
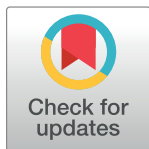


## RESEARCH ARTICLE

## Multiple-to-multiple path analysis model

Yujie Du, Junli Du\*, Xi Liu, Zhifa Yuan \*

College of Sciences, Northwest A&amp;F University, Yangling, P. R. China

\* [djl602@nwafu.edu.cn](mailto:djl602@nwafu.edu.cn) (JD); [liml75@126.com](mailto:liml75@126.com) (ZY)

## Abstract

One-to-multiple path analysis model describes the regulation mechanism of multiple independent variables to one dependent variable by dividing the correlation coefficient and the determination coefficient. How to analyse more complex regulation mechanisms of multiple independent variables to multiple dependent variables? Similarly, according to multiple-to-multiple linear regression analysis, multiple-to-multiple path analysis model was proposed in this paper and it demonstrated more complex regulation mechanisms among multiple independent variables and multiple dependent variables by dividing the generalized determination coefficient. Differently, three other types of paths were generated in multiple-to-multiple path analysis model in that the correlation among multiple dependent variables was considered. Then, the decision coefficient of each independent variable was constructed for dependent variables system, and its hypothesis testing statistics were given. Finally, the research example of the wheat breeding rules in arid area demonstrated that the multiple-to-multiple path analysis considering more correlation information can get better results.

 OPEN ACCESS

**Citation:** Du Y, Du J, Liu X, Yuan Z (2021) Multiple-to-multiple path analysis model. PLoS ONE 16(3): e0247722. <https://doi.org/10.1371/journal.pone.0247722>

**Editor:** Mohammadreza Hadizadeh, Central State University & Ohio University, UNITED STATES

**Received:** November 25, 2020

**Accepted:** February 11, 2021

**Published:** March 4, 2021

**Copyright:** © 2021 Du et al. This is an open access article distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

**Data Availability Statement:** The data underlying the results presented in the study are available from: Zhang Z, Wang D. *Wheat drought-resistant ecological breeding*. Xi'an: Shaanxi People's Education Press; 1992. China.

**Funding:** This work was financially supported by Chinese Universities Scientific Fund (Grant Nos. 2452015082 and Z1090219004). The funders had no role in decision to publish, or preparation of the manuscript.

**Competing interests:** The authors have declared that no competing interests exist.

## 1 Introduction

The regression analysis, as one of the most widely used statistical methodologies, focuses on studying the relations between dependent variables and independent variables. However, the regression analysis worries less about the correlation mechanisms that may exist among the independent variables [1]. In 1918–1921, the issue was addressed by the biological geneticist Sewall Wright through developing the path analysis method [2, 3]. Sewall Wright's path analysis mainly emphasizes decomposing the correlation and total determination in terms of model parameters, and drawing the path diagram. The path diagram is a pictorial representation of a system of simultaneous equations, which presents the picture of the relationships that are assumed and is more clearly than the equations [4]. The concrete decomposition result is to distinguish the three types of effects: direct, indirect and total effects, which can lead to a more comprehensive understanding of the relation between variables. Usually, the indirect effects of a variable are mediated by at least one intervening variable [4]. In fact, the decomposed indirect effects quantify the regulation of variables with correlation. The quantitative expression of regulatory mechanism can make the analysis more thorough and clear. Therefore, the path analysis was later applied in multiple science research fields, such as behavioural science, social science, economics, biology, agriculture, medical science and so on [5–18]. This method seems to be more and more widely used at present.

In terms of methodology research, the path analysis was generalized to the structural equation models (SEMs) through combining the principle of factor analysis and was used to analyse the relations between multivariate blocks of data [19, 20]. The decision coefficient was constructed in the specified path analysis model with no latent variables, which included one dependent variable (as result) and multiple independent variables (as causes), based on the decomposition of total determination coefficient [21]. Here, the specific path analysis model was called one-to-multiple path analysis model with the nature of standard multiple linear regression. The decision coefficient of each independent variable equals to the sum of its direct determination and the correlation indirect determination with the other independent variables. The decision coefficient can express the magnitude and direction of each independent variable influencing the variation of dependent variable. Still further, the importance of each independent variable for dependent variable can be ranked according to the decision coefficient result, which shows that the decision coefficient has the significance of making decisions. Subsequently, the statistical test of the decision coefficient was proposed [22]. The decision coefficient improves the one-to-multiple path analysis model to a certain extent. Later, the one-to-multiple path analysis model was applied in the lint yield of upland cotton research and the KEGG gene pathway regulation mechanisms research [23–25].

However, the causal system including multiple independent variables (as “causes”) and multiple dependent variables (as “results”) are often encountered in practice research. For instance, the different pathways contain the same genes in the KEGG pathway, which demonstrated that the same genes can lead to the different gene functions. Here, multiple identical genes and multiple different gene functions constitute a multiple-to-multiple system. Analysis of the regulatory relationship between genes and gene functions is helpful to the modification and change of gene structure. Similar to this, in breeding field, multiple biological shapes to multiple yield indicators also constitute a multiple-to-multiple system. Determining the importance of multiple biological shapes to multiple yield indicators is helpful to improve the yield and quality of crops. It is assumed that such a causal system does not contain latent variables. Then, the one-to-multiple path analysis model can be used to analyse the importance of each independent variable to one dependent variable and the regulations among multiple independent variables. But, it is frustrating that the results of multiple single one-to-multiple path analysis are often contradictory, so that decision makers feel confused when making decisions. Therefore, it is urgent to find a more suitable model to provide more clear decision-making suggestions for decision-makers in such a more complex system.

In this paper, we attempt to propose the multiple-to-multiple path analysis model according to the multiple-to-multiple linear regression analysis, including multiple independent variables and multiple dependent variables and no latent variables. This model considers the correlation among multiple dependent variables caused by multiple common independent variables on the basis of one-to-multiple path analysis model. The other three types of paths generated besides the two types of paths in one-to-multiple path analysis model. The decomposition of the generalized determination coefficient showed the regulation mechanisms among the multiple independent variables and multiple dependent variables along these five types of paths. And the decision coefficient of each independent variable was used to judge its importance for all dependent variables system. Finally, the effectiveness of the model was verified by an example of the wheat breeding rules in arid area.

## 2 Method

### 2.1 Equations and models

The multiple-to-multiple linear regression model is the basis of the multiple-to-multiple path analysis, so it was introduced firstly. Define the following assumptions: the dependent variable

of linear regression is  $Y = (Y_1, Y_2, \dots, Y_p)^T$  and the independent variable is  $X = (X_1, X_2, \dots, X_m)^T$ . Suppose the joint distribution of  $X_{m \times 1}$  and  $Y_{p \times 1}$  is:

$$\begin{bmatrix} X \\ Y \end{bmatrix} \sim N_{m+p} \left( \begin{bmatrix} \mu_x \\ \mu_y \end{bmatrix}, \begin{bmatrix} \sum_x & \sum_{xy} \\ \sum_{yx} & \sum_y \end{bmatrix} \right) = N_{m+p}(\mu, \Sigma) \tag{1}$$

$\Sigma_{xy} \neq 0$ , when both have been normalized, the joint distribution above becomes

$$\begin{bmatrix} X \\ Y \end{bmatrix} \sim N_{m+p} \left( \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \rho_x & \rho_{xy} \\ \rho_{yx} & \rho_y \end{bmatrix} \right) = N_{m+p}(0, \rho) \tag{2}$$

Among them,  $\rho_x, \rho_{xy}$  and  $\rho_y$  are the correlation arrays of  $X, X$  and  $Y, Y$  respectively.  $\rho$  is the correlation matrix of  $[X^T, Y^T]^T$ . Under the above assumption, the normalized multiple-to-multiple linear regression model is:

$$\begin{bmatrix} Y_{i1} \\ Y_{i2} \\ \vdots \\ Y_{ip} \end{bmatrix} = \begin{bmatrix} \beta_{11}^* & \beta_{12}^* & \dots & \beta_{1p}^* \\ \beta_{21}^* & \beta_{22}^* & \dots & \beta_{2p}^* \\ \vdots & \vdots & \ddots & \vdots \\ \beta_{m1}^* & \beta_{m2}^* & \dots & \beta_{mp}^* \end{bmatrix}^T \begin{bmatrix} x_{i1} \\ x_{i2} \\ \vdots \\ x_{im} \end{bmatrix} + \begin{bmatrix} \epsilon_{i1} \\ \epsilon_{i2} \\ \vdots \\ \epsilon_{ip} \end{bmatrix} = \begin{bmatrix} \beta_1^{*T} x_i \\ \beta_2^{*T} x_i \\ \vdots \\ \beta_p^{*T} x_i \end{bmatrix} + \begin{bmatrix} \epsilon_{i1} \\ \epsilon_{i2} \\ \vdots \\ \epsilon_{ip} \end{bmatrix}$$

$$\Rightarrow Y_i = \beta^{*T} x_i + \epsilon_i \Rightarrow Y = \beta^{*T} X + \epsilon \tag{3}$$

In (3),  $Y_i = [Y_{i1}, Y_{i2}, \dots, Y_{ip}]^T$ ,  $x_i = [x_{i1}, x_{i2}, \dots, x_{im}]^T$ ,  $\beta^*$  is the regression parameter of the model,  $i = 1, 2, \dots, n$ . Let  $n$  be the number of observations. We assumed that  $\epsilon \sim N_p(0, \Sigma_\epsilon)$  is the regression residual and has nothing to do with the value of  $X$ .

