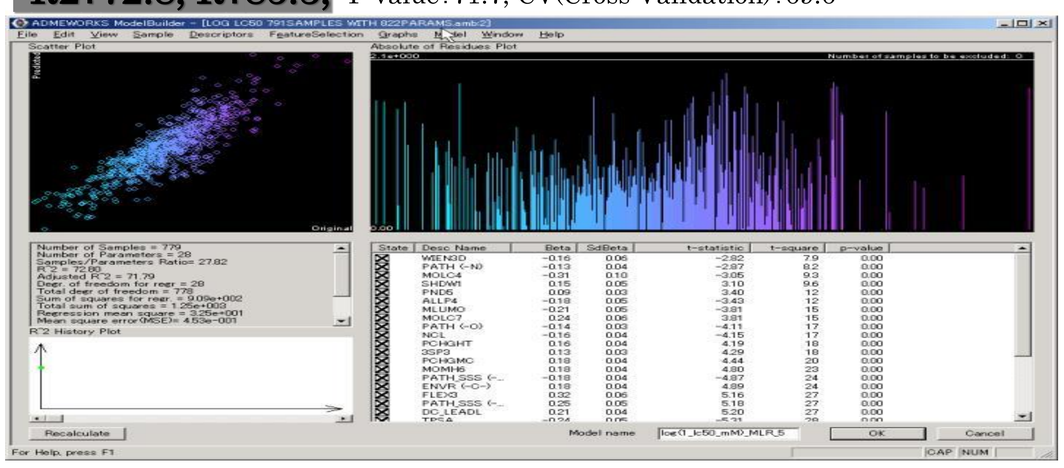
Original sample data;

Fish: 96 hours LC50、Number of samples: 791、Log(1/LC50_Mm) (Max/Min) : 6.376 / -2.963

◇ Data analysis by ordinal linear regression

Number of samples : 779, Number of used parameters : 28, Reliability ratio : 27.8

R2 : 72.8, R : 85.3, F-value : 71.7, CV(Cross Validation) : 69.6

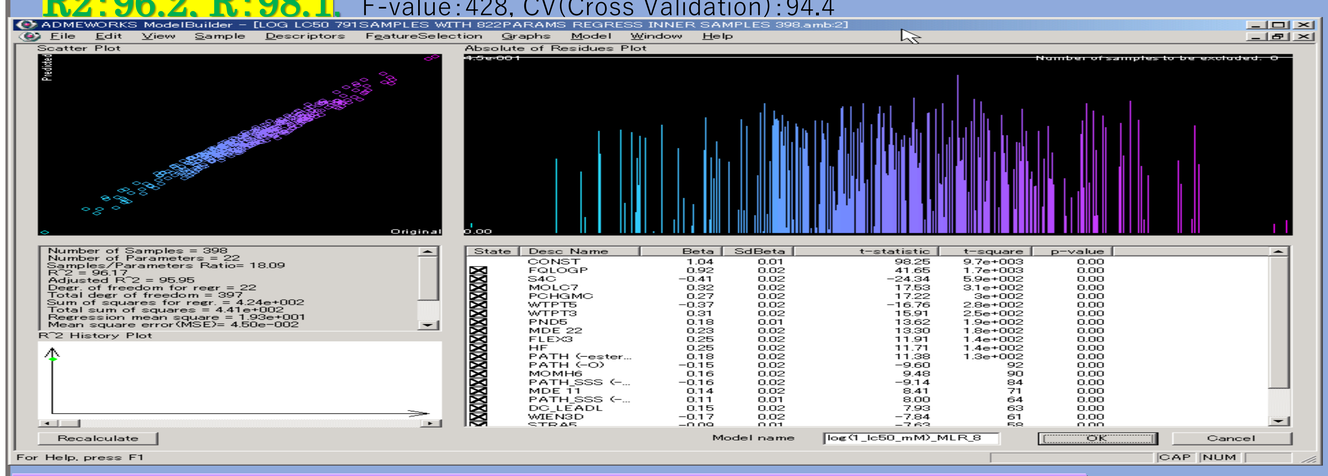


◇Fitting KY- method Step1 (Inner sample set)

Step1 : Inner sample set

Number of samples : 398, Used parameters : 22, Reliability ratio : 18.1,

R2 : 96.2, R : 98.1, F-value : 428, CV(Cross Validation) : 94.4



Experiment and results: Fish toxicity data analysis was performed by two methods. One was normal linear multiple regression and the other was the regression KY-method.

