The Erroneous Application and Countermeasures of Return Analysis in Sports Research

Xiuying Han
College of physical education, Shandong University of Finance And Economics, Jinan, Shandong, 250014, China

***Abstract*—On the basis of widely reading papers published in the 13 core Chinese journals recently and with methods of literature review, expert interview and questionnaire, this paper points out the problem concerning regression in sports research, such as the model examination, the content of the case study, gradual regression and the invented variables. This paper also analyzes the problems and gives the right solutions to improve the application effect and provide a reference for further study in future.**

Keywords—regression; sports research; application; wrong areas

I. INTRODUCTION

In order to learn more about the statistical research in sports education, we investigated 144 papers in 13 core periodicals of sports published during 1998 - 2008. In these papers, the extensively adopted statistical methods are mean number, standard deviation, t-test, correlation analysis, multivariate regression analysis, type analysis, PCA and mathematical model prediction. These 11 years are peak time in P.E statistical research and application, and research interests lie in practical studies in physical fitness research, physical education teaching, training and sports medicine. However, there are less papers illustrating problems in regression in sports research and still less papers on solving these problems in accordance with the circumstances of sports studies. Regression is a primary method in mathematical statistics. It is mainly applied to analyze correlations between variables, predict research results, evaluate governmental policies as well as test and develop theories [1]. Regression analysis is widely applied in sports research and brings out productive research results. However, we found in many papers that if authors had partial or unscientific understandings in regression and had no awareness of the problems in practical research, they would draw unreasonable conclusions or fail to reveal some hidden rules or even make huge mistakes in policy decisions. After looking through papers in Chinese core periodicals on sports research, we analyze research problems of regression, and present scientific treatment aiming to provide a reference on a more scientific application of regression analysis.

II. PROBLEMS IN MODEL TESTING

Generally speaking, regression models need to meet three tests, namely sports science test, statistical validation, prediction test of models. But presently tests of models in sports research usually only go through statistical tests, sports science test and prediction test of models being neglected [2].

A. Sports Significance Test

Sports significance test mainly inspects rationality of parameter estimate in sports science. Theoretically expected value is compared with parameter estimate in symbols, size and correlations between parameters.

First, symbols of parameter estimate need to be tested. For instance, in the following model of football players' competitiveness:

Competitiveness = 5.02－10.08×explosive leg strength＋0.64×cardiovascular endurance＋2.37×agility

In this regression model, the parameter estimate before explosive leg strength is negative, meaning the stronger the explosive leg strength is, the weaker the competitiveness is, which is contrary to the actual situation and is meaningless. This model apparently fails to meet the test, partly due to the default linear model when modeling, consequently the above model needs further tests and more analyses.

When all the symbols of parameter estimate are correct, more tests are still required for the size of the estimate and correlations between all the estimate. In the following model for family demands of sportswear:

Ln (average expenditure in purchasing sportswear) =-3.69 ＋1.20 ln (average income)－6.40ln (sportswear price)

As this model is a log-linear model, so in this model, parameters before "average income" and "sportswear price" have definite economic significance for sports, and their symbols of parameters are correct with generally appropriate numerical range. But according to economic significance for sports, the sum of the two parameter estimates should be around 1%, because when income and price both increase by 1%, average expenditure in purchasing sportswear should also increase by 1%. Apparently the parameter estimate of this model can not pass the test, so the reason for this mistake should be found out and a new model need to be set up.

Only if parameter estimates pass all tests of sports significance, can models be tested afterwards. Sports significance test of parameter estimates is of primary importance, as if the model is unscientific in sports significance, whatever the quality in other aspects, the model is of no practical value. In other words, an regression model can be

made only if it accords with sports phenomenon, otherwise, conclusions drawn from this model are contrary to the reality making statistics misleading and casting a negative impact and bringing severe consequences to the society.

B. Statistics Test

Statistics tests are guided by statistical theories at the purpose of testing statistical quality of the model. The most widely applied statistical tests are goodness-of-fit test, significance test of variables and equations, multicollinearity test explaining variables. However serial correlation test for stochastic error and Engle's ARCH test are seldom used. In all the 144 papers we investigated, there exists a common problem that tests of regression models are not comprehensive enough, and are limited to goodness-of-fit test and significance test. Their tests are problematic in some aspects. The present paper puts forward some proposals so as to remedy some defects in regression tests of the sports research.

