



"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 outlier 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 \diamondsuit Fitting : Regression analysis Constantly achieve perfect (100%) classification under any conditions Highly overlapped class sample data set 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 • Quite large number of sample data set (tens and several thousands of) 2. Starting sample set was divided into **Repeat these operation, until** Repeat these calculation, until no more can this operation small and clean sample set small and hierarchical samples 'inlier' and 'outlier' sample set small and hierarchical sample set all samples are correctly classified Sample space : highly overlapped space **Spatial region on sample space** It is very difficult to achieve a high coefficient of determination and Starting sample space Discriminant function: Linear and non-linear correlation coefficient by this sample sp Both side of sample space **Pure and no-overlapping on** Large sample number R = 0.R=0.7 Experimented Experimented utlier) is the noise sample group, it is a good sampl group that exists on the regression line near the portion in blue color (inlier). The main cause ····+anxn+Const. Y(all) = a1x1 + a2x2 +Linear and non-linear discriminant on Steps to the K-step methods **AP and AN models** Inlier samples were selected out From the starting sample set **High** reliability Not to be classified High reliability Grey Zone To the inlier sample : Nove from the out 50% Inlier samples (Is) Grey Zone 50% **Positive Zone** Experimented experimented $Y(selected) = \alpha 1 \times 1 + \alpha 2 \times 2 + \cdots + \alpha n \times n + Const.$ $Y(all) = \beta 1 \times 1 + \beta 2 \times 2 + \cdots + \beta n \times n + Const.$ Variation of the "KY-methods on binary classifier"

Variation of the "KY-methods on regression methods"

Binary classification; 3 approaches 1. Two model KY- discriminant metho 2. One model KY- discriminant method 3. Model free KY- discriminant method

Fitting (multi regression); 3 approaches 1. KY-fitting with discriminant method 2. Three zone KY-fitting method 3. Model free KY-fitting method

<u>- 🗆 ×</u>

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

Scatter Plot Absolute of Residues Plot		ř.		N	lumber of sample	a to be excluded: 0	
				N	iumbet of sample.	a to be excluded: 0	
	2.1e+000 Number of samples to be excluded: 0						
Number of Samples = 779 State Desc Name Number of Parameters = 28 Image: Samples/Parameters Ratio= 27.82 Image: State Image: State Samples/Parameters Ratio= 27.82 Image: State Image: State Image: State R 2 = 72.80 Image: State Image: State Image: State Adjusted R 2 = 71.79 Image: State Image: State Image: State Degr. of freedom for regr = 28 Image: State Image: State Image: State Tatel Image: State Image: State Image: State	Beta -0.16 -0.13 -0.31 0.15 0.15	SdBeta 0.06 0.04 0.10 0.05	t-statistic t- -2.92 -2.97 -3.05 3.10	square 79 82 93 96	p-value 0.00 0.00 0.00 0.00 0.00		
ium of squares for regr. # 9.05e+002 ALLP4 lotal sum of squares 1 25e+003 MLUMO Regression mean square = 3.25e+001 MLUMO dean square error (MSE)= 453e-001 MLUMO '2 History Plot NCL PATH (-0) NCL PCHGHT 3SP3 PCHGHT	-018 -021 024 -014 -016 016 013 019	0.05 0.05 0.05 0.06 0.03 0.04 0.04 0.04	-3.43 -3.81 3.81 -4.11 -4.15 4.19 4.29	12 15 15 17 17 18 18	0.00 0.		
PATH_SSS (0.18 -0.18 0.18 0.32 0.25 0.21 -0.24	0.04 0.04 0.04 0.06 0.05 0.05	4.80 -4.87 4.89 5.16 5.18 5.20 -5.21	23 24 24 27 27 27 27	0.00 0.	*	
Recalculate	Mo	del name	log (1_lc50_mM0_MLR_)	5 [OK	Cancel	

Solution Step (Inner sample set)

F-value: 428, CV(Cross Validation): 94.4

solute of Residues PI

Desc Name

CONST

MOLC7

WTPT5

WTPT3

MDE 22

FLE×3

PND5

FQLOGP

PCHGM0

Step1: Inner sample set

96.2. R:98.

Number of Samples = 398 Number of Parameters = 22 Samples/Parameters Ratio= 18.09 R² = 96.17 Adjusted R² = 95.95 Degr. of freedom for regr = 22 Total degr of freedom = 397 Sum of squares for regr. = 4.24e+002 Total sum of squares = 4.41e+002 Regression mean square = 1.93e+001 Mean square error(MSE)= 4.50e-002

<u>File Edit View Sample Descriptors FeatureSel</u>

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

XXXXXXXXXX

-

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

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.



Beta 1.04 0.92 -0.41 0.32

-0.37 0.31

0.18

idBeta

0.01 0.02 0.02 0.02 0.02 0.02 0.02

t-statistic

98.25 41.65

-24.34 17.53

-16.76 15.91

13.30

11.91

11.71

t-square

.7e+003

3e+002

2.8e+002 2.5e+002

1.9e+002 1.8e+002 1.4e+002

9.7e+003

5.9e+002 3.1e+002

p-value

0.00

0.00

0.00 0.00

0.00 0.00

0.00