CSCanada
Studies in Mathematical Sciences
Vol. 3, No. 1, 2011, pp. 58-63
DOI: 10.3968/j.sms.1923845220110301.4Z483

ISSN 1923-8444 [Print]
ISSN 1923-8452 [Online]
www.cscanada.net
www.cscanada.org

Study on Linear Correlation Coefficient and Nonlinear Correlation Coefficient in Mathematical Statistics

WANG Ting[1], ZHANG Shiqiang[1,*]

[1]Department of mathematics; Lab. of forensic medicine and biomedicine information, Chongqing Medical University, China, 400016
[*]Corresponding author. E-mail: math808@sohu.com
Received 20 June 2011; accepted 26 July 2011

Abstract: In the two-dimensional or multidimensional experimental data in the traditional statistics, there is usually a linear relationship, or a similar linear relationship between independent variables and the dependent variable. Commonly the linear correlation coefficient is used to measure the degree of linear between the independent variables and the dependent variable. However, for the two-dimensional or multidimensional experimental data, there may be a simple linear relationship between independent variables and the dependent variable, or a simple non-linear relationship, or both linear and non-linear relationship. Article [1] found that the traditional correlation coefficient (linear correlation coefficient) r is only suitable for simple linear relationship. On the basis of article [1], this article discusses the linear correlation coefficient r, analyzes nonlinear correlation coefficient r_{nl}, and gives a new definition of the correlation coefficient R. The new correlation coefficient R can not only describe the case of a simple linear relationship, but also describe the case of a simple nonlinear relationship and the case of both simple linear relationship and nonlinear relationship. That is to say, the new correlation coefficient R can describe the internal law of any experimental data.
Keywords: Mathematical statistics; Correlation coefficient; Linear correlation coefficient; Nonlinear correlation coefficient

WANG Ting, ZHANG Shiqiang (2011). Study on Linear Correlation Coefficient and Nonlinear Correlation Coefficient in Mathematical Statistics. *Studies in Mathematical Sciences, 3*(1), 58-63. Available from: URL: http://www.cscanada.net/index.php/sms/article/view/j.sms.1923845220110301.4Z483. DOI: http://dx.doi.org/10.3968/j.sms.1923845220110301.4Z483.

INTRODUCTION

In the traditional statistics, there is usually a linear relationship, or a similar linear relationship between independent variables and the dependent variable. Commonly the linear correlation coefficient is used to measure the degree of linear between the independent variables and the dependent variable. If the independent variables and the dependent variable show a clear nonlinear relationship, linear correlation coefficient should not be used to measure the relationship between the independent variable and dependent variable. Unfortunately, when the independent variable and the dependent variable show a clear nonlinear relationship, a lot of the literature still uses the linear correlation coefficient to measure the relationship between the independent variable and dependent variable. The reason is that the traditional statistics textbooks discuss

that there is usually a linear relationship, or a similar linear relationship between the independent variable and the dependent variable and the linear correlation coefficient is usually abbreviated as the correlation coefficient. So people are used to dealing with nonlinear problems in linear ways.

Article[1] discussed the linear correlation coefficient in statistics and pointed out that when the formula of linear correlation coefficient applied to linear regression model is perfect, it has its own particularity which is that it can only be used in linear regression mathematical model and can not be extended to nonlinear regression mathematical model. No matter there is a significant linear relationship or a clear nonlinear relationship between independent variables and the dependent variable, can we find a unified correlation coefficient to measure the independent variables and the relationship between the dependent variable? This article will discuss the issue.

1. INTRODUCTION TO LINEAR CORRELATION COEFFICIENT IN TRADITIONAL STATISTICS

In statistics, for a two-dimensional experimental data, if x indicates independent variable, y the dependent variable, we can use a collection $\{(x_i, y_i)|i = 1, 2, 3, \ldots, n\}$to represent the set of experimental data, in which n indicates that the set of experimental data are n pairs, x_i indicates that the experimental data of the independent variable x, y_i indicates that the experimental data of the dependent variable y.

If there is mainly linear relationship between independent variable x and dependent variable y in the set of experimental data, you can find the linear regression model $\hat{y} = a_1 + b_1 x$using the least square method.In the model $\hat{y}_i$ indicates that the predicted value of the experimental data $y_i$of the dependent variable y. How to measure the linear regression mathematical model of the merits of it? In statistics$r$ is used to measure it. Linear correlation coefficient r is calculated as follows:

$$r^2 = \frac{\left[\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})\right]^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{1}$$

here $\bar{x} = \sum_{i=1}^{n} x_i, \bar{y} = \sum_{i=1}^{n} y_i$.

Linear correlation coefficient can be introduced to another calculated as r:

$$r^2 = 1 - \frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

2. INTRODUCTION TO NONLINEAR CORRELATION COEFFICIENT

For a set of two-dimensional experimental data $\{(x_i, y_i)|i = 1, 2, 3, \ldots, n\}$, if the independent variable x and dependent variable y of the set of experimental data are mainly the nonlinear relationship, you can use the least squares method or other ways to find the corresponding linear regression model. How to measure the pros and cons of non-linear regression mathematical model? Obviously can not be used to measure it.

