Send Orders for Reprints to reprints@benthamscience.ae

Statistical Methods of Causal Inference in Sports Science Research

Jinchao Li*

School of Physical Education, Huaibei Normal University, Huaibei 235000, Anhui, China

Abstract: Among the complex large quantity of sports phenomenon and sports issues, we need to discuss the causal relationship and essential rule in between. Mathematical statistics is one of the common and practical methods for causal inference. In order to better understand, select, and use mathematical statistic methods, and to avoid deviation of research result due to misuse or even abuse of statistic methods, this paper, through literature and comparative analysis, enumerates the basic theory, model of calculation formula, applicable condition, and operation process of common causal relationship statistical methods, including partial correlation analysis, multiple regression analysis, path analysis, and structural method model, with raising limitation of these methods at the same time. These methods have the same logic, but different principles and calculation process, with one of them a panacea. Therefore, research result of high reliability and validity can only be drawn by full understanding of the applicable conditions of the various methods and comprehensive use of various methods based on actual condition.

Keywords: Causal inference, statistical methods, sports.

1. INTRODUCTION

Sports science is the discipline that studies sports science system and its development direction. It has the properties of natural science as well as social science. Sports science is a comprehensive science that explores the nature and laws of sports with scientific method; researches various sports phenomenon to maximize the sports capabilities of human; and educates and improves health level by physical exercises. Sports means nothing when no human engaged in. In this case, research and experimental research all focus more on the collection of attitude, behavior, and view of human to draw causal result through statistic analysis.

The methods to draw causal conclusions include quantitative and qualitative methods, and each of them has its own limitation, advantage, and disadvantage. The combination of various methods is often used in researches to avoid results with low reliability and validity due to single method. Statistical methods are no exception. This paper, taking common used causal inferences in sports science research as examples, briefly introduced the theory, assumption, model, condition, and operation process of these methods.

2. LOGIC THEORY OF CAUSAL RELATIONSHIP

Cause and effect is a pair of categories that reveal the successive and constraint relationship among things in universal connection in the objective world. Cause the phenomenon that leads to certain phenomena, while effect is the phenomena caused by certain cause.

Causal relationship must meet the following three conditions: first, causal relationship is unidirectional relationship, while associative relationship is a two-way relationship; second, cause and effect variables is successive in time, i.e. effect comes after cause; third, different relationship between cause and effect variables depends on the third variable.

As high frequency statistical methods in science research, partial correlation, multiple regression, path analysis and structural equation modeling, are in the same logic thinking, and based on the five reasoning methods of inductive method (Mueller Five Methods). Muller Five Methods can be expressed as follows: a) the method of agreement: If there is only one condition appeared in various situation of the studied phenomenon, that only condition is the cause of the phenomenon. b) The method of difference: When examining the two situations where the studied phenomenon appear and not appear, determine whether there is only one condition that is different. If yes, there is causal relationship between the condition and the studied phenomenon. c) the joint method of agreement and difference: If there is only one single common case in the positive scenario where the studied phenomenon appears, and there is no the mentioned common case in the negative scenario where the studied phenomenon does not appear, then this common case is the cause of the studied phenomenon. d) The method of concomitant variation: In the situation where other conditions remain unchanged, if one phenomenon changes along with another phenomenon, then the latter one is the cause of the former one. e) The method of residues: If a complex phenomenon is caused by certain complex cause, the remaining portion has causal relationship when the part with determined causal relationship is deducted.

1874-110X/15

3. STATISTICAL METHODS TO DETERMINE CAUSAL RELATIONSHIP

3.1. Partial Correlation Analysis Theory

First of all, we need to understand some theories about general correlation analysis. The main purpose of correlation analysis is to study the close degree of the variables, and to infer general correlation condition based on sample information. The indicator that reflects the close degree of variables is correlation coefficient r, with value between -1 and +1. When the value is close to -1 or +1, the relationship is close, while when the value is close to 0, the relationship is not close. There are several methods for the calculation of coefficient, among which, Pearson correlation coefficient applies to even interval measurement, and Spearman and Kendall rank correlation coefficients apply for non-parameter measurement.

Pearson correlation coefficient is calculated as follows:

$$r = \frac{L_{xy}}{\sqrt{L_{xx} \bullet L_{yy}}} = \frac{\sum xy - \frac{\sum x \sum y}{n}}{\sqrt{\sum x^2 - (\sum x)^2 / n} \bullet \sqrt{\sum y^2 - (\sum y)^2 / n}} \quad (1)$$

The application of correlation analysis must meet the following conditions:

X and Y must be two random variables; for any value of X, the conditional distribution of Y is normal; and vice versa; the joint distribution of X and Y is a two-dimensional normal distribution.