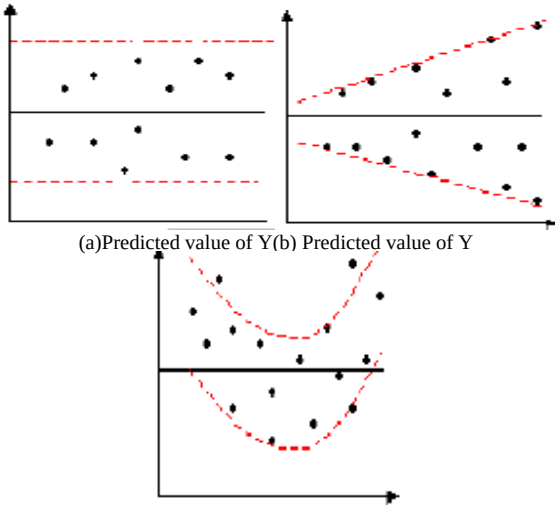
The effect of regression analysis is evaluated from the following three perspectives:

1) Coefficient of determination. In sports research, it is generally agreed the higher the value ofR2 is, the better GFI becomes. But there is no definite limit as to how high the value of R2 can be. In some papers, the value of R2 is quite onfusing lack of unified standard. So how high the value can be considered better GFI?

In sports research, the value of R2 is determined by the kind of data in regression analysis, being time series data or cross section data. For time series data, the value of R2 can be reasonably as low as 0.4 or 0.5, judging by the actual situation.

2) Residual standard deviation. Residual standard deviation can be applied to estimate precision of equation. The fundamental purpose of regression analysis is to estimate the variation of dependent variables through equations. Residual standard deviation can help estimate confidence interval of the predicted value for the corresponding explanatory variables. After being examined in the context of research interests, whether the length of the confidence is reasonable or not is revealed. A large confidence interval means poor regression effect, so the equation needs to be improved.

3) Residual plot. By observing the regularity of residual distribution in residual plot, we can find system errors in the equation. When residual plot is presented in rectangular coordinate system, usually the predicted value of Y represents horizontal axis, and error between Y and Y, i.e. E is presented on vertical axis. Scatter diagram of the residual errors are hence displayed in the plot.



(a)Predicted value of Y(b) Predicted value of Y

(c) Predicted value of Y

FIGURE I. RESIDUAL PLOT

The residual distribution of an ideal model should be random and irregular. Scatters in figure (a) distribute at random, which means residual error and independent variable are independent of each other, i.e. the model is very scientific. Scatters in figure (b) distribute quite regularly, which indicates heteroscedasticity. And in figure (c), scatters appear to be fluctuated and hence independent variables show autocorrelation. Models built in the latter two situations should be reconsidered and revised.

C. Prediction Test of Model

In some papers, tests of regression analysis models were merely performed within the original sample. Although the test accuracy was high, the test for the varied sample size was neglected. This problem can be satisfactorily solvenrvations, i.e. super-sample model. Specific test method is as follows:

1) Stability test of model. In this test, the stability of the model can be judged by analyzing the influence of sample size variation on estimate of parameter. More specifically, we add (or subtract) 1 to 2 observation data to (or from) the original sample data of estimated model parameter, which is re-estimated, and compare the newly estimated model to the former one[4].

If symbols of parameter stays the same and the value has no substantial changes, the original model is of relatively high stability and reliability.

2) Prediction test of model. We applied models with estimated parameter to prediction of actual activity, and compare results estimated by models with the fact so as to test effectiveness of the model. Specifically, observation data for explaining variables, within or out of the sample range, are substituted into the model, work out the theoretical value of explaining variables, and compare the theoretical value of the explained variables with the actual occurrence value. If there are only small differences between the theoretical and actual

value, then the model can better reflect the real economic situation, and has high application value. On the contrary, the model can not effectively simulate operation rules of the actual situation, with low application value and should be abandoned.

After passing all the above tests, a satisfying regression model is set up and can be used for a scheduled purpose.

III. Problems in Sample Size

In sports research, model parameters are estimated with the help of sample observation, so they are quite dependent on samples. However collecting and organizing data can be very difficulty, so proper and minimum sample size can satisfy the need of building models and also relieve the burden of collecting data. In some studies, there exists some problems in sample size, for example, less sample size than the amount of variables. The following proposals are put forward to solve these problems.

