Acta Technica *62* (2017), No. 7A, 247–256 © 2017 Institute of Thermomechanics CAS, v.v.i.

A two factor influence analysis method based on ridge estimation least square algorithm

Cang Wang[1], Shikun Zhang[2,*]

Abstract. In order to improve effectiveness of analysis result for influence of sports awareness of adolescents, a kinds of analysis method about influences of sports media on sports awareness of adolescents based on least squares algorithm (LS) of ridge estimation is proposed. Loss of superiority for LS estimation is analyzed under the ill condition of design matrix for multiple regression equation. Ill reason of design matrix for multiple regression equation in scientific research of sports is discussed. Non-ideal reason of LS estimation due to the kind of data is visually analyzed. The method of biased estimation is proposed in two perspectives so as to improve LS. Then, above-mentioned algorithm is used to evaluate analysis result of sports awareness influence on adolescents. In addition, opinions and suggestions are given.

Key words. Ridge estimation, Least squares algorithm, Sports media, Sports awareness.

1. Introduction

Problem of sports awareness for adolescents gradually becomes a key point concerned by the society. It is shown in current researches that parenting pattern, social support, and response methods, and other factors will influence on level of sports awareness for adolescents. Fewer researchers consider to taking advantages of inherent potentials of adolescents to improve their level of sports awareness. Seligman, founder of positive psychology, believes that positive psychological qualities are strong weapons for human beings to defeat psychological illness. In addition, he believes that human beings have inherent power o defend psychological illness. Positive psychological qualities refer to relatively stable positive psychological traits

[1] School of Media and Arts, Chongqing University of Posts and Telecommunications, Chongqing, 400065, China

[2] Department of Physical Education, College of Sports and Recreation, National Taiwan Normal University, Taiwan Taipei, 10610

*. Corresponding Author

of individual is formed based on congenital quality and acquired environment education. These traits influence positive orientation of individual on individual cognition, feeling, and response, which is basis for realization of inner power and potential for individual.

Symptom Check List 90 (hereinafter referred to as SCL-90) among numerous researches about sports awareness of adolescents is an important research tool. If "adolescent" and "SCL90" are regarded as key words to be searched on China Journal Net, 1130 relevant researches will be found in recent 5 years, including 166 master's thesis and a doctoral dissertation. SCL-90 is prepared by Derogatis, and others, which includes 90 problems described in feeling, emotion, thought, consciousness, behavior, and other aspects and is used to measure 9 factors of somatization, compulsion, interpersonal relationship, depression, anxiety, hostility, terror, paranoid, and psychosis. People to be test can conduct five-grade self-assessment from 1 to 5 for description of all items according to self conditions. The higher the score is, the worse the psychological problem is. Scoring indicators generally include total points (which is added by scores of 90 problems), total average scores (which is obtained by using total scores to divide by 90), positive items (No. of "symptom" is generally subject to the standard about that factor should be ≥ 2 or ≥ 3), and scores of all factors.

Based on above-mentioned indicators, multiple linear regression algorithm is introduced in the thesis so as to analyze sports awareness of adolescents; analysis model for sports awareness of adolescents is constructed; LS solution is introduced so as to conduct model analysis for sports awareness of adolescents. Meanwhile, in order to solve problems of high state space dimension in standard LS solution, complex calculation, and low precision, method of multiple linear regression is used to construction LS regression scheme so as to realize improvement of algorithm performance.

2. Influence of illness of design matrix on LS estimation

2.1. *Least square estimation (LS estimation)*

Considered linear model[1]:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + e\,. \tag{1}$$

After obtaining observation value of samples, LS estimation of regression coefficient β is:

$$\hat{\beta} = (x'x)^{-1} xy\,. \tag{2}$$

Where $\hat{\beta} = (\hat{\beta}_1, \hat{\beta}_2, \cdots \hat{\beta}_p)'$, x indicates matrix of $n(p+1)$; $y = (y_1, y_2 \cdots y_n)$, n indicates sample size. If x and y are subject to standardization, obtained standard regression equation will be:

$$\tilde{y} = \tilde{\gamma}_1 \tilde{x}_1 + \tilde{\gamma}_2 \tilde{x}_2 + \cdots + \tilde{\gamma}_p \tilde{x}_p\,. \tag{3}$$

They will be correlation matrix after standardization. If x is normal, LS estimation will be good.