The correlation between multiple variables is complex: there is simple correlation between any two variables; two variables with correlation are not necessarily in causal relationship; two variables in causal relationship must have correlation. The correlation is interspersed with influence brought by other variables. Therefore, the simple correlation cannot fully reflect the pure correlation between two variables. For example, in the research of factors affecting performance in athletic competition, other factors are fixed to calculate the correlation between two variables. This coefficient is called partial correlation coefficient. Partial correlation coefficients can be interpreted as the relationship between certain dependent variable and the independent variable when other dependent variables are fixed.

The formula for partial correlation (two dependent variables) is as follows:

$$r_{y2 \cdot 1} = \sqrt{\frac{R^2_{y \cdot 12} - r^2_{y1}}{1 - r^2_{y1}}} \quad (2)$$

Where: $r_{y2 \cdot 1}$ refers to the partial correlation coefficient of the second independent variable when the first independent variable is fixed; $R^2_{y \cdot 12}$ refers to the coefficient of determination with two independent variables; r^2_{y1} refers to the coefficient of determination of the dependent variable Y and the first independent variable. Similarly, you can also calculate the partial correlation coefficient of the first independent variable when the second independent variable is fixed.

3.2. Theory of Multiple Regression Analysis

3.2.1. Model and Assumption

Single linear regression model mainly studies the relationship between two variables. However, phenomenon in sports is complex, with the combined results of many factors. In this case, we often adopt multiple linear regression analysis in sports science research, i.e. with two or more independent variables. Linear regression analysis studies if there is certain linear relationship between one or more independent variables and one dependent variable.

General mathematical model of multiple regression models is:

$$Y_i = \beta_0 + \beta_1 \chi_{i1} + ... + \beta_p \chi_{p1} + \varepsilon_i \quad (3)$$

Where, i = 1, 2, n, respectively, χ_{i1} χ_{i2} ···· χ_{ip} are the values of the independent variable χ_1 χ_2 ··· χ_p for i^{th} observation.

The effect of regression equation can be examined by analysis of variance .ANOVA test is as follows: $L_{yy} = U + Q$, Among them,

$$L_{yy} = \sum (y - \bar{y})^2 = \sum y^2 - (\sum y)^2 / n$$
$$U = \sum (\hat{y} - \bar{y})^2 = b_1 L_{1y} + b_2 L_{2y} \quad (4)$$
$$Q = \sum (y - \hat{y})^2 = L_{yy} - U$$

Calculate the value F, Check F distribution table for given α, and critical value Fα (k, n-k-1), then list the analysis of variance table:

Sources of variation	Quadratic sum	DOF	Mean square	F value
Regression	$U = \sum b_i L_{iy}$	k	U/k	(N-k-1) U/kq
Residual	$Q = L_{yy} - U$	N-k-1	Q/(n-k-1)	
$\sum$	$L_{yy} = \sum (y - \bar{y})^2$	N-1		

Compare F and Fα to determine if the effect of regression equation is significant.

Application of multiple regression models must meet the following assumptions:

χ_i may be arbitrarily determined, or deliberately selected variable. It is used as an independent variable to explain the cause for the dependent variable Y, therefore, it is also known as the explanatory variables; For each i, ε_i is a normal independent distribution with mean value of 0 and

variance of σ^2; Each factor is independent from other factors; the independent variable and dependent variable are in linear relationship. To sum up, the four conditions of linear regression are "independence", "linear", "normality", and "equal variance".

3.2.2. Result Evaluation

After determining the regression equation, two aspects can be used to determine if the regression equation effectively reflect the relationship of the variations.

1) Evaluation of Residual Plots

If the points on the residual plot spread at both sides of the O-line, without any regularity, then the regression result is satisfactory from the residual terms. If there is an increasing or decreasing change trend on the residual system plot, then the multiple regression model is not established.

2) Technical Indicators Checking

a) R2 determining factor: Refers to the percentage of total sum of squares of deviations can be explained by the regression line. Therefore, it is a index that reflects the fitting degree of regression linear equation. The closer the points are to the regression line, the closer R2 is to 1, indicating a good fit.

b) Analysis of variance (ANOVA): Use F statistic to test the significance of the entire linear regression equation. The difference in Y is divided into two parts, of which one part can be explained by the variance of X, but the other cannot. These two parts are divided by the variance of degree of freedom, and tested by F value. If there is no significant difference, then the linear relationship is not significant. If there are significant differences, then the fitting result has linear relationship.

c) Significance test of partial regression coefficient: The test is to ascertain whether each independent variable is important to the dependent variable. In the test of partial regression coefficients, assume the overall regression coefficient is zero. When the test shows certain independent variable coefficient is not significant, the corresponding independent variable is considered no effect in the regression equation, and should be removed from the regression equation to establish a simpler regression equation. T statistic is used in the partial regression coefficient test.

d) Normality test of residuals (Residuals): Normality test of residual is very important for the model because the linear model is based on the assumption that the residual is normally distributed. If it is found by test that the residual is not normal, there is no need for further regression analysis. That is to say, only when residual is at or near normal distribution, further analysis can be done.

