*Advances in Computer Science Research (ACSR), volume 81*
2017 International Conference on Computational Science and Engineering (ICCSE 2017)

# The Best Screen Analysis of Small Sample Multivariate Gradation Regression in Statistical Research

Xiao Xiaonan

Xiamen University Tan Kah Kee College,Zhangzhou Fujian China

xiaoxn@xujc.com

**Keywords**: Small sample, Gradation regression, Screen, Significance level, Critical value

**Abstract**: Gradation regression analysis generally needs adequate samples from which we can obtain more reliable statistical results. However, in some statistical analysis, owing to some practical difficulties, we cannot obtain satisfactory observation numbers of indexes. Therefore, this article makes use of screen method of gradation regression by stages and in batches and deals with this kind of problem successfully. The practical use shows that the effect is satisfactory.

## Introduction

The screen method establishment of small multivariate gradation regression by stages and in batches: Under the circumstances when sample information is limited or sometimes sample number is smaller than the variable number, we adopt the ordinary gradation regression method to analyze the action of each explanation variable $x_1,x_2,\ldots, x_m(m>N)$ to dependent variable Y. The result is not reliable. If we adopt the following method, the result will be satisfactory.

First, we randomly divide every explanation variable($x_1,x_2,\ldots,x_n$) into several batches, and make each batch independent variable number no larger than the half of sample number. And then, under the same significance level of a, we begin to conduct gradation regression by stages and in batches.

At the first stage, we make the regression analysis to each batch significance action to Y. At the second stage we again randomly divide each batch which has been picked out variable($x_{i1},x_{i2},\ldots, x_{ik}$) into several batch and continue to make regression until further picking out the factor which has significant action to y….this going to the end until we have selected several explanation variables which has significant action to y. This method is usually used to make up for the shortage of the sample number[1-2].

## Practical example

In the research of mine silicosis, it often involves the relationship between the phagocyte's rate of living in dust solution and the rate of content of chemical composition. As conditions are limited, we can only get the content of chemical composition. As conditions are limited, we can only investigate eight typical mine areas. We acquire the twenty-six kinds

chemical composition ($Na_2O$,MgO,$Al_2O_3$,$SiO_2$,P,$K_2O$,CaO,$TiO_2$,$Fe_2O_3$,Cr,Mn,V,Ca,Ni,Cu,Zn,Ga,Rb,Sr,Y, Zr,Nb,Ba,La,Pb,Th) of each mine area's dust and the rate of cell living in containing dust's cultivation solution for the phagocyte.

Because the independent variable number is far larger than the sample number ,we directly use twenty-six independent Variable Y. The result is not satisfactory. In view of this, we can adopt this method by which this article shows to analyze the relationship between the rate of cell living Y has the content of twenty six independent Variable $x_i$($i$=1,2,…， 26) to make gradation regression to dependent variable Y. The result is not satisfactory. In view of this,we can adopt this method by which this article shows to analyze the relationship between the rate of cell living Y and the content of twenty of twenty-six chemical solution of dust $x_i$($i$=1,2,…， 26).

At first, we conduct logarithm transformation to original number under the same significance level α=0.15 and critical value $F_1$=$F_2$=2.50,and the carry on gradation regression analysis by stages and in batches .The result is indicated in Table 1.

Table1 The result of 26 variables gradation reqression analysis bt stages and in batches(α=0.15,$F_1$=$F_2$=2.50,$N$=8)

| *stages* | *batches* | *Independent uariable in each batch* | *Selected independent variable* | *Remainder standard deviation* | *Complex correlation efficent* | *Sifnificance F Value* | *P Value* |
|---|---|---|---|---|---|---|---|
| | *1* | $x_1$ $x_2$ $x_3$ $x_4$ $x_5$ | $x_1$ $x_2$ | *3.17632* | *0.88973* | *9.49712* | *<0.05* |
| | *2* | $x_6$ $x_7$ $x_8$ $x_9$ $x_{10}$ | $x_6$ | *5.16749* | *0.58151* | *3.06558* | *<0.15* |
| | *3* | $x_{11}$ $x_{12}$ $x_{13}$ $x_{14}$ | $x_{12}$ | *5.19072* | *0.57636* | *2.98642* | *<0.15* |
| *1* | *4* | $x_{15}$ $x_{16}$ $x_{17}$ $x_{18}$ | $x_{15}$ $x_{16}$ $x_{18}$ | *2.08320* | *0.96348* | *17.2605* | *<0.01* |
| | *5** | $x_{19}$ $x_{20}$ $x_{21}$ $x_{22}$ | | | | | |
| | *6** | $x_{23}$ $x_{24}$ $x_{25}$ $x_{26}$ | | | | | |
| | *1* | $x_1$ $x_5$ $x_6$ | $x_1$ $x_5$ | *3.17632* | *0.88973* | *9.49712* | *<0.05* |
| *2* | *2* | $x_{12}$ $x_5$ $x_{16}$ $x_{18}$ | $x_{12}$ $x_{15}$ $x_{16}$ | *1.79290* | *0.79390* | *23.76933* | *<0.01* |
| *3* | *1* | $x_1$ $x_5$ $x_{12}$ $x_{15}$ $x_{16}$ | $x_1$ $x_5$ | *3.17632* | *0.88973* | *9.49712* | *<0.05* |

*Referring fo no variable to enter under critical value 2.50

Whereas in the third stage the selected variable are $x_1$,$x_5$,$x_{12}$,$x_{15}$,$x_{16}$,the corresponding to skew correlation coefficients are $R_1$=-0.6918 、 $R_5$=-0.85768 、 $R_{12}$=0.07587 、 $R_{15}$=0.36401,$R_{16}$=0.26366. So we can observe that the action of P and $Na_2O$ are the best significant, the second is the Cu、Zn、V.P and $Na_2O$, and the living of phagocyte are reversely related to the rate of living. In the end, we can obtain regression equation[3].