If there is illness in design matrix (which is also called multi-collinearity) in actual application, LS estimation will have no good property anymore[1][3], which may make the analyzer draw wrong conclusion. Common results are in the following aspects: excessively large estimation error to coefficient; unstable coefficient estimation; great changes of coefficient at the time of increasing and decreasing samples; occurrence of coefficient symbol which are against actual conditions.

2.2. Mean square error (MSE) of estimated value

Definition of MSE: if $\tilde{\theta}$ is estimated value of parameter θ, MSE$\tilde{\theta}$ = E$(||\tilde{\theta}-\theta||2)$ can be called MSE for estimated value of θ. (symbol ||a|| indicates length[3] of vector α in mathematical statistics)

MSE$\tilde{\theta}$ = E$(||\tilde{\theta}-\theta||2)$ is a measurement for deviation size of estimated value and truth value of parameter. Specifically, in terms of a good estimation, MSE$\tilde{\theta}$ should not be excessively large. In order to clearly explain problems, MSE$\tilde{\theta}$ is further disintegrated.

$$\begin{aligned} \mathrm{MSE}\tilde{\theta} =& \mathrm{E}[(\tilde{\theta}-\theta)'(\tilde{\theta}-\theta)] = \mathrm{E}[(\tilde{\theta}-\mathrm{E}\tilde{\theta}) + (\mathrm{E}\tilde{\theta}-\theta)]'[(\mathrm{E}\tilde{\theta}-\theta) + (\mathrm{E}\tilde{\theta}-\theta)] \\ =& tr[cov(\tilde{\theta})] + ||\mathrm{E}\tilde{\theta}-\theta||2 \end{aligned} \tag{4}$$

If θ is written as ($\tilde{\theta}$1, $\tilde{\theta}$2, $\tilde{\theta}$p), the first item of Equation (4) can be written as that measurement of $\sum\limits_{i=1}^{p} var(\tilde{\theta}_i)$ is variance of estimated value for all components of $\tilde{\theta}i$. The second item of Equation (4) can be written as that measurement of $\sum\limits_{i=1}^{p} (E\tilde{\theta}_i - \theta_i)^2$ is variance of estimated value for all components of $\tilde{\theta}$I as well. Theoretically, the two items can be considered as good estimation after they reach the minimum.

2.3. MSE of LS estimation

MSE ($\hat{\beta}$) of LS estimation is discussed based on disintegration of MSE$\tilde{\theta}$. Standard form of regression equation (1) is identical to regression equation (3). If y∼N(xβ, σ2/n), MSE($\hat{\beta}$) = E$||\hat{\beta}-\beta||2$ has testified that E$||\hat{\beta}-\beta||2 = \sigma$tr$(x'x) - 1$ and D$||\hat{\beta}-\beta||2 = 2\sigma 2tr(x'x) - 2$ in statistical theory. Its theoretical basis is shown in [3].

If characteristic roots of x' and x separately are λ1, λ2, and λp and characteristic roots of $(x'x)^{-1}$ and $(x'x)^{-2}$ separately are λ_i^{-1} λ_i^{-2} which are obtained through linear algebra, thus

$$E(||\hat{\beta}-\beta||2) = \sigma 2 \sum_{i=1}^{p} \lambda i^{-1}\,. \tag{5}$$

$$D(||\hat{\beta}-\beta||2)=2\sigma 4\sum_{i=1}^{p}\lambda i^{-2}\,. \tag{6}$$

Error of coefficient estimation is measured in Equation (5); stability of $\hat{\beta}$ (or fluctuation condition of $\hat{\beta}$) is measured in Equation (6). If design matrix is normal, LS estimation is absolutely appropriate. However, if X is ill which means x' and x are characteristic roots approaching 0. Visually, Equation (5) and Equation (6) are especially large, which shows that MSE of LS estimation is excessively large and coefficient is unstable, thus LS estimation loses superiority.