### 2.2 Regression hypothesis testing

Path analysis can only be carried out when the standardized regression equation is significant. Therefore, we need to perform the following four types of hypothesis tests for regression analysis before path analysis.

**2.2.1 Hypothesis testing of generalized complex correlation coefficient  $r_{xy}$ .** In multiple-to-multiple standardized linear regression equations, the joint distribution of  $X$  and  $Y$  is showed as formula (2), then the generalized determination coefficient is defined as [26]:

$$R^2 = 1 - v_{xy} = 1 - \frac{|R|}{|R_{xx}| |R_{yy}|} = 1 - |I_p - B| \approx tr(B) - \sum_{i \neq l} \lambda_i^2 \cdot \lambda_l^2 \tag{4}$$

In Eq (4),  $v_{xy}$  is the likelihood ratio statistics for testing independence of  $X$  and  $Y$ . And  $R_{xx} = \hat{\rho}_x, R_{yy} = \hat{\rho}_y, R$  in  $|R|$  is the correlation matrix of  $X$  and  $Y$ .  $B = \sum_y^{-1} \sum_{yx} \sum_x^{-1} \sum_{xy} = R_{yy}^{-1} U$  is the sample linear correlation matrix of  $X$  and  $Y$ ,  $U$  is the regression square sum matrix.  $\lambda_i^2$  and  $\lambda_l^2$  are non-zero characteristic roots of  $B$ .  $r_{xy} = R_{(x_1 x_2 \dots x_m)(y_1 y_2 \dots y_p)} = \sqrt{R^2}$  is the generalized complex correlation coefficient of  $X$  and  $Y$ . The invalid assumption of  $r_{xy}$  is  $H_0: \Sigma_{xy} = 0$ . When  $p > 2$  and  $m > 2$ , we can use Bartlett's approximate chi-square test:

$$V = - \left( n - 1 - \frac{p + m + 1}{2} \right) \ln v_{xy} \sim \chi^2(pm) \tag{5}$$

**2.2.2 Hypothesis testing of regression equation  $\hat{y}_\alpha = \mathbf{b}_\alpha^T \mathbf{x}$  ( $\alpha = 1, 2, \dots, p$ ).** The invalid hypothesis is  $H_0 : \beta_\alpha^* = 0$  and the corresponding  $F$  test statistic is:

$$F = \frac{R_{(\alpha)}^2/m}{(1 - R_{(\alpha)}^2)/(n - m - 1)} \sim F(m, n - m - 1) \tag{6}$$

$R_{(\alpha)}^2$  is the determination coefficient of  $X$  to  $Y_\alpha$

**2.2.3 Hypothesis testing of components  $\mathbf{b}_{j\alpha}^*$  in  $\mathbf{b}_\alpha^*$ .** The invalid hypothesis is  $H_{0j\alpha} : \beta_{j\alpha}^* = 0$  and the  $t$  test statistic is

$$t_{j\alpha} = \frac{b_{j\alpha}^*}{\sqrt{c_{jj} \hat{\sigma}_\alpha^2}} \sim t(n - m - 1) \tag{7}$$

In (7),  $\sum_e = \frac{Q_e}{n-m-1}$ ,  $\hat{\sigma}_\alpha^2$  is the  $\alpha$ -th element on the main diagonal of  $\sum_e$ ,  $c_{jj}$  is the  $j$ -th element on the main diagonal of  $R_{xx}^{-1}$ .

**2.2.4 Hypothesis testing of  $\mathbf{b}_{xy}^*$  [27].** The invalid hypothesis is  $H_0 : (\beta_{j1}^*, \beta_{j2}^*, \beta_{j3}^*)^T = 0$ .

The  $F$  test statistic is:

$$F = (n - m - 2) \times \frac{1 - \sqrt{\Lambda_{H_{0j}}}}{\sqrt{\Lambda_{H_{0j}}}} \sim F[2, 2(n - m - 2)] \tag{8}$$

In (8),  $\Lambda_{H_{0j}} = \frac{|Q_e|}{|Q_{H_{0j}e}|}$ . After the above four hypothesis tests, if the standardized multiple linear regression equation is significant, it is meaningful to perform path analysis.

**2.3 Path analysis of  $Y_\alpha = \beta_\alpha^T \mathbf{x} + \varepsilon_\alpha$  ( $\alpha = 1, 2, \dots, p$ )**

The first step of multiple-to-multiple path analysis is to conduct one-to-multiple path analysis for each dependent variable and all independent variables. According to the established multiple-to-multiple linear regression equation, the path analysis model is performed. The correlation coefficient  $r_{jy_\alpha}$  of each dependent variable  $Y_\alpha$  ( $\alpha = 1, \dots, p$ ) and all independent variables  $X = (X_1, X_2, \dots, X_m)^T$  and their determination coefficient  $R_{(\alpha)}^2$  were divided following completely the previous one-to-multiple path analysis model on the basis of standardized linear regression equation [25]. Still further, the decision coefficient was constructed using the existing method [21]. According to the theoretical study of multiple linear regression analysis, the system of regular equations  $R_{xx} \mathbf{b}^* = R_{xy}$  about the least squares estimation of  $\beta^*$  can be rewritten as:

$$R_{xx} (b_1^*, b_2^*, \dots, b_p^*)^T = (R_{xy_1}, R_{xy_2}, \dots, R_{xy_p}) = R_{xy}$$

So

$$R_{xx} b_\alpha^* = R_{xy_\alpha}, \alpha = 1, 2, \dots, p \tag{9}$$

In Eq (9),  $R_{xx} = \hat{\rho}_x$ ,  $R_{xy_\alpha} = \hat{\rho}_{xy_\alpha}$ . The specific path diagram of the one-to-multiple path analysis model is shown in Fig 1.

2.3.1 The division and path of  $r_{jy_\alpha}$  ..

$$\begin{aligned}
 r_{jy_\alpha} &= b_{j\alpha}^* + \sum_{k \neq j} r_{jk} b_{k\alpha}^* \\
 j &= 1, 2, \dots, m; k = 1, 2, \dots, m; \alpha = 1, 2, \dots, p; \\
 b_{\varepsilon_\alpha \rightarrow y_\alpha}^* &= \sqrt{1 - R_{(\alpha)}^2}
 \end{aligned}
 \tag{10}$$

Obviously, the correlation efficient  $r_{jy_\alpha}$  was divided into  $m$  terms. There are two types for this  $m$  term:  $b_{j\alpha}^*$  is formed by the path  $y_\alpha \leftarrow x_j$ , so  $b_{j\alpha}^*$  is called the direct effect of  $x_j$  on  $y_\alpha$ ; and  $r_{jk} b_{k\alpha}^* (k \neq j)$  is formed by  $x_j \leftrightarrow x_k \rightarrow y_\alpha$  which is the effect of  $x_j$  on  $y_\alpha$  through the correlation with  $x_k$  and called the indirect effect. Its magnitude can be obtained by multiplying the path coefficients  $b_{k\alpha}^*$  by the correlation coefficient  $r_{jk}$ , including  $m-1$  items. Finally,  $r_{jy_\alpha}$  is the total effect of  $x_j$  on  $y_\alpha$ , which is the sum of the direct effect and all the indirect effects.