Suppose obtained the nonlinear mathematical model of the experimental data set $\{(x_i, y_i)|i = 1, 2, 3, \ldots, n\}$ obtained by the least square method or other methods is: $\hat{y} = f(x)$. That the collection composed by the

experimental data y_i and the prediction data $\hat{y}_i$ is

$$\{(y_i, \hat{y}_i)|i = 1, 2, 3, \cdots, n\} \equiv \{(y_i, z_i)|i = 1, 2, 3, \cdots, n\}$$

If the non-linear mathematical model of $\hat{y} = f(x)$ is a very good fit, the experimental data y_i and the prediction data z_i should exist the very good linear regression model

$$\hat{z} = a_2 + b_2 y$$

Follow the derivation of linear correlation coefficient r of formula (1),we can deduce the formula of linear correlation coefficient of the collection

$$\{(y_i, \hat{y}_i)|i = 1, 2, 3, \cdots, n\} \equiv \{(y_i, z_i)|i = 1, 2, 3, \cdots, n\}$$

composed of the experimental data y_i and the prediction data $\hat{y}_i = z_i$is

$$r_{nl}^2 = \frac{\left[\sum_{i=1}^{n} (y_i - \bar{y})(z_i - \bar{z})\right]^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2 \sum_{i=1}^{n} (z_i - \bar{z})^2}, \tag{2}$$

here $\bar{y} = \sum_{i=1}^{n} y_i$, $\bar{z} = \sum_{i=1}^{n} z_i$.

Formula (2) can serve as a measure of the merits of non-linear regression mathematical model.We defined r_{nl} of formula (2) as nonlinear correlation coefficient of the experimental data collection $\{(x_i, y_i)|i = 1, 2, 3, \ldots, n\}$. Imitation of statistics, the deduce method of of linear correlation coefficient, you can also launch another formula of calculated linear correlation coefficient r_{nl}:

$$r_{nl}^2 = 1 - \frac{\sum_{i=1}^{n} (\hat{z}_i - z_i)^2}{\sum_{i=1}^{n} (z_i - \bar{z})^2}$$

3. INTRODUCTION TO THE NEW DEFINITION OF THE CORRELATION COEFFICIENT

For a set of two-dimensional experimental data $\{(x_i, y_i)|i = 1, 2, 3, \ldots, n\}$, if there is mainly linear relationship between independent variable x and dependent variable y in the set of the experimental data, you can find the least square method for the linear regression model $\hat{y} = a_1 + b_1 x$,in the model $\hat{y}_i$ indicates that the predicted value of the experimental data y_i of the dependent variable y. How to measure the linear regression mathematical model of the merits of it? We can use the following linear correlation coefficient r to measure.

$$r^2 = 1 - \frac{\sum_{i=1}^{n} (\hat{y}_i - y_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}$$

If there is mainly non-linear relationship between independent variable x and dependent variable y in the set of the experimental data,you can find the least square method or other method for the corresponding

mathematical model of nonlinear regression. How to measure the pros and cons of non-linear regression mathematical model? We can use nonlinear correlation coefficient to measure it.

$$r_{nl}^2 = 1 - \frac{\sum_{i=1}^{n} (\hat{z}_i - z_i)^2}{\sum_{i=1}^{n} (z_i - \bar{z})^2}$$

In solving practical problems, for a set of two-dimensional experimental data $\{(x_i, y_i)|i = 1, 2, 3, \ldots, n\}$, there may be a simple linear relationship, maybe a simple non-linear relationship, or both linear and non-linear relationship. For the latter, there may be major factors showing linear relationship, nonlinear relations are showing a secondary factor; nonlinear relationship may be the main factors showed a linear relationship showing a secondary factor. How to accurately find the main factors rendering the regression model corresponding to it?

Compare the formula (1) calculated the linear correlation coefficient with the formula (2) calculated the non-linear correlation coefficient, then you can find the linear correlation coefficient of experimental data $\{(x_i, y_i)|i = 1, 2, 3, \ldots, n\}$is from formula (1) calculated the linear correlation coefficient£while the linear correlation of the collection

$$\{(y_i, \hat{y}_i)|i = 1, 2, 3, \cdots, n\} \equiv \{(y_i, z_i)|i = 1, 2, 3, \cdots, n\}$$

composed by the experimental data y_i and the forecast data $\hat{y}_i$ is from the formula (2) calculated the non-linear correlation coefficient.

By comparing the above, note that the process of the calculation of r_{nl} is not completed in one step, it is from the start a collection of original experimental data $\{(x_i, y_i)|i = 1, 2, 3, \ldots, n\}$ draw a nonlinear model $\hat{y} = f(x)$,and then from the set that composed of the experimental data y_i and forecast data data$\hat{y}_i$:

$\{(y_i, \hat{y}_i)|i = 1, 2, 3, \cdots, n\} \equiv \{(y_i, z_i)|i = 1, 2, 3, \cdots, n\}$ draw the linear model $\hat{z} = a_2 + b_2 y$.