3.3. Path Analysis

3.3.1 Model and Theory of Path Analysis

In fact, a variable may be the independent variable that affects the other variable, or the dependent variable that is affected by other p variables. There is direct relationship as well as indirect relationship among variables, forming a recursive relationship. The purpose of path analysis is to determine if the causal relationship between variables exists. In the path analysis, the recursive causal model is not determined by the statistical method; it is analyzed by the research theory. Path analysis is to test the rationality of the assumption and indicate the influence degree between variables based on data. Path analysis can be considered as a combination of several regression analyses. It can be divided into one-way path analysis model and the non-single path analysis model.

Take one-way path analysis model for example, list the model (see Fig. **1**) and the path equation.

Path equations of the above figure:

$$X_3 = a_3 + b_{31}X_1 + b_{32}X_2 + b_{3e_3}e_3 \quad (5)$$

$$X_4 = a_4 + b_{41}X_1 + b_{42}X_2 + b_{43}X_3 + b_{4e_4}e_4 \quad (6)$$

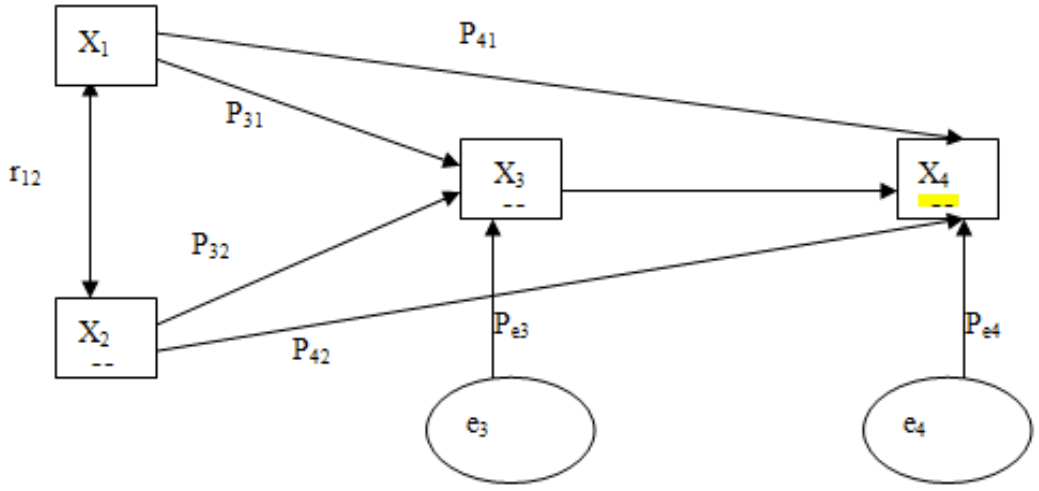

Fig. (1). Single path analysis.

Double arrow means the two variables is related, but not in causal relationship; Pij is the path coefficient, indicating the degree of causal relationship between variables. In the index of path coefficient, the former refers to effect variable, while the latter refers to cause variable; Single arrow indicates the causal direction of the variables.

Assumptions of path analysis model are: The relationship between variables is linear and additive; each variable is in controlled interval; Precedent variable of the endogenous variable is independent from its residual value; the residual values are independent from each other.

3.3.2. Result Evaluation

Path coefficients and residual path coefficients are main indicators that reflect the causal relationship between the variables; if the path coefficient is small, then the causal relationship reflected by the path may not exist. For an ideal path model, the residual path coefficient should be small; if it is large, then some important variable may be missed in the path equation and the causal path need to be re-evaluated.

3.4. Structural Equation Modeling

3.4.1. Basic Theory of Structural Equation Modeling

Structural equation modeling tests the causal relationship between observed variables and latent constructs, and among multiple potential constructs (Crowley & Fan, 1997). Therefore, some scholars believe that structural equation modeling is the sum of confirmatory factor analysis, path analysis, and multiple regression analysis (Schreiberetal. 2006). Unlike traditional regression analysis, structural equation analysis can simultaneously handle multiple dependent variables, and compare and evaluate different theoretical models. Different from traditional exploratory factor analysis, a particular factor structure in structural equation model can be used to test if it is consistent with the data. Through multi-group analysis by structural equation, we can understand whether the relationship between the variables in the different groups remain unchanged, and is there a significant difference between the mean of each factor.