2.3.2 The division and path of  $R_{(\alpha)}^2$ .

$$\begin{aligned}
 R_{(\alpha)}^2 &= b_{\alpha}^{*T} R_{xx} b_{\alpha}^* = \sum_{j=1}^m b_{j\alpha}^{*2} + \sum_{j=1}^{m-1} 2 b_{j\alpha}^* r_{jk} b_{k\alpha}^* \\
 &= \sum_{j=1}^m R_{j(\alpha)}^{*2} + \sum_{k>j} R_{jk(\alpha)}^*
 \end{aligned}
 \tag{11}$$

Among (11):  $R_{(\alpha)}^2$  is the total coefficient of determination of  $X$  for  $Y_\alpha$ .  $R_{j(\alpha)}^{*2} = b_{j\alpha}^{*2}$  and its corresponding path is  $y_\alpha \leftarrow x_j \rightarrow y_\alpha$ . It is called the direct determination coefficient of  $x_j$  to  $y_\alpha$ . The corresponding path of  $R_{jk(\alpha)}^* = 2b_{j\alpha}^* r_{jk} b_{k\alpha}^*$  is  $y_\alpha \leftarrow x_j \leftrightarrow x_k \rightarrow y_\alpha$ . It is called the correlation determination coefficient of  $x_j$  through the correlation with  $x_k (k \neq j)$  to  $y_\alpha$ .

2.3.3 The decision coefficient  $R_{\alpha(j)}$  and hypothesis test [22]. The comprehensively determine ability of  $x_j$  to  $y_\alpha$  can be represented by the decision coefficient based on the division of  $R_{(\alpha)}^2$ . Its specific expression and hypothesis test are:

$$\begin{aligned}
 R_{\alpha(j)} &= 2b_{j\alpha}^* r_{jy_\alpha} - b_{j\alpha}^{*2} = R_{j(\alpha)}^{*2} + \sum_{k \neq j} R_{jk(\alpha)}^* \\
 t_{\alpha(j)} &= \frac{R_{\alpha(j)}}{S_{R_{\alpha(j)}}} = \frac{R_{\alpha(j)}}{2|r_{jy_\alpha} - b_{j\alpha}^*| \sqrt{\frac{c_{jj}(1 - R_{(\alpha)}^2)}{n - m - 1}}} \sim t(n - m - 1) \\
 j &= 1, 2, \dots, m; \alpha = 1, 2, \dots, p
 \end{aligned}
 \tag{12}$$

The definition indicates that  $R_{\alpha(j)}$  equals to the sum of the direct determination coefficient  $R_{j(\alpha)}^{*2} = b_{j\alpha}^{*2}$  and the correlation determination coefficient  $R_{jk(\alpha)}^* = 2b_{j\alpha}^* r_{jk} b_{k\alpha}^* (k \neq j)$ . In fact, the decision coefficient is the sum of all determination coefficients related to  $x_j$ . The decision coefficient was used to determine the main decision variables and restrictive variables affecting  $Y_\alpha$ .

2.4 Multiple-to-multiple path analysis central theorem

The second step is to conduct multiple-to-multiple path analysis. And the innovation is that the correlation between  $Y$  caused by the common cause  $X$  is considered and three other types of paths are generated. For convenience of observation, let  $p = 3, m = 3$  as an example to make a multiple-to-multiple path analysis diagram as Fig 2. But, the theoretical analysis is based on  $m$  independent variables and  $p$  dependent variables.

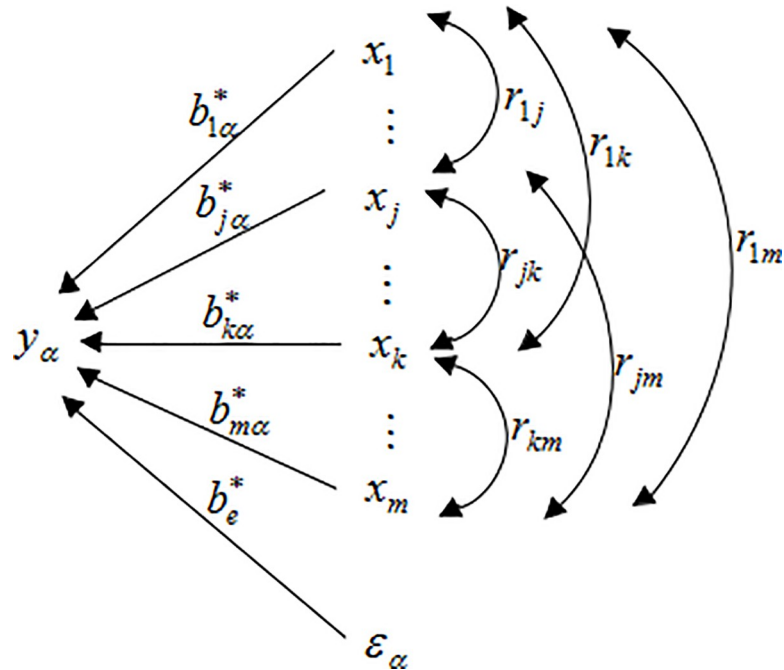


Fig 1. One-to-multiple path analysis diagram.

<https://doi.org/10.1371/journal.pone.0247722.g001>

The multiple-to-multiple path analysis model considered the correlation among different dependent variables compared to the one-to-multiple path analysis model. Accordingly, the central theorem of multiple-to-multiple path analysis is proposed. Based on model (3), for two different  $Y_\alpha$  and  $Y_t$ , their models are:

$$\begin{cases} Y_\alpha = \beta_\alpha^{*T} X + \epsilon_\alpha, \epsilon_\alpha = Y_\alpha - E(Y_\alpha | X = x) \\ Y_t = \beta_t^{*T} X + \epsilon_t, \epsilon_t = Y_t - E(Y_t | X = x) \end{cases} \quad (13)$$

In (13),  $\epsilon_\alpha$  and  $\epsilon_t$  are independent of each other and have nothing to do with the value of  $X$ . Since  $Y_\alpha$ ,  $Y_t$  and  $X$  have been standardized, the correlation coefficients of  $Y_\alpha$  and  $Y_t$ , and the

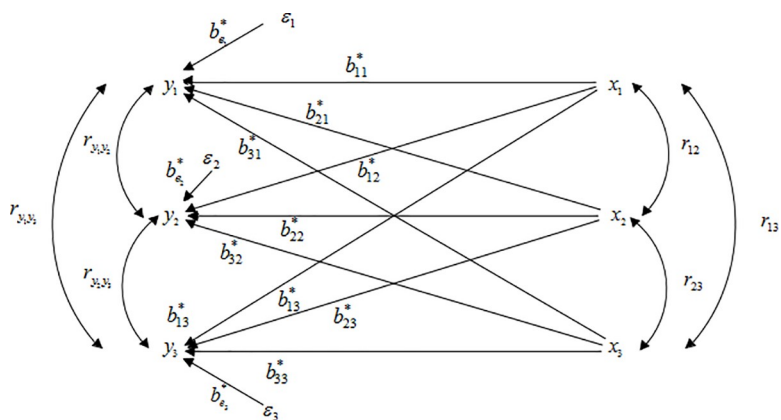


Fig 2. Multiple-to-multiple path analysis diagram.

<https://doi.org/10.1371/journal.pone.0247722.g002>

corresponding path theoretically is:

$$\begin{aligned}
 & \rho_{Y_\alpha Y_t} \\
 &= \text{Cov}(\beta_\alpha^{*T} X + \varepsilon_\alpha, \beta_t^{*T} X + \varepsilon_t) = \beta_\alpha^{*T} \text{Cov}(X, X) \beta_t^* = \beta_\alpha^{*T} \rho_x \beta_t^* \\
 &= \sum_{j=1}^m \beta_{j\alpha}^{*T} \beta_{jt}^{*T} + \sum_{k \neq j} \left( \beta_{j\alpha}^* \rho_{jk} \beta_{kt}^* + \beta_{k\alpha}^* \rho_{kj} \beta_{jt}^* \right) \\
 &= \rho_{y_\alpha(x) y_t(x)} \\
 & \quad y_\alpha(x) \leftrightarrow y_t(x)
 \end{aligned} \tag{14}$$

Considering the sample case, Eq (14) is:

$$\begin{aligned}
 & r_{y_\alpha y_t} \\
 & \quad y_\alpha \leftrightarrow y_t \\
 &= \text{Cov}(b_\alpha^{*T} x + \varepsilon_\alpha, b_t^{*T} x + \varepsilon_t) = b_\alpha^{*T} R_{xx} b_t^* \\
 &= \sum_{j=1}^m b_{j\alpha}^* b_{jt}^* + \sum_{k \neq j} \left( b_{j\alpha}^* r_{jk} r_{kt}^* + b_{k\alpha}^* r_{kj} b_{jt}^* \right)
 \end{aligned} \tag{15}$$

Among them,  $j = 1, 2, \dots, m$ ;  $k = 1, 2, \dots, m$ ; and  $\alpha = 1, 2, \dots, p$ ;  $t = 1, 2, \dots, p$ . Eq (14) and Eq (15) are called the central theorem of multiple-to-multiple path analysis.