A. Minimum Sample Size

Confined by the condition, in sports research, sample size can not be large enough. Then what is the minimum of sample size? Does it make sense that regression can be carried out with a certain size of sample? Apparently it is not the case.

The so called "minimum sample size", is decided under the guide of the principle of least squares, and maximum likelihood. Whatever the quality of parameter of estimate is, theoretically, the minimum of sample size is $n \geqslant k+1$. In other words, sample size should be not less than the amount of explaining variables in the model, including the constant term, and this is also the minimum sample size.

B. Sample Size Meeting the Basic Requirements

When the condition $n \geqslant k+1$ is met, parameter estimate is determined. But when the quality of estimate is unsatisfactory, succeeding activities for building models can't be carried out properly. For instance, statistical test of parameters requires large size of samples. Z test can not be applied when $n < 30$. T test is the most common method in testing significance of variables. Experience has shown that when $n - k \geqslant 8$, distribution of t is stable and test is effective. So, generally when $n \geqslant 30$ or at least $n \geqslant 3\ (k+1)$ basic requirements of model estimate can be satisfied.

IV. Problems in Stepwise Regression

In regression analysis, screening of variables or validation of model is of primary importance. Usually when performing regression analysis in sports research, stepwise regression is employed to screen variables, and first degree original variable. The screening process is conducted by computers and stepwise regression is made by default. This way of regression analysis is reasonable in a statistical sense, but not in the sense of sports significance.

A. Determining Mathematical Form of Variables

Regression analysis in sports research is conducted by analyzing the original variable linearly. As in the reality, relations between variables are not always linear, and variables often influence each other[5], in fact when regression is applied, quadratic term of original variables and other function terms are also taken into consideration. In general, quadratic term reflects U-shaped relationship or inverted U-shaped relationship, and product term reflects interactive relationship between variables.

Introduction of dummy variables can help introduce quadratic term of independent variables or other equations into linear regression analysis. If we introduce quadratic term, set x1, x2 be the original independent variables, set $x3 = x12$, $x4 = x22$, $x5 = x1 \times x2$, then linear regression analysis is conducted with x1, x2, x3, x4, x5 being independent variables. But if we have p original variables, then there'll be $p(p+1)/2$ quadratic terms, and more equations. When there are too many independent variables, it is unrealistic to take all of them into consideration, so in the actual research, we should first conduct qualitative analysis according to specialized knowledge, then introduce some quadratic terms or other equations into the model, then stepwise regression and diagnostic analysis should be conducted, and "try-analyze-improve" the model repeatedly in order to acquire a preferable model.

B. Stepwise Regression Procedures

After mathematic form of variables is fixed, variables should be selected on the stepwise regression principle with subjective analysis. The regression model has to be tried-analyzed-improved repeatedly. And this process shouldn't be completed merely by computers. Specific procedures are as follows:

1) Select the most correlated variables from all the explaining variables, and build a unitary regression model.

2) Introduce the second variable into the model, set up $k-1$ bivariate regression model (with k variables) and choose a better model from these models. When selecting, every explaining variable should be of appreciable impact in this model, parameter symbols be correct, and the value of R2 is increased.

3) Introduce the third variable into the chosen bivariate regression model, and the procedures keep on processing, until no new variables can be introduced.

When introducing regression equation with new explaining variables:

(1) The newly introduced explaining variable can be adopted if it can improve the value of goodness of fit, R2, being statistically significant, on the premise that the variable accord with sports significance.

(2) The newly introduced variable should be abandoned, if it can not improve goodness of fit, and has no obvious significance on other variables.

(3) The newly introduced explaining variable brings collinearity if it improves goodness of fit, R2, greatly influences symbols and value of other parameters and has low statistical significance. Apply the above mentioned method, investigate the degree of the linear correlation, and estimate its sports significance. Then of the two variables with the highest collinearity, abandon the one with lower significance on the explained variable, and with minor sports significance. Keep the more significant one.

V. Introduction of Dummy Variables

In regression analysis, independent variables are basically quantitative variables, but that doesn't necessarily mean qualitative variables can not be selected as independent variables. Qualitative variables are usually nonnumeric, and represents some quality, such as male or female, urban or rural resident, normal or abnormal calimate, stable or unstable government strategy. Nowadays, when regression analysis is conducted, regression equation is also built according to possible values of qualitative variables. For example, we build equations for both male athlete and female one, and we seldom comparatively analyze these equations in a quantitative way.