While correlation coefficient r of the linear function is obtained only from the original experimental collection $\{(x_i, y_i)|i = 1, 2, 3, \ldots, n\}$ directly after fitting a linear model $\hat{y} = a_1 + b_1 x$.

So you can give a new definition of the correlation coefficient.It can be used to accurately find the main factors corresponding regression model main factors.

If a mathematical model$\hat{y} = f(x)$of the original set of experimental data $\{(x_i, y_i)|i = 1, 2, 3, \ldots, n\}$ is obtained in the linear or nonlinear regression method, then you can define the formula of the new correlation coefficient of the regression model as

$$R = \frac{\sum_{i=1}^{n} (y_i - \bar{y})(z_i - \bar{z})}{\left[\sum_{i=1}^{n} (y_i - \bar{y})^2 \sum_{i=1}^{n} (z_i - \bar{z})^2\right]^{\frac{1}{2}}} \tag{3}$$

Here $z_i = \hat{y}_i = f(x_i)$, $\bar{y} = \sum_{i=1}^{n} y_i$,$\bar{z} = \sum_{i=1}^{n} z_i$.

Accordance with the above definition of the new correlation coefficient,between the independent variables and the dependent variable of the original set of experimental data $\{(x_i, y_i)|i = 1, 2, 3, \ldots, n\}$, it can be used the new correlation coefficient to measure ,not only shows clear linear relationship but also nonlinear relationship.

After the new correlation coefficient is given, dealing with practical problems become more convenient.For example, for a given set of original experimental data $\{(x_i, y_i)|i = 1, 2, 3, \ldots, n\}$, when the linear model can

be used both to describe their internal rules, they may also use the nonlinear model to describe it, using the new correlation coefficient R, it can be found the best regression model of describing the rules of the experimental data.

4. APPLICATION EXAMPLE

A set of raw statistical data shown in Table 1

Table 1
Raw Statistical Data

x	1	2	3	4	5	6	7	8
y	63.9	36.0	17.1	10.5	7.3	4.5	2.8	1.7

Does it show clear linear relationship or clear nonlinear relationship between the independent variables and the dependent variable of raw statistical data?Now fitting out of a linear model and two nonlinear model[1,2,3,4]as follows: $\hat{y}_1$=52.3892-7.6476x; $\hat{y}_2$= exp(4.5261-0.5062x); $\hat{y}_3$=113.7959 exp(-0.5812x)

Table 2
Three Mathematical Model Projections and the New Correlation Coefficient ***R***

x	1	2	3	4	5	6	7	8	R
$\hat{y}_1$	44.742	37.094	29.446	21.799	14.151	6.054	-1.144	-8.791	0.865
$\hat{y}_2$	55.696	33.572	20.237	12.198	7.353	4.432	2.671	1.610	0.995
$\hat{y}_3$	63.636	35.588	19.902	11.130	6.234	3.481	1.946	1.088	0.999

Compared the new correlation coefficient R of table 2,it shows a clear nonlinear relationship between the independent variables and the dependent variable of raw statistical data.Set of experimental data describing the internal laws of the regression model should be nonlinear model.

At this point, if you compare the new correlation coefficient data of the two nonlinear model in table2, you will know that the best regression model of the independent variables and the dependent variable of raw statistical data is nonlinear model $\hat{y}_3$=113.7959exp(-0.5812x).

CONCLUSION

In the two-dimensional or multidimensional experimental data in the traditional statistics£there usually is a linear relationship, or a similar linear relationship between independent variable and the dependent variable. Commonly used the linear correlation coefficient to measure the degree of linear between the independent variables and the dependent variable. However, for the two-dimensional or multidimensional experimental data, there may be a simple linear relationship between independent variables and the dependent variable, or a simple non-linear relationship, or both linear and non-linear relationship. For the latter, there may be major factors showing linear relationship, nonlinear relations showing a secondary factor; or nonlinear relationship may be the main factors and linear relationship shows a secondary factor. The traditional correlation coefficient (linear correlation coefficient) r is only suitable for simple linear relationship. This

article gives a definition of the correlation coefficient R. The new correlation coefficient R can not only describe the case of a simple linear relationship, but also describe the case of a simple nonlinear relationship and the case of both simple linear relationship and nonlinear relationship. In all, that the new correlation coefficient R can describe the internal law of any experimental data.

REFERENCES

[1] ZHANG Shiqiang, LU Jieneng, ZHANG Lei and JIANG Zheng (2009). Study of the Correlation Coefficients in Mathematical Statistics. *Mathematics in Practice and Theory, 39*(19), 102-107 (in Chinese).

[2] ZHANG Shiqiang (2009). Modelling Method of Combination Forecast Model Based on Data Mining. *Chinese Journal of Health Statistics, 26*(5), 470-472 (in Chinese).

[3] ZHANG Shiqiang (2002). Approach on the Fitting Optimization Index of Curve Regression. *Chinese Journal of Health Statistics, 19*(1), 19-22 (in Chinese).

[4] ZHANG Shiqiang (1997). The New Approximate Regressive Method for Common Non-linear Function Model and Its Application. *Chinese Journal of Health Statistics, 14*(1), 20-23 (in Chinese).