The central theorem demonstrated that  $\rho_{Y_\alpha Y_t}, r_{y_\alpha y_t}$  equal to the sum of  $m^2$  items composite path coefficient. Wherein, the direct path  $y_\alpha \leftarrow x_j \rightarrow y_t$  has  $m$  items. Due to the correlation among independent variables  $x_j \leftrightarrow x_k (k \neq j)$ , the two types of indirect paths were formed as  $y_\alpha \leftarrow x_j \leftrightarrow x_k \rightarrow y_t, y_\alpha \leftarrow x_k \leftrightarrow x_j \rightarrow y_t$ . And  $x_j \leftrightarrow x_k (k \neq j)$  has  $C_m^2 = \frac{1}{2} m(m - 1)$  items. So the total composite path number is  $m + 2 \times \frac{1}{2} m(m - 1) = m^2$  items. In addition, the central theorem also showed that three other types of paths generated when the correlation between different dependent variables  $y_\alpha$  and  $y_t$  was considered, which was caused by the common  $X$ . Therefore, there are five types of paths in multiple-to-multiple path analysis, plus the two types of paths in one-to-multiple path analysis.

In fact, the correlation coefficient in the multiple-to-multiple path analysis central theorem is theoretically the regression square sum matrix  $U$  in multiple-to-multiple standardized linear regression. Under the least squares estimation,  $U$  can be expressed as follows [16]:

$$\begin{aligned}
 U_{Y \leftarrow X \rightarrow Y} &= U_{\hat{y} \leftarrow X \rightarrow \hat{y}} = b^{*T} R_{xy} = b^{*T} R_{xx} b^* \\
 &= \begin{bmatrix} b_1^{*T} R_{xx} b_1^* & b_1^{*T} R_{xx} b_2^* & \dots & b_1^{*T} R_{xx} b_p^* \\ b_2^{*T} R_{xx} b_1^* & b_2^{*T} R_{xx} b_2^* & \dots & b_2^{*T} R_{xx} b_p^* \\ \vdots & \vdots & & \vdots \\ b_p^{*T} R_{xx} b_1^* & b_p^{*T} R_{xx} b_2^* & \dots & b_p^{*T} R_{xx} b_p^* \end{bmatrix} = \begin{bmatrix} R_{(1)}^2 & r_{y_1 y_2(x)} & \dots & r_{y_1 y_p(x)} \\ \hat{y}_1 \leftarrow X \rightarrow \hat{y}_1 & \hat{y}_1 \leftarrow X \rightarrow \hat{y}_2 & & \hat{y}_1 \leftarrow X \rightarrow \hat{y}_p \\ r_{y_2 y_1(x)} & R_{(2)}^2 & \dots & r_{y_2 y_p(x)} \\ \hat{y}_2 \leftarrow X \rightarrow \hat{y}_1 & \hat{y}_2 \leftarrow X \rightarrow \hat{y}_2 & & \hat{y}_2 \leftarrow X \rightarrow \hat{y}_p \\ \vdots & \vdots & & \vdots \\ r_{y_p y_1(x)} & r_{y_p y_2(x)} & \dots & R_{(p)}^2 \\ \hat{y}_p \leftarrow X \rightarrow \hat{y}_1 & \hat{y}_p \leftarrow X \rightarrow \hat{y}_2 & & \hat{y}_p \leftarrow X \rightarrow \hat{y}_p \end{bmatrix} \tag{16}
 \end{aligned}$$

In (16),  $b_\alpha^{*T} R_{xx} b_t^* = r_{y_\alpha y_t(x)}, \alpha \neq t$  is the correlation coefficient between  $y_\alpha$  and  $y_t$  caused by the common cause  $X$ . Here,  $U$  is the determination coefficient matrix of  $X$  to  $Y$ .  $R_{(\alpha)}^2 = r_{y_\alpha y_\alpha(x)}^2$  is the coefficient of determination of  $X$  to  $Y_\alpha$ .  $\sqrt{R_{(\alpha)}^2} = r_{y_\alpha(x_1 x_2 \dots x_m)} = r_{y_\alpha(x)}, \alpha = 1, 2, \dots, p$  is the complex correlation coefficient of  $X$  to  $Y_\alpha$ . And in statistics,  $r_{Y_\alpha Y_t}$  is the correlation coefficient

of  $Y_\alpha$  and  $Y_t$ , and has nothing to do with  $X$  in the calculation.  $r_{\hat{y}_\alpha \hat{y}_t(\alpha)}$  is the determining part of  $Y_\alpha$  and  $Y_t$  to  $r_{Y_\alpha Y_t}$  due to the common cause  $X$ .

### 2.5 The division of $R^2 \approx tr(B)$ and its corresponding path

The generalized determination coefficient has been defined using formula (4) before, which was used to reflect the comprehensive determination of all independent variables to all dependent variables [26]. Because the non-zero eigenvalue  $\lambda_t^2 (t = 1, 2, \dots, k)$  of  $B$  is small and  $0 < \lambda_k^2 \leq \lambda_{k-1}^2 \leq \dots \leq \lambda_1^2 \leq 1$ , the result of  $\sum_{t \neq 1} \lambda_t^2 \lambda_1^2$  is small enough to make  $R^2 \approx tr(B)$ . In fact,  $tr(B)$  is the overestimation of  $R^2$  here. According to  $R^2 \approx tr(B)$ , the generalized determination coefficient  $R^2$  was divided as follows:

$$\begin{aligned}
 R^2 \approx tr(B) &= \sum_{\alpha=1}^p \theta_{\alpha\alpha} R_{\alpha(\alpha)}^2 + 2 \sum_{t>\alpha} \theta_{\alpha t} r_{y_\alpha y_t(x)} \\
 &= \sum_{\alpha=1}^p \theta_{\alpha\alpha} \left( \sum_{j=1}^m b_{j\alpha}^{*2} + 2 \sum_{\substack{j=1 \\ k>j}}^{m-1} b_{j\alpha}^* r_{jk} b_{k\alpha}^* \right) \\
 &+ \sum_{t>\alpha} \theta_{\alpha t} \left[ \sum_{j=1}^m 2 b_{j\alpha}^* b_{jt}^* + \sum_{k \neq j} \left( 2 b_{j\alpha}^* r_{jk} b_{kt}^* + 2 b_{k\alpha}^* r_{kj} b_{jt}^* \right) \right] \\
 &= \sum_{j=1}^m \sum_{\alpha=1}^p \theta_{\alpha\alpha} R_{j(\alpha)}^2 + \sum_{k<j} \sum_{\alpha=1}^p \theta_{\alpha\alpha} R_{jk(\alpha)} \\
 &+ \sum_{t>\alpha} \theta_{\alpha t} \left[ \sum_{j=1}^m R_{j(\alpha t)} + \sum_{k \neq j} \left( R_{jk(\alpha t)} + R_{kj(\alpha t)} \right) \right] \tag{17}
 \end{aligned}$$

Among (17),  $\theta_{\alpha\alpha}$  is the element in matrix  $R_{yy}^{-1}$ .  $R_{j(\alpha)}^2 = b_{j\alpha}^{*2}$  is the direct determination coefficient of  $x_j$  on  $y_\alpha$  and the effect path is  $y_\alpha \leftarrow x_j \rightarrow y_\alpha$   $j = 1, 2, \dots, m; \alpha = 1, 2, \dots, p$ .  $R_{jk(\alpha)} = 2b_{j\alpha}^* r_{jk} b_{k\alpha}^*$  is the indirect determination coefficient of  $x_j$  and  $x_k$  on  $y_\alpha$ , the effect path is  $y_\alpha \leftarrow x_j \leftrightarrow x_k \rightarrow y_\alpha$ ,  $jk$  has  $\frac{1}{2}m(m - 1)$  items.  $R_{j(\alpha t)} = 2b_{j\alpha}^* b_{jt}^*$  is the direct determination coefficient of  $x_j$  on  $y_\alpha$  and  $y_t$ , which is caused by the correlation of  $y_\alpha$  and  $y_t$  because of the common cause  $x_j$ . The effect path is  $y_\alpha \leftarrow x_j \rightarrow y_t$ ,  $j = 1, 2, \dots, m, \alpha t$  has  $\frac{1}{2}p(p - 1)$  items.  $R_{jk(\alpha t)} = 2b_{j\alpha}^* r_{jk} b_{kt}^*$  is the indirect determination coefficient of  $x_j$  and  $x_k$  on  $y_\alpha$  and  $y_t$ . The effect path is  $y_\alpha \leftarrow x_j \leftrightarrow x_k \rightarrow y_t$ . When  $\alpha < t$ ,  $y_\alpha$  and  $y_t$  have  $\frac{1}{2}p(p - 1)$  items; when  $j \neq k$ ,  $jk$  has  $\frac{1}{2}m(m - 1)$  items.  $R_{kj(\alpha t)} = 2b_{k\alpha}^* r_{kj} b_{jt}^*$  is the indirect determination coefficient of  $x_j$  and  $x_k$  on  $y_t$  and  $y_\alpha$ . The effect path is  $y_\alpha \leftarrow x_k \leftrightarrow x_j \rightarrow y_t$ , when  $\alpha < t$ ,  $y_\alpha$  and  $y_t$  have  $\frac{1}{2}p(p - 1)$  items; when  $j \neq k$ ,  $kj$  has  $\frac{1}{2}m(m - 1)$  items. Therefore, the total number of items divided is:

$$\begin{aligned}
 &p \left( m + \frac{1}{2}m(m - 1) \right) + \frac{1}{2}p(p - 1) \left( m + \frac{1}{2}m(m - 1) \right) \\
 &= \frac{pm}{4} (m + 1)(p + 1)
 \end{aligned}$$