But in reality, when the involved qualitative variables are independent, the better solution is to introduce dummy variables into the equation. Dummy variables are artificial ones reflecting qualitative factors' change with value of 0 or 1 and they are traditionally represented by the letter D. For instance:

$$D1 = \begin{cases} 1 \ urban \ resident \\ 0 \ rural \ resident \end{cases}$$

$$D2 = \begin{cases} 1 \ male \\ 0 \ female \end{cases}$$

$$D3 = \begin{cases} 1 \ athlete \\ 0 \ non-athlete \end{cases}$$

The introduction of dummy variables can quantify the influence of qualitative variables or qualitative factors on independent variables, and can reflect correlations between variables scientifically, improve equation accuracy and deal with abnormal data more easily.

In principle, if qualitative variables have m different qualities or mutually exclusive types, only m－1 dummy variables can be applied, or else complete multicollinearity will emerge. So qualitative factor "resident" with two different qualities "urban" and "rural", and m=2 in this occasion, so m—1＝2—1＝1 dummy variable can be applied.

$$D = \begin{cases} 1 \ urban \ resident \\ 0 \ rural \ resident \end{cases}$$

If there are n qualitative variables, and each has two different qualitative factors, then we should introduce n dummy variables. So when urban, rural and gender differences are considered, we should let dummy variables be:

$$D1 = \begin{cases} 1 \ urban \ resident \\ 0 \ rural \ resident \end{cases}$$

$$D2 = \begin{cases} 1 \ male \\ 0 \ female \end{cases}$$

When dummy variables are introduced, not only effect of various factors are reflected but also interaction effect are reflected by introducing product term of dummy variables and other ones. At this time, corresponding equations are integrated into one model, making comparison between qualities easier. Meanwhile as all data are collected and taken into calculation, utilization of data is significantly improved.

VI. Conclusion and Suggestion

A. *Conclusion*

First, statistics was mechanically applied in many studies we investigated, though sometimes there was nothing wrong with the method itself, it was not reasonably applied in reality; hence it was hard to acquire better effect. In sports research, researchers are not only required to know research questions well, but also to be acquainted well with the statistical method. Surely statistical analysis itself is a comprehensive process, and we can not apply formulas mechanically. Second, when researchers try to convert a specific sports research project into a statistical question with regression analysis, he has to take many crucial steps. More specifically, he needs to specify the research object (statistical ensemble), master the nature of the study, apply appropriate method, select variables and samples then goes further with statistical calculation, building equations, hypothesis testing and result analysis. These procedures comprise the whole process of statistical regression analysis. Anyone failure of these steps may directly lead to mistakes in result analysis. Third, if the overall research object is not definite, it is hard to define what the sample represents, making sampling meaningless. According to the statistical theory of regression, indefinite research object making sampling design as well as drawing conclusions impossible.

B. *Suggestion*

First, in regression analysis, building model and subsequent analysis process are centered on relationship between independent and dependent variables, so identify this relationship is crucial to regression analysis. Second, basically, deep insight into the nature and basic idea of regression equation makes possible the clear understanding of relationship between variables in equation. Third, when applying least square method for building equation, the inverse function can't control or predict or else errors can be huge. But if regression equation is set with distance method, inverse function can be applied for prediction and control. Last, more research into application of regression analysis should be conducted. In statistics regression analysis is quite profound and full of

variety. It is not possible for a researcher to understand and master it comprehensively, but it is necessary to apply it correctly into sports research.

REFERENCES

[1] Shi Shuhua, Shi Junxin, Zhang Jing et al. Study of Making Normal Values of Preschooler's Four Sports Events with Linear Regression Model, J. Chinese Journal of Child Health Care, 2001, 9 (1):20-28.

[2] Zheng Huilian. Child Health Care, People's Medical Publishing House, Beijing, 1993:50-70

[3] Sun Jingshui. Econometrics, Tsinghua University Press, Beijing, 2004:167-170.

[4] Zhu Jun. Principles of Linear Model Analysis. Beijing: Science Press, 1999.i.

[5] Yang Wenli. Introduction to Linear Models, Beijing Normal University Press, Beijing, 1998.

[6] Liu Wei. Unbefitting Application of Liner Model in Scientific Research of Physical Education and Sports, J. Journal of Shanghai Physical Education Institute, 2001,25(4):23-26