$Y$=115.89588-2.33334$x_1$-3.52800$x_5$

## Discussion and conclusion

The key method in this article is that I divided total correlation matrix $R_{(P+1)\times(p+1)}$ into several main sub-correlation matrix, which corresponds to the number of batches. And then we make the transformation of introduction and reject variable to each main correlation

matrix $R^{(1)}_{(k+1)\times(k+1)}$ by using K variable of corresponding introduction. It is evident that the correlation matrix is still a sub-matrix of $R_{(P+1)\times(p+1)}$.We continue to repeat above mentioned steps on $R^{(1)}_{(k+1)\times(k+1)}$ .We finally can gains factors which are significant to the dependent variable. Their correlation efficiently forms the sub-matrix $R^{m}_{(s+1)\times(s+1)}$of $R_{(P+1)\times(p+1)}$,where m is stage number. $R^{m}_{(s+1)\times(s+1)}$ can be regarded as the factor of inconsiderable dependent variable which has no significant action. It forms the new correlation matrix. That's to say, it has been made for selection. Also i can be regarded as the result which has been finished even introduction and reject transformation to corresponding (p—s) variables in total correlation matrix $R_{(P+1)\times(p+1)}$ in the course of the gradation regression. In short,the correlation $R^{m}_{(s+1)\times(s+1)}$ contains a great deal of information to matrix which has a significant contribution to *Y*. Therefore, the regression analysis of sample by stages and in batches can reflect overall condition[4-5].

If we adopt gradation regression in batches after making assemble analysis in association with system assemble method. Its result (see Table2) will be in accordance with the result of Table 1.

In gradation in batches the significance level of each stage and each batch must keep accordance with the critical value. Sometimes according to the practical needs, it can be led into the standard of introduction variable in certain batches so that the key factors are not rejected.

The screen method in batches in this review has no relationship to the order of line of the independent variable. When we do screen in batches, owing to the fact that the content of sample each batch is twice as big as the independent variable number, this analysis result is more solid and reliable. To be sure, if we expect the effect in batches to be better, we should raise the precision of observation number as carefully as possible[6-7].

Au the results this article be moved on ZD—3100 micromputer.

Table 2 The result of gradation regression in batches offer twenty variable turning into fifteen kinds(α=0.15,F1=F2=2.50,N=8)

| *stages* | *batches* | *Independent uariable in each batch* | *Selected independent variable* | *Remainder standard deviation* | *Complex correlation efficent* | *Sifnificance F Value* | *P Value* |
|---|---|---|---|---|---|---|---|
| | *1* | $x_1\ x_2\ x_4\ x_5\ x_6$ | $x_1\ x_5$ | *0.04421* | *0.87955* | *8.54305* | *<0.05* |
| *1* | *2* | $x_7\ x_8\ x_{10}\ x_{11}$ | $x_5$ | *0.06940* | *0.57397* | *2.94782* | *<0.01* |
| | *3* | $x_{12}\ x_{13}\ x_{14}\ x_{15}$ | $x_{13}$ | *0.06818* | *0.59489* | *3.28638* | *<0.05* |
| | *4** | $x_{19}\ x_{22}$ | | | | | |
| *2* | *1* | $x_1\ x_5\ x_8\ x_{13}$ | $x_1\ x_5$ | *0.04421* | *0.87955* | *8.54305* | *<0.05* |

*Referring fo no variable to enter under critical value 2.50

## References

[1] Frauenfelder P, Schwab C, Todor R A. Finite elements for elliptic problems with stochastic coefficients [J].*Computer Methods in Applied Mechnics and Engineering,* 2005, 194 (2/5): 205-228.

[2] Shi Kaiquan, Yao Bingxue.Funtion. S-rough sets and las identification [J].*Science in China Series F:Information Sciences*, 2008, 51 (5): 499-510.

[3] Anderson T W. *An Introduction to Multivariate Statistical Analysis* [M].New York: John Wiley & Sons, 1990. 485~513.

[4] Yuan C G,William G. Approximate solutions of stochastic differential delay equations with Markocvian switching [J].*Comput Appl* Math, 2006, 194: 207-226.

[5] Lucin L C.*Asymptotic Methods in Statistical Decision Theory* [M].New York: Spring-Verlag, 1988. 264~298.

[6] Kalbfleisch J G. *Probability and Statistical Inference*, Volume 2: Statistical Inference [M].

[7] Korezlioglu H ,Mazziotto G,Szpirglas J. *Filtering and Control of Random Processes* [M]. Berlin: Springer-Verlag, 1984.164~198.