Formula (17) demonstrates that the generalized determination coefficient  $R^2$  was divided successfully along the five types of paths stated in multiple-to-multiple path analysis central



theorem. The specific path vector structure is:

$$\begin{aligned}
 & \underset{p \times 1}{y} \overset{R^2}{\leftrightarrow} \underset{m \times 1}{x} \approx tr(B) \\
 & = (\theta_{11}, \theta_{22}, \dots, \theta_{pp}) \times \left( \sum_{j=1}^m \begin{bmatrix} R_{j(1)}^2 \\ R_{j(2)}^2 \\ \vdots \\ R_{j(p)}^2 \end{bmatrix}_{\substack{y \leftarrow x_j \rightarrow y \\ p \times 1}} + \sum_{k>j} \begin{bmatrix} R_{jk(1)} \\ R_{jk(2)} \\ \vdots \\ R_{jk(p)} \end{bmatrix}_{\substack{y \leftarrow x_j \leftrightarrow x_k \rightarrow y \\ p \times 1}} \right) \\
 & + (\theta_{12}, \theta_{13}, \dots, \theta_{(p-1)p}) \times \left( \sum_{j=1}^m \begin{bmatrix} R_{j(12)} \\ R_{j(13)} \\ \vdots \\ R_{j((p-1)p)} \end{bmatrix}_{\substack{y_a \leftarrow x_j \rightarrow y_t \\ (a < t)}} + \sum_{k>j} \begin{bmatrix} R_{jk(12)} + R_{kj(12)} \\ R_{jk(13)} + R_{kj(13)} \\ \vdots \\ R_{jk((p-1)p)} + R_{kj((p-1)p)} \end{bmatrix}_{\substack{y_a \leftarrow \begin{pmatrix} x_j \leftrightarrow x_k \\ x_k \leftrightarrow x_j \end{pmatrix} \rightarrow y_t}} \right) \tag{18}
 \end{aligned}$$

### 2.6 The generalized decision coefficient $R_{y(j)}$

**2.6.1 The definition of  $R_{y(j)}$ .** In order to describe the comprehensive decision-making ability of  $x_j$  to  $Y$ , the generalized decision coefficient  $R_{y(j)}$  was defined as follows:

$$\begin{aligned}
 R_{y(j)} & = (\theta_{11}, \theta_{22}, \dots, \theta_{pp}) \times \left( \begin{bmatrix} R_{j(1)}^2 \\ R_{j(2)}^2 \\ \vdots \\ R_{j(p)}^2 \end{bmatrix}_{\substack{y \leftarrow x_j \rightarrow y \\ p \times 1}} + \sum_{k \neq j} \begin{bmatrix} R_{jk(1)} \\ R_{jk(2)} \\ \vdots \\ R_{jk(p)} \end{bmatrix}_{\substack{y \leftarrow x_j \leftrightarrow x_k \rightarrow y \\ p \times 1}} \right) \\
 & + (\theta_{12}, \theta_{13}, \dots, \theta_{(p-1)p}) \times \left( \begin{bmatrix} R_{j(12)} \\ R_{j(13)} \\ \vdots \\ R_{j((p-1)p)} \end{bmatrix}_{\substack{y_a \leftarrow x_j \rightarrow y_t \\ (a < t)}} + \sum_{k>j} \begin{bmatrix} R_{jk(12)} + R_{kj(12)} \\ R_{jk(13)} + R_{kj(13)} \\ \vdots \\ R_{jk((p-1)p)} + R_{kj((p-1)p)} \end{bmatrix}_{\substack{y_a \leftarrow \begin{pmatrix} x_j \leftrightarrow x_k \\ x_k \leftrightarrow x_j \end{pmatrix} \rightarrow y_t}} \right) \\
 & = R_{y(j)I} + R_{y(j)II}, j = 1, 2, \dots, m \tag{19}
 \end{aligned}$$

Obviously, the generalized decision coefficient is the sum of the products of  $R_{j(\alpha)}^2$ ,  $R_{jk(\alpha)}$ ,  $R_{j(\alpha t)}$  and  $R_{jk(\alpha t)} + R_{kj(\alpha t)}$  related to  $x_j$  in the division and the corresponding elements in  $R_{yy}^{-1} =$

$(\theta_{at})_{p \times p}$  on the basis of  $R^2 \approx tr(B)$ . In (19),  $R_{y(j)}$  is divided into two parts:  $R_{y(j)I}$  and  $R_{y(j)II}$ .  $R_{y(j)I}$  is the determination part of  $x_j$  and  $x_j \leftrightarrow x_k$  to  $Y_\alpha$ .  $R_{y(j)II}$  is the determination part of  $x_j$  and  $x_j \leftrightarrow x_k$  to  $Y_\alpha$  and  $Y_t (\alpha \neq t)$  due to the common  $X$ . In a word, the generalized decision coefficient includes not only the direct determination of  $x_j$  to  $Y_\alpha$ ,  $Y_\alpha$  and  $Y_t (\alpha \neq t)$ , but also the indirect determination of  $x_j \leftrightarrow x_k (k \neq j)$  to  $Y_\alpha$  and  $Y_\alpha$  and  $Y_t (\alpha \neq t)$ . Specially, the indirect determination considers the correlation among the independent variables and the correlation among the dependent variables at the same time. Therefore, the decision coefficient  $R_{y(j)}$  can be used to express the comprehensive decision ability of  $x_j$  to  $Y$ .

**2.6.2 The hypothesis testing of  $R_{y(j)}$ .** The invalid hypothesis is  $H_0: E(R_{y(j)}) = 0$  and the corresponding t test statistic is:

$$t_j = \frac{R_{y(j)}}{\sqrt{\sum_{\alpha=1}^p (R'_{y(j)\alpha})^2 \frac{c_{jj}(1 - tr(B))}{n - m - 1}}} \sim t(n - m - 1) \tag{20}$$

In (20),  $R'_{y(j)\alpha} = \frac{\partial R_{y(j)}}{\partial b_{j\alpha}^*}$ .

### 3 Application

#### 3.1 Datasets

In order to demonstrate the effectiveness of the multiple-to-multiple path analysis, the wheat data in arid areas to explore breeding rules was selected to discuss. In detail, the wheat data included thirty-five varieties. These data were obtained in a completely randomized block test, and each sample was set with three repetitions [28]. In multiple-to-multiple path analysis, three indexes closely related to wheat yield was selected as dependent variables: panicles per plant ( $y_1$ ), grain number per panicle ( $y_2$ ) and 1000-grain weight ( $y_3$ ), and three other indexes were selected as independent variables: bio-mass per plant ( $x_1$ ), single stem grass weight ( $x_2$ ) and economic coefficient ( $x_3$ ). Here, economic coefficient refers to the ratio of economic yield to biological yield of wheat.