3. Cause and identification method for illness of design matrix in scientific research of sports

3.1. Cause analysis

There are lots of factors leading to illness of design matrix for coefficient of multiple regression equation. Common causes for illness of design matrix in sports field are analyzed in the thesis with the following reasons:

(1) Limitation of data collection. In comparison with researches of other disciplines, analysis of sports awareness has self complexity and peculiarity. Therefore, data collection usually is under limitation of various objective conditions, such as non-repeatability of sports awareness. It can be expressed in statistical language that P collected indicators (variables) are approximately on the Rn plane[3] which is lower than dimension P. In principle, more data can be collected to break through colinearity of ill data. However, objectively, there are lots of difficulties in collecting data. Although it is feasible to collect more data, it may lead to new problems, such as high leverage point, high influence point, and etc, which may cause trouble to analyzer as well,

(2) To a certain degree, there is linear correlation between independent variables of regression equation in objectiveness. In comparison with other disciplines, analysis of sports awareness has self complexity and peculiarity. Human body with complex connections are considered as carrier at the time of completion of sports awareness and various physiological and biochemical responses in the process of movement. Human body is a complex system. Researches in the complex system are not perfected, which brings difficulties to selection of indicators. For example, correlation between indicators which are required to be selected is unclear.

(3) Lots of pseudo variables. Researches of lots of problems in the training process of sports awareness involve in conducting quantitative analysis to qualitative variable. If several qualitative indicators are used to establish regression equation, the common method is to take advantages of variables "0 and 1" (which are called pseudo variables). If lots of pseudo variables are selected, it may lead to completely colinearity of design matrix. In general, No. of independent variables should subtract 1 so as to obtain No. of pseudo variables.

(4) Infirm theoretical basis of adolescents. A variable or several variables with actual correlation is introduced to regression equation due to limit of theoretical knowledge level for analyzer, leading to illness of design matrix. Or correlation between selected indicators is not considered due to carelessness of analyzer, which may lead the serious consequence as well.

In addition, due to rapid development of computer science, lots of adolescents excessively depend on computer, especially at the time of handling large-scale regression problem of multiple variables. They just input subjectively selected variables to the computer without considering it from the perspective of professional knowledge. Therefore, variables with colinearity in objectiveness may be selected in regression equation, leading to illness of design matrix.

3.2. Identification methods for illness of design matrix

There are lots of identification methods for illness of design matrix. Several common identification methods are introduced in the thesis from the perspective of application.

(1) Identification method for correlation coefficient. The specific method is to identify it through analyzing correlation between indicators with professional knowledge. If correlation coefficient between indicators reaches 0.75, it is usually considered that it is high correlation[3], which will lead to illness of design matrix. There is one point needing to be concerned, which is that correlation identification method can only be used to identify relationship between two indicators and it can not be used to identify collinear relationship between multiple indicators.

(2) Contradictory identification method between F test and t test. In terms of test for regression equation, if selected variable of F test is found to have significant relationship with dependent variable, several or all single variables will be found to be not significant in t test, which shows that contradiction between F test and t test is a good sign[4] for multiple colinearity. It can be judged that design matrix is ill.

(3) Identification method for characteristic root (which is also called identification method of principle component). If Equation (3) (correlation matrix) is subject to spectral factorization, obtained characteristic roots will separately be $\lambda 1$, $\lambda 2$, and λp. If one of them or several of them approach 0, it shows that there is colinearity[4] between original independent variable, which leads to illness of design matrix.

4. Improvement methods for LS estimation

It is known that bad effect of LS estimation for illness of design matrix is reflected in MSE, which means that MSE($\hat{\beta}$) is large. The reason is that there is characteristic root in $x'x$ approaching 0. The intuitive idea for improvement of LS is to conduct proper conversion for $x'x$ so as to break through colinearity and to improve the degree of characteristic root approaching to 0. Ridge estimation is introduced as follows from the perspective of reducing MSE. In addition, biased estimation of principle components are introduced from the prospective of eliminating multi-collinearity between independent variables so as to improve LS estimation.