3.4.2. Assumptions of Structural Equation Modeling

Using assumptions of structural equation modeling must meet the following conditions: a) A reasonable sample size. Small sample size may lead to failure on the convergence of calculation, thereby affects the parameter estimation; In particular, when the data is not normally distributed, such as in poor quality or contaminated, larger sample size is needed. b) Consecutive normal endogenous variables. c) Model identification. Compare the number of input that can be used and the number of parameters that need to estimate; unrecognized model would bring failure to parameter estimation. d) Complete data or proper handling of incomplete data. e) Explanation of model and the theory of causal relationship, i.e. the logic of the assumption test.

3.4.3. Operation of Structural Equation Modeling

First, select variables and determine the direction of relationship between variables based on literature review, to design hypothetical theory model. Then, conduct exploratory factor analysis in spss16.0, extract factor, and verify the result of exploratory factor analysis through structural equation modeling. Draw a path diagram (i.e.: a conceptual model) in LISREL 8.80 (STUDENT EDITION), conduct model test, and check the test result. Check if the data fitting of the model is ideal based on the various fitting indicators reported in the LISREL report. If not, amend the model according to the model amending indicators provided in LISREL. Last, repeat the above steps on the modified model until the appropriate model is found. If the modified model is not satisfying, new data must be collected.

4. LIMITATIONS AND APPLICATION OF CAUSAL INFERENCE STATISTICAL METHODS

Sports science studies people, who have biological and social properties. Among the large quantity of random phenomenon and variables in sports, some are normally distributed, while some are not. For example, the values of some power projects (including push-ups and chin-up) are not normally distributed. There are many indicators, including will, mental and emotional factors, technical conditions, and tactical awareness are difficult to quantify. Therefore, to quantify these indicators itself is not reasonable. In addition, a variety of factors in sports tend to be in cross-coverage. The more factors you take, the more you sacrifice the independence of variables. Therefore, the complexity of sports science adds difficulties to the rational use of statistical methods.

Rubin noted that statistical methods can reduce, but not completely eliminate the influence that confounding variables on the results. Statistical methods that can be used for causal inference also have their own limitations and shortcomings. For example, in the assumption test, when testing the significance of the mean, the statistical test methods are different in different conditions where the variance is known or unknown, same or different. Despite these facts, simple application of common formula can only draw statistical results, but not professional results of practical sports.

In the multiple regression analysis, the relationship between sample size n and the number of independent variable k, and the correlation of independent variables should be paid particular attention. If the proportion of the sample size n and the number of independent variables is not appropriate, the solved bi cannot correctly reflect the relationship between the variables and Y. Generally, n should be 5 to 10 times of K; if the independent variables are high correlated, the importance of the independent variables to the dependent variables cannot be correctly reflected. There are some limitations to use structural equation modeling, for example, the conclusion drawn from the given model may not be explained. Sample size could be large; a lot of problems have no reasonable answers and guidelines to follow.

To sum up, these methods have the same logic, but different principles and calculation processes; therefore, research result of high reliability and validity can only be

drawn by full understanding of the applicable conditions of the various methods and comprehensive use of various methods based on actual condition.

CONFLICT OF INTEREST

The author confirms that this article content has no conflicts of interest.

ACKNOWLEDGEMENTS

Declared none.

REFERENCES

[1] F. Ping, X. Duanqin and C. Xuemei, "Development and Application of Structural Equation Model", *Advances in Psychological Science*, vol. 3, 2002.

[2] H. Hai, L. Youfeng and C. Zhiying, "Statistical Analysis -- SPSS for windows", *Beijing: People's Posts and Telecommunications Press*, 2001.

[3] H. Jietai, W. Zhonglin and C. Zijuan, "Structural Equation Model and Its Application", *Beijing: Education Science Press*, 2004.

[4] P.W. Holland, "Statistics and causal inference", *Journal of the American Statistical Association*, vol. 81, pp. 945-970, 1986.

[5] X. Tao, L. Feng, "Causal inference statistical methods under social science background", *Beijing Normal University*, vol. 1, 2009.

[6] L. Hua, "Several Issues Need to be Noted for Statistical Methods in Sports Research", *Academic Journal of Chengdu Institute of Physical Education*, vol. 1, 1993.

[7] Si Dihu, Wang Shengli. Essence of Causal Relationship in Sports Sociological Study . Journal of Physical Education, 2004, (4).

[8] Y. Jing, "On the Application of Causal Relationship Statistical Model in Sports Scientific Research", *Academic Journal of Guangzhou Physical Education Institute*, vol. 4, 2010.

Received: June 10, 2015 Revised: July 29, 2015 Accepted: August 15, 2015