#### 3.2 Calculation and results

Firstly, the phenotypic correlation matrix of the sample was calculated and expressed as Eq (21). The number of observations for each variable is  $n = 105$ .

$$R = \begin{matrix} & x_1 & x_2 & x_3 & y_1 & y_2 & y_3 \\ \begin{matrix} x_1 \\ x_2 \\ x_3 \\ y_1 \\ y_2 \\ y_3 \end{matrix} & \begin{bmatrix} 1 & 0.711 & -0.367 & 0.013 & 0.225 & 0.028 \\ 0.711 & 1 & -0.418 & -0.477 & 0.259 & 0.238 \\ -0.367 & -0.418 & 1 & -0.055 & 0.173 & 0.327 \\ 0.013 & -0.477 & -0.055 & 1 & -0.255 & -0.383 \\ 0.225 & 0.259 & 0.173 & -0.255 & 1 & -0.058 \\ 0.028 & 0.238 & 0.327 & -0.383 & -0.058 & 1 \end{bmatrix} \end{matrix} \tag{21}$$

Then, we establish a multiple-to-multiple standardized multiple linear regression equation

and calculate the corresponding parameters, the results were written as follow:

$$\begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \hat{y}_3 \end{bmatrix} = b^{*T}x = \begin{bmatrix} 0.6767x_1 - 1.0631x_2 - 0.2510x_3 \\ 0.1323x_1 - 0.3121x_2 - 0.3520x_3 \\ -0.2150x_1 - 0.3121x_2 - 0.4986x_3 \end{bmatrix} \tag{22}$$

Right after, four types of hypothesis testing based on the established standardized multiple linear regression model were conducted as follow:

1. The hypothesis testing of generalized complex correlation coefficient  $r_{xy}$ .  
Likelihood ratio statistics of  $X$  and  $Y$  is  $v_{xy} = 0.2987$ , so  $\chi^2 = 121.4357^{**} > \chi^2(3 \times 3)$ , and  $R^2 = 1 - v_{xy} = 0.7013$ ,  $r_{xy} = r_{(x_1, x_2, x_3)(y_1, y_2, y_3)} = \sqrt{R^2} = 0.8374^{**}$ . The results showed that the linear regression of  $Y$  to  $X$  was extremely significant.
2. The hypothesis testing of regression equation  $\hat{y}_\alpha = b_\alpha^{*T}x (\alpha = 1, 2, 3)$ .  
The values of F test statistics are  $F_1 = 37.9188^{**}$ ,  $F_2 = 25.394^{**}$ ,  $F_3 = 14.408^{**}$ , respectively. They were all greater than  $F_{0.01}(3, 101) = 4.007$ , which showed that each standardized regression equation was extremely significant.
3. The hypothesis testing of components  $b_{j\alpha}^*$  in  $b_\alpha^*$ .  
The results of hypothesis testing of components  $b_{j\alpha}^*$  in  $b_\alpha^*$  were listed in [Table 1](#).

Among them, except  $x_1$  was not significant to  $y_2$  and  $y_3$ , the others were extremely significant.

4. The hypothesis testing of  $b_{x_j y}^*$

The results are  $F_1 = 8.670^{**}$ ,  $F_2 = 25.394^{**}$ ,  $F_3 = 10.384^{**}$ , and the test results were all extremely significant.

Except  $x_1$  is not significant to  $y_2$  and  $y_3$ , the above test results showed that the established multiple-to-multiple standardized linear regression equations were extremely significant. The path analysis and decision analysis can be performed subsequently.

Secondly, one-to-multiple path analysis of  $Y_\alpha = \beta_\alpha^{*T}x + \varepsilon_\alpha (\alpha = 1, 2, \dots, p)$  was conducted according to the theory before (Method, Part 2.3). The detailed division results of the correlation coefficient and the determination coefficient were listed in [Table 2](#) and [Table 3](#). The decision analysis was also conducted and the results were also listed in [Table 3](#).

The t test statistics values of decision coefficient hypothesis testing were listed in [Table 4](#).

In one-to-multiple path analysis, the results of correlation coefficient division showed that the total effect of biomass per plant ( $x_1$ ), single stem grass weight ( $x_2$ ) and economic coefficient ( $x_3$ ) are all positive and the largest to panicles per plant ( $y_1$ ), grain number per panicle ( $y_2$ ), 1000-grain weight ( $y_3$ ), respectively. Differently, the direct effect of  $x_1$  to  $y_1$  is the positive and

**Table 1. t test statistics value of  $b_{j\alpha}^*$ .**

	$y_1$	$y_2$	$y_3$
$x_1$	6.8994 <sup>**</sup>	1.0212	-1.8092
$x_2$	-105852 <sup>**</sup>	2.3527 <sup>**</sup>	4.9257 <sup>**</sup>
$x_3$	-3.3060 <sup>**</sup>	3.5101 <sup>**</sup>	5.4202 <sup>**</sup>

<https://doi.org/10.1371/journal.pone.0247722.t001>

**Table 2. The division results of the correlation coefficient about  $Y_\alpha = \beta_\alpha^T x + \varepsilon_\alpha (\alpha = 1, 2, 3)$ .**

	$x_j$ to $y_\alpha$	Direct effect	$x_j \leftrightarrow y_\alpha$	$r_{jk} b_{ka}^*$	Indirect effect	$\sum_{j \neq k} r_{kj} b_{ja}^*$	Total effect
1	$x_1$ to $y_1$	0.6767**	$x_1 \leftrightarrow x_2 \rightarrow y_1$	-0.7559	-0.66638(3)	0.2328(1)	0.013(1)
			$x_1 \leftrightarrow x_3 \rightarrow y_1$	0.0921			
	$x_2$ to $y_1$	-1.0631**	$x_2 \leftrightarrow x_1 \rightarrow y_1$	0.4811	0.5860(1)	-0.3115(3)	-0.477(3)
$x_2 \leftrightarrow x_3 \rightarrow y_1$			0.1049				
	$x_3$ to $y_1$	0.251**	$x_3 \leftrightarrow x_1 \rightarrow y_1$	-0.2483	0.1960(2)	0.1970(2)	-0.055(2)
			$x_3 \leftrightarrow x_2 \rightarrow y_1$	0.4444			
2	$x_1$ to $y_2$	0.1323(3)	$x_1 \leftrightarrow x_2 \rightarrow y_2$	0.2219	0.0927(1)	0.0455(2)	0.225(2)
			$x_1 \leftrightarrow x_3 \rightarrow y_2$	-0.1292			
	$x_2$ to $y_2$	0.3121(2)	$x_2 \leftrightarrow x_1 \rightarrow y_2$	0.0941	-0.0530(2)	0.0914(1)	0.259(1)
$x_2 \leftrightarrow x_3 \rightarrow y_2$			-0.1471				
	$x_3$ to $y_2$	0.3520(1)	$x_3 \leftrightarrow x_1 \rightarrow y_2$	-0.0486	-0.1791(3)	-0.2763(3)	0.173(3)
			$x_3 \leftrightarrow x_2 \rightarrow y_2$	0.1305			
	$x_1$ to $y_3$	-0.2150*(3)	$x_1 \leftrightarrow x_2 \rightarrow y_3$	0.4262	0.2432(1)	-0.074(2)	0.028(3)
$x_1 \leftrightarrow x_3 \rightarrow y_3$			-0.1830				
3	$x_2$ to $y_3$	0.5994**(1)	$x_2 \leftrightarrow x_1 \rightarrow y_3$	-0.1529	-0.3613(3)	0.1757(1)	0.238(2)
			$x_2 \leftrightarrow x_3 \rightarrow y_3$	-0.2084			
	$x_3$ to $y_3$	0.4986**(2)	$x_3 \leftrightarrow x_1 \rightarrow y_3$	0.0789	-0.1716(2)	-0.3914(3)	0.327(1)
			$x_3 \leftrightarrow x_2 \rightarrow y_3$	-0.2505			

<https://doi.org/10.1371/journal.pone.0247722.t002>

**Table 3. The division results of determination coefficient  $Y_\alpha = \beta_\alpha^T x + \varepsilon_\alpha (\alpha = 1, 2, 3)$ .**

	$y_\alpha \leftarrow x_j \rightarrow y_\alpha$	Direct determination	$y_\alpha \leftarrow x_j \leftrightarrow x_k \rightarrow y_\alpha$	$r_{jk} b_{ka}^*$	indirect determination	Decision coefficient
1	$y_1 \leftarrow x_1 \rightarrow y_1$	0.4579	$y_1 \leftarrow x_1 \leftrightarrow x_2 \rightarrow y_1$	-1.0230	-0.8983	-0.4404**(3)
			$y_1 \leftarrow x_1 \leftrightarrow x_3 \rightarrow y_1$	0.1247		
	$y_1 \leftarrow x_2 \rightarrow y_1$	1.1302	$y_1 \leftarrow x_2 \leftrightarrow x_1 \rightarrow y_1$	-1.0230	-1.2461	-0.1159(2)
$y_1 \leftarrow x_2 \leftrightarrow x_3 \rightarrow y_1$			-0.2231			
	$y_1 \leftarrow x_3 \rightarrow y_1$	0.0630	$y_1 \leftarrow x_3 \leftrightarrow x_1 \rightarrow y_1$	0.1247	-0.0984	-0.0354(1)
			$y_1 \leftarrow x_3 \leftrightarrow x_2 \rightarrow y_1$	-0.2231		
2	$y_2 \leftarrow x_1 \rightarrow y_2$	0.0175	$y_2 \leftarrow x_1 \leftrightarrow x_2 \rightarrow y_2$	0.0587	0.0245	0.0420*(2)
			$y_2 \leftarrow x_1 \leftrightarrow x_3 \rightarrow y_2$	-0.0342		
	$y_2 \leftarrow x_2 \rightarrow y_2$	0.0974	$y_2 \leftarrow x_2 \leftrightarrow x_1 \rightarrow y_2$	0.0587	-0.0331	0.0643**(1)
$y_2 \leftarrow x_2 \leftrightarrow x_3 \rightarrow y_2$			-0.0918			
	$y_2 \leftarrow x_3 \rightarrow y_2$	0.1239	$y_2 \leftarrow x_3 \leftrightarrow x_1 \rightarrow y_2$	-0.0342	-0.1260	-0.0021(3)
			$y_2 \leftarrow x_3 \leftrightarrow x_2 \rightarrow y_2$	-0.0918		