4.1. Ridge estimation

It can be known in Equation (3) that $\hat{\beta}$ estimation is $\hat{\beta} = (\tilde{x}'\tilde{x})^{-1}\tilde{x}\tilde{y}$; it is assumed that a small positive No. $k(0 < k < 1)$ is added on main diagonal element $\tilde{x}'\tilde{x}$ so as to improve characteristic root of $x'x$ which approaches 0 for reducing and stabilizing estimated MSE of coefficient. According to idea of ridge estimation, estimation expression of β is $\hat{\beta}(k) = (\tilde{x}'\tilde{x}+kIp)-1\tilde{x}'\tilde{y}$. In order to clearly observe its structure, the equation is expanded to be:

$$\begin{pmatrix} \tilde{\beta}_1(k) \\ \tilde{\beta}_2(k) \\ \vdots \\ \tilde{\beta}_3(k) \end{pmatrix} = \begin{pmatrix} \gamma_{11} + k & \gamma_{12} \cdots\cdots \gamma_{1p} & & \\ \gamma_{21} & \gamma_{22} + k \cdots\cdots \gamma_{2p} & & \\ \vdots & \vdots & \ddots & \vdots \\ \gamma_{p1} & \gamma_{p2} \cdots\cdots\cdots \gamma_{pp} + k & & \end{pmatrix}^{-1} \begin{pmatrix} \gamma_1 y \\ \gamma_2 y \\ \vdots \\ \gamma_p y \end{pmatrix}. \tag{7}$$

It is testified in statistical theory that there is always a proper k which makes MSE($\hat{\beta}k$) reach the minimum and makes the conclusion of MSE($\hat{\beta}k$) <MSE$\hat{\beta}$ be established when $k > 0$[6].

4.2. Selection of K value for ridge parameter

A small positive k which is introduced in ridge parameter is called ridge parameter. Determination of its value is up to sample data, thus it is hard to be determined. A common method to determine ridge parameter is introduced as follows, which is ridge trace method [4].

Ridge trace refers to a trace described by taking different $k(0 < k < 1)$ $k$value as x-coordinate and taking $\hat{\beta}i(k)$ as y-coordinate. When k value is the optimal? It is indicated in Literature that described ridge trace is in stable state and there is no unreasonable symbol and quick rise for residual sum of squares. K value at the moment is the selected one. Due to complex calculation of ridge trace, a convenient ridge trace formula is given as follows so as to avoid complex adverse calculation:

$$\hat{\beta}(k) = (\tilde{x}'\tilde{x} + kIp) - 1\tilde{x}'\tilde{y} = \sum_{i=1}^{p} (\frac{1}{\lambda_i + k})\varphi_i \varphi_i' x' y\,. \tag{8}$$

Where λ_i and φ_i are characteristic root of x,x and corresponding characteristic vector of characteristic root.

5. Experimental analysis

There is are only 2 major which position talents on the "compound" training target in 2007 edition of training scheme. However, the idea of "professional" talents is abandoned in all majors of training scheme in the thesis. Training specification of "innovative awareness", "practical ability", "understanding to cutting-edge knowledge in field", and others is used, which specifically shows that "compound" requirement cycle of current society to practical ability, comprehensive quality, scientific research innovation, and later development potential of talents in reality. On the basis, majors with traditional technology advantages emphasize scientific training and humanistic literacy training for adolescents according to its disadvantages, consciously abandon the idea that professional talents just serve physical education and competitive sports, and emphasize guidance for public fitness and social sports in guidance position of social service. In addition, the capacity requirement of integrating martial arts into performance is proposed in martial arts major and traditional national sports major so as to seek for development and growth point of majors in new period for responding demand change of training for sports awareness for talents in sports trade.