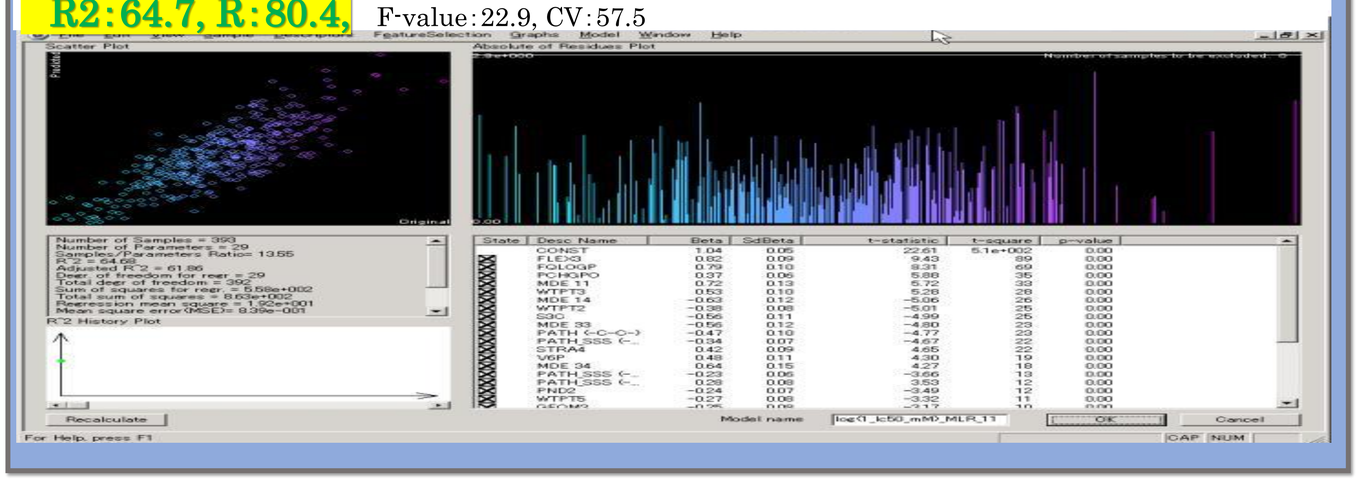
1. Result by the normal linear multiple regression method using all samples.
   Number of total samples: 779, Number of parameters: 28, Reliability index: 27.8,
   R: 85.3, R2: 72.8, F value: 71.7, CV: 69.6
2. Results by the "multiple regression KY-method".
   a) Number of in-lier samples; 398, Number of parameters: 22, Reliability index: 18.1,
   R; 98.1, R2; 96.2, F value: 428, CV: 94.4
   b) Number of out-lier samples: 393, Number of parameters: 29, Reliability index: 13.6,
   R: 80.4, R2: 64.7, F value: 22.9, CV: 57.5

◇Fitting KY method Step1 (Outer sample set)

Step1 : Outer sample set

Number of samples : 393, Used parameters : 29, Confidance ratio : 13.6,

R2 : 64.7, R : 80.4, F-value : 22.9, CV : 57.5



Summary: Currently, we have entered the era of big data where the number of samples extremely large. However the data analysis method currently deployed can not correspond to the big data era. Therefore, it is necessary to develop new and state-of-the-art data analysis methods. The multiple regression KY-method discussed in this poster is developed as a data analysis method corresponding to the big data era.