(Continued)

Table 3. (Continued)

	$y_{\alpha} \leftarrow x_j \rightarrow y_{\alpha}$	Direct determination	$y_{\alpha} \leftarrow x_j \leftrightarrow x_k \rightarrow y_{\alpha}$	$r_{jk} b_{k\alpha}^*$	indirect determination	Decision coefficient
	$y_3 \leftarrow x_1 \rightarrow y_3$	0.0462	$y_3 \leftarrow x_1 \leftrightarrow x_2 \rightarrow y_3$	-0.1833	-0.1046	-0.0584(2)
	$y_3 \leftarrow x_1 \leftrightarrow x_3 \rightarrow y_3$		0.0787			
3	$y_3 \leftarrow x_2 \rightarrow y_3$	0.3593	$y_3 \leftarrow x_2 \leftrightarrow x_1 \rightarrow y_3$	-0.1833	-0.4331	-0.0738(3)
	$y_3 \leftarrow x_2 \leftrightarrow x_3 \rightarrow y_3$		-0.2498			
	$y_3 \leftarrow x_3 \rightarrow y_3$	0.2486	$y_3 \leftarrow x_3 \leftrightarrow x_1 \rightarrow y_3$	0.0787	-0.1711	0.0775*(1)
	$y_3 \leftarrow x_3 \leftrightarrow x_2 \rightarrow y_3$		-0.2498			

<https://doi.org/10.1371/journal.pone.0247722.t003>

the largest, while the indirect effect is negative and the smallest. The direct effect of  $x_2$  to  $y_2$ ,  $x_3$  to  $y_3$  are not the largest, but the total effect becomes the largest through the correlation regulation by the indirect effect. The results of the determination coefficients division and the decision coefficients showed that for  $y_1$ ,  $x_1$  is a very significant restrictive factor; for  $y_2$ ,  $x_2$  is a very significant positive factor and  $x_1$  is a significant positive factor; for  $y_3$ ,  $x_3$  is a significant positive factor. These results meant that single stem grass weight ( $x_2$ ) and economics coefficient ( $x_3$ ) need to be increased in order to increase grain number per pancicle ( $y_2$ ) and 1000-grainweight ( $y_3$ ), but panicles per plant ( $y_1$ ) will decrease according due to the negative correlation  $x_2$ ,  $x_3$  and  $y_1$ . Meanwhile, biomass per plant ( $x_1$ ) should be decreased in order to increase the panicles per plant ( $y_1$ ), but grain number per pancicle ( $y_2$ ) will decrease here. The contradictory decision-making results of different independent variables ( $x_i$ ) to different dependent variables ( $y_i$ ) often lead to the confusion of breeders.

Therefore, after the one-to-multiple path analysis, the multiple-to-multiple path analysis was practiced by taking into account the correlation between the dependent variables. According to formula (17–19), the generalized determination coefficient  $R^2$  was divided and the results were listed in Table 5.

The specific calculation of path vector structure is as follows:

$$R^2 \approx tr(B) = (1.2883, 1.1030, 1.2086) \begin{bmatrix} 0.5297 \\ 0.1717 \\ 0.2997 \end{bmatrix} + (0.3583, 0.5142, 0.2012) \begin{bmatrix} -0.3331 \\ -0.6323 \\ 0.3861 \end{bmatrix} = 0.8670 \tag{23}$$

From the previous calculation, we can get  $tr(B) = 0.8671$ . The above division of the generalized coefficient of determination is reasonable according to  $R^2 \approx tr(B)$ . The decision analysis of the model was carried out continually. The decision coefficient of each independent variable

Table 4. t test statistics value of  $R_{\alpha(j)}^*$ .

	$y_1$	$y_2$	$y_3$
$x_1$	-3.39774**	2.3205*	-1.2288
$x_2$	-0.74458	4.5633**	-0.7716
$x_3$	-0.9793	-0.0636	2.4488*

<https://doi.org/10.1371/journal.pone.0247722.t004>

Table 5. The division results about three other types paths of the generalized determination coefficients.

a	$y_{\alpha} \leftarrow x_j \rightarrow y_t$	direct determination	$y_{\alpha} \leftarrow x_j \leftrightarrow x_k \rightarrow y_{\alpha}$	$r_{jk} b_{k\alpha}^*$
1	$y_1 \leftarrow x_1 \rightarrow y_2$	0.1791	$y_1 \leftarrow x_1 \leftrightarrow x_2 \rightarrow y_2$	0.3003
			$y_1 \leftarrow x_1 \leftrightarrow x_3 \rightarrow y_2$	-0.1748
			$y_1 \leftarrow x_2 \leftrightarrow x_1 \rightarrow y_2$	-0.2000
1	$y_1 \leftarrow x_2 \rightarrow y_2$	-0.6636	$y_1 \leftarrow x_2 \leftrightarrow x_3 \rightarrow y_2$	0.3128
			$y_1 \leftarrow x_3 \leftrightarrow x_1 \rightarrow y_2$	0.0244
			$y_1 \leftarrow x_3 \leftrightarrow x_2 \rightarrow y_2$	0.0655
1	$y_1 \leftarrow x_3 \rightarrow y_2$	-0.1767	$y_1 \leftarrow x_1 \leftrightarrow x_2 \rightarrow y_3$	0.5768
			$y_1 \leftarrow x_1 \leftrightarrow x_3 \rightarrow y_3$	-0.2477
			$y_1 \leftarrow x_2 \leftrightarrow x_1 \rightarrow y_3$	0.3253
2	$y_1 \leftarrow x_2 \rightarrow y_3$	-1.2744	$y_1 \leftarrow x_2 \leftrightarrow x_3 \rightarrow y_3$	0.4431
			$y_1 \leftarrow x_3 \leftrightarrow x_1 \rightarrow y_3$	-0.0396
			$y_1 \leftarrow x_3 \leftrightarrow x_2 \rightarrow y_3$	0.1258
2	$y_1 \leftarrow x_3 \rightarrow y_3$	-0.2503	$y_2 \leftarrow x_1 \leftrightarrow x_2 \rightarrow y_3$	0.1128
			$y_2 \leftarrow x_1 \leftrightarrow x_3 \rightarrow y_3$	-0.0484
			$y_2 \leftarrow x_2 \leftrightarrow x_1 \rightarrow y_3$	-0.0955
3	$y_2 \leftarrow x_1 \rightarrow y_3$	-0.0569	$y_2 \leftarrow x_2 \leftrightarrow x_3 \rightarrow y_3$	-0.1301
			$y_2 \leftarrow x_3 \leftrightarrow x_1 \rightarrow y_3$	0.0556
			$y_2 \leftarrow x_3 \leftrightarrow x_2 \rightarrow y_3$	-0.1764