Professional basic courses adhere to amendment theory of training core qualities and capacities of adolescents and teaching of relevant basic methods and classic knowledge and abandon problems of original early major orientation, narrow scope of knowledge, and etc., which contributes to broadening disciplinary horizon and speciality range for laying good foundation for major study and later development of sports awareness for adolescents.

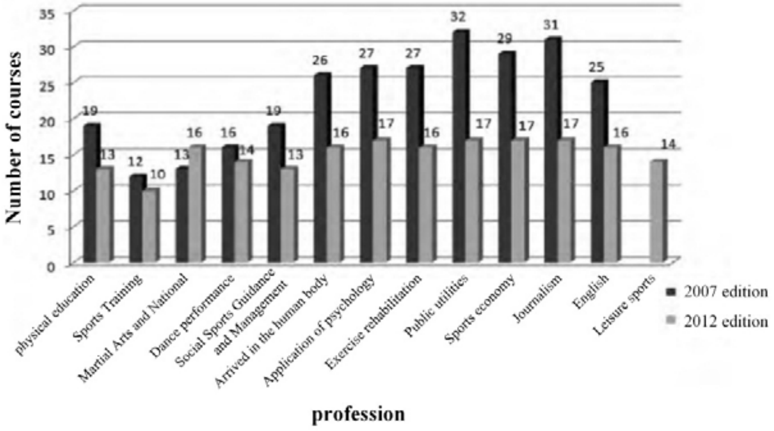

Fig. 1. Analysis of development for sports awareness in training scheme

On the basis, the method of self selection is uniformly set in 2007 edition of scheme for optional courses. Major development direction or different major skills of all majors are shown through design of 2 or 3 modules, which to a certain degree avoids the phenomena of course arrangement according to people for sports colleges triggered by narrow professional background of teachers and staff shortage, and etc. at the time of providing independently selected space of adolescents. It is shown in Table 1 that selected space is averagely improved by 30% in training scheme in the thesis. In addition, the following chronic disease is eradicated. For example, diversified development of sports awareness for adolescents can not be improved, because arrangement of several optional courses is equal to arrangement of compulsive courses.

Table 1. Comparison of evaluation result

Name of major	Standard LS algorithm			LS algorithm of ridge estimation		
	Traditional media	Media	Selected space	Traditional media	Media	Traditional media
Physical Education	14	34	2.42	14	50.2	3.62
Sports Training	24	41	1.71	16	52	3.21
Martial Arts and Traditional National Sports	8	14.5	1.82	18	27	1.50
Dance performance	16	23.6	1.47	16	30.2	1.91
Social Sports Guidance and Management	39.5	69.6	1.75	29	69	2.34
Sports Science of Human Body	28	32	1.13	29	54	1.93
Applied Psychology	24	30	1.24	34	47.3	1.35
Sports Rehabilitation	34	41.2	1.23	21	44	2.04
Management of Public Services	27	27	1.01	32	42	1.28
Management of Public Services	30	30	1.02	21	41	2.04
Journalism	25	40	1.55	23	41	1.85
English	12	27	2.24	21	40	2.01
Recreational Sports	14	26	51	8	38	2.05

6. Conclusion

(1) Improvement of social responsibility consciousness. In comparison with traditional media, media has short time and quick changes. Meanwhile, in order to win more users, there are some problems in operation, such as recreational and pornographic news reports, flush headlines, inaccurate news report due to exaggeration of major events, and etc. of all large websites, which leads to deviation to adolescents at the time of receiving sports information. Therefore, social responsibility consciousness of media has to be improved and supervision strength of public opinions on media should be strengthened. (2) Encouragement of booming development of media and encouragement of spread for positive energy of sports information. Booming development of media has become historical trend and time tide, thus development of media can not be doubted or stopped due to some problems in the spread process of sports media. On the contrary, development of media should be supported and encouraged. Meanwhile, spread of positive energy for sports should be encouraged in the spread process of sports information for media so as to convey active and positive sports information to adolescents for guiding them to build up right sports value. (3) Improvement of media literacy for adolescents. Media literacy refers to the capacity of correctly and constructively enjoying mass communication resources, the capacity of fully using media resources to perfect oneself, and the capacity of participating in

social progress. Social literacy has three meanings: firstly, media should be used to obtain knowledge; secondly, value and meaning spread by media should be judged; finally, media information should be used o perfect and develop oneself. "That how to correctly identify and select media and information in miscellaneous, complicated, and numerous information becomes a kind of cultural confusion in information age", said by a journalism expert Dong Changlan. Therefore, college, family, and society should form a resultant force so as to help adolescents to improve their media literacy and to avoid that sports nature and themselves are lost in complicated sports information.

References

[1] L. S. Gresham, D. L. Zirkle, S. Tolchin, et al.: *Partnering for injury prevention: evaluation of a curriculum-based intervention program among elementary school children*[J]. Journal of pediatric nursing, *16* (2001), No. 2, 79–87.

[2] L. S. Olive, D. G. Byrne, R. B. Cunningham, et al.: *Effects of physical activity, fitness and fatness on children's body image: The Australian LOOK longitudinal study*[J]. Mental Health and Physical Activity, *5* 2012, No. 2, 116–124.

[3] G. W. Imbens: *Nonparametric estimation of average treatment effects under exogeneity: A review[J]. The review of Economics and Statistics*, (86) 2004, No. 1, 4–29.

[4] D. R. Hardoon, S. Szedmak, J. Shawe-Taylor: *Canonical correlation analysis: An overview with application to learning methods*[J]. Neural computation, *16* 2004, No. 12, 2639–2664.

[5] R. De Oliveira, A. Armenta, P. Concejero, et al.: *Customer cognitive style prediction model based on mobile behavioral profile: U.S.* Patent Application 13/177, 615[P]. 2011-7-7.

[6] C. Crespo, M. Kielpikowski, P. E. Jose, et al.: *Relationships between family connectedness and body satisfaction: A longitudinal study of adolescent girls and boys*[J]. Journal of Youth and Adolescence, *39* (2010), No. 12, 1392–1401.

[7] G. Rajlic: *Comparison of the approaches to assessing statistical interactions: an application to risk factors for adolescent problem behaviour*[D]. University of British Columbia, 2014.

[8] M. L. Butson, E. Borkoles, C. Hanlon, et al.: *Examining the role of parental self-regulation in family physical activity: A mixed-methods approach*[J]. Psychology & health, *29* 2014, No. 10, 1137–1155.

[9] C. F. Matthies: *Evidence-based approaches to law enforcement recruitment and hiring studies of the Los Angeles Police Department*[D]. The Pardee RAND Graduate School, (2011).

[10] S. L. Brantley, J. P. Megonigal, F. N. Scatena, et al.: *Twelve testable hypotheses on the geobiology of weathering*[J]. Geobiology, *9* 2011, No. 2, 140–165.

[11] D. P. Miller: *Accessibility of summer meals and the food insecurity of low-income households with children*[J]. Public health nutrition, *19* (2016), No. 11, 2079–2089.

[12] W. Lu, E. L. J. McKyer, C. Lee C., et al.: *Perceived barriers to children's active commuting to school: a systematic review of empirical, methodological and theoretical evidence*[J]. International Journal of Behavioral Nutrition and Physical Activity, *11* 2014, No. 1, 140.

[13] Z. Lv, A. Halawani, S. Feng, H. Li H., & S. U. Réhman: *Multimodal hand and foot gesture interaction for handheld devices.* ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM), *11* (2014), No. 1s, 10.

[14] W. S. Pan, S. Z. Chen, Z. Y. Feng, *Automatic Clustering of Social Tag using Community Detection.* Applied Mathematics & Information Sciences, *7* 2013, No. 2, 675–681.

[15] Y. Y. Zhang, J. W. Chan, A. Moretti, and K. E. Uhrich: *Designing Polymers with Sugar-based Advantages for Bioactive Delivery Applications*, Journal of Controlled Release, *219*, 2015, 355–368.

Received May 7, 2017