<https://doi.org/10.1371/journal.pone.0247722.t005>

to  $Y = (y_1, y_2, y_3)^T$  was calculated as follows:

$$\begin{aligned}
 R_{y(1)} &= (\theta_{11}, \theta_{22}, \theta_{33}) \times \left( \begin{array}{c} \left[ \begin{array}{c} R_{1(1)}^2 \\ R_{1(2)}^2 \\ R_{1(3)}^2 \end{array} \right]_{\substack{y \\ 3 \times 1} \leftarrow \substack{x_1 \rightarrow y \\ 3 \times 1}} + \left[ \begin{array}{c} R_{12(1)} \\ R_{12(2)} \\ R_{12(3)} \end{array} \right]_{\substack{y \\ 3 \times 1} \leftarrow \substack{x_1 \leftrightarrow x_2 \rightarrow y \\ 3 \times 1}} + \left[ \begin{array}{c} R_{13(1)} \\ R_{13(2)} \\ R_{13(3)} \end{array} \right]_{\substack{y \\ 3 \times 1} \leftarrow \substack{x_1 \leftrightarrow x_3 \rightarrow y \\ 3 \times 1}} \end{array} \right) \\
 &+ (\theta_{12}, \theta_{13}, \theta_{23}) \left( \begin{array}{c} \left[ \begin{array}{c} R_{1(12)} \\ R_{1(13)} \\ R_{1(23)} \end{array} \right]_{\substack{y_2 \leftarrow x_1 \rightarrow y_t \\ (x < t)}} + \left[ \begin{array}{c} R_{12(12)} + R_{21(12)} \\ R_{12(13)} + R_{21(13)} \\ R_{12(23)} + R_{21(23)} \end{array} \right]_{\substack{y_2 \leftarrow \begin{pmatrix} x_1 \leftrightarrow x_2 \\ x_2 \leftrightarrow x_1 \end{pmatrix} \rightarrow y_t}} + \left[ \begin{array}{c} R_{13(12)} + R_{31(12)} \\ R_{13(13)} + R_{31(13)} \\ R_{13(23)} + R_{31(23)} \end{array} \right]_{\substack{y_2 \leftarrow \begin{pmatrix} x_1 \leftrightarrow x_3 \\ x_3 \leftrightarrow x_1 \end{pmatrix} \rightarrow y_t}} \end{array} \right) \\
 &= -0.5916 + 0.2060 = -0.3856^{**} \tag{24}
 \end{aligned}$$

Similar available:  $R_{y(2)} = -0.1157$ ,  $R_{y(3)} = 0.0906^*$ . According to the decision coefficient, the t test about  $R_{y(j)}$  is further conducted, and the result is  $t_1 = -4.3943^{**}$ ,  $t_2 = 0.9293$ ,  $t_3 = 2.0785^{**}$ . In addition, it should be noted that the determination coefficients of  $x_j$  and  $x_j \leftrightarrow x_k$  to

$y_\alpha$  have been calculated by one-to-multiple path analysis model (Table 3). The comparison of the results of Table 3 and those of Table 5 demonstrated that great changes have taken place in the regulation of  $x_j$  to  $Y$  when the correlation among dependent variables was considered. Firstly, the direct and indirect regulations of  $x_j$ ,  $x_j \leftrightarrow x_k$  to  $Y$  also were greatly affected by the correlation among  $Y$  because of common  $X$ . As shown in Table 3, the direct determination of  $x_2$  to  $y_1$ ,  $y_2$  were both positive, respectively ( $R_{1(2)}^2 = 1.1302 y_1 \leftarrow x_2 \rightarrow y_1$ ;  $R_{2(2)}^2 = 0.0974 y_2 \leftarrow x_2 \rightarrow y_2$ ). But in Table 5, the direct determination of  $x_2$  to  $y_1$  and  $y_2$  became negative ( $R_{1(2)}^2 = -0.6636 y_1 \leftarrow x_2 \rightarrow y_2$ ). This change was due to the consideration of the negative and large correlation of  $y_1$  and  $y_2$  ( $r_{y_1 y_2} = -0.255$ ). Similarly, the direct determination of  $x_2$  to  $y_2$  and  $y_3$  was still changed ( $R_{2(2)}^2 = 0.3741 y_2 \leftarrow x_2 \rightarrow y_3$ ), compared to the previous determination coefficient ( $R_{2(2)}^2 = 0.0974 y_2 \leftarrow x_2 \rightarrow y_2$ ;  $R_{3(2)}^2 = 0.3593 y_3 \leftarrow x_2 \rightarrow y_3$ ). Different from the above, this change was small and both were positive. This phenomenon showed that the small correlation of  $y_2$  and  $y_3$  ( $r_{y_2 y_3} = -0.058$ ) had little influence on the direct determination of  $x_2$  to  $y_2$  and  $y_3$ . The direct determination of  $x_3$  to  $y_2$  and  $y_3$  was exactly like the direct determination of  $x_2$  to  $y_2$  and  $y_3$ . The indirect determination due to the correlation of  $x_j \leftrightarrow x_k$  also changed a lot because of consideration of the correlation among  $Y$ . For example, the indirect determination of  $x_1 \leftrightarrow x_2$  to  $y_1$  and  $y_3$  was  $0.5768(y_1 \leftarrow x_1 \leftrightarrow x_2 \rightarrow y_3)$  and  $0.3253(y_1 \leftarrow x_2 \leftrightarrow x_1 \rightarrow y_3)$ . It's strange that the original indirect determination of  $x_1 \leftrightarrow x_2$  to  $y_1$ ,  $x_1 \leftrightarrow x_2$  to  $y_3$  were  $-1.023(y_1 \leftarrow x_1 \leftrightarrow x_2 \rightarrow y_1)$  and  $0.1833(y_3 \leftarrow x_1 \leftrightarrow x_2 \rightarrow y_3)$ , respectively. It is obvious that the strong negative correlation of  $y_1$  and  $y_3$  ( $r_{y_1 y_3} = -0.383$ ) led to the change of indirect regulation. These big changes were enough to show the importance of considering the correlation among  $Y$ . There were similar changes in the direct determination of  $x_1 \leftrightarrow x_2$  to  $y_1$  and  $y_2$  ( $y_1 \leftarrow x_1 \leftrightarrow x_2 \rightarrow y_2$ ,  $y_1 \leftarrow x_2 \leftrightarrow x_1 \rightarrow y_2$ ) and  $x_2 \leftrightarrow x_3$  to  $y_1$  and  $y_3$  ( $y_1 \leftarrow x_2 \leftrightarrow x_3 \rightarrow y_3$ ,  $y_1 \leftarrow x_3 \leftrightarrow x_2 \rightarrow y_3$ ). Secondly, the decision coefficients results showed that  $x_1$  is the very significant restrictive decision factor of  $Y = (y_1, y_2, y_3)^T$  ( $R_{y(1)} = -0.3856^{**}$ ). But  $x_1$  is the significant positive decision factor to  $y_2$  ( $R_{y_2(1)} = 0.0420^*$ ) and is not significant to  $y_3$  ( $R_{y_3(1)} = -0.0584$ ). This phenomena seemed to be caused by the very significant negative decision making effect of  $x_1$  to  $y_1$ , and strong negative correlation between  $y_1$  and  $y_2$  ( $r_{y_1 y_2} = -0.255$ ),  $y_1$  and  $y_3$  ( $r_{y_1 y_3} = -0.383$ ). For  $x_2$ , there is no point in making a decision. ( $R_{y(2)} = -0.1157$ ). And  $x_3$  became a significant positive decision factor to  $Y = (y_1, y_2, y_3)^T$  ( $R_{y(3)} = 0.0906^*$ ). However,  $x_3$  is significant only to  $y_3$  ( $R_{y_3(3)} = 0.0775^*$ ), and is not significant to  $y_1$ ,  $y_2$  in one-to-multiple path analysis. Obviously, the correlation among  $Y$  due to common  $X$  makes a big difference in the decision making. The results showed that the economic coefficient ( $x_3$ ) should be increased, the biomass per plant ( $x_1$ ) should be appropriately reduced and the single stem grass weight ( $x_2$ ) should remain unchanged in the process of wheat breeding. These results were in accordance with the existing documents results [28]. In short, the consideration of the correlation among  $Y$  caused a big change of the direct determination, the indirect determination and the decision analysis results of  $x_j$  to  $Y$ . And the greater the correlation among  $Y$  is, the greater the impact on regulation.

## 4 Discussion

In this article, the multiple-to-multiple path analysis model was proposed based on multivariate linear regression analysis, which can be regarded as a generalization of one-to-multiple path analysis model based on univariate linear regression analysis. The innovation of this model is the multiple-to-multiple path analysis central theorem. The correlation among  $Y$  caused by common  $X$  was considered in the system analysis including multiple independent variables and multiple dependent variables. As Fig 2 shown, the other three types of paths ( $y_\alpha \leftarrow x_j \rightarrow y_\beta$ ,  $y_\alpha \leftarrow x_j \leftrightarrow x_k \rightarrow y_\beta$ ,  $y_\alpha \leftarrow x_k \leftrightarrow x_j \rightarrow y_\beta$ ) generated in multiple-to-multiple path analysis

model besides the two types of paths ( $y_{\alpha} \leftarrow x_j \rightarrow y_{\alpha}$ ,  $y_{\alpha} \leftarrow x_j \leftrightarrow x_k \rightarrow y_{\alpha}$ ) in one-to-multiple path analysis. Along these five types of paths, the generalized determination coefficient  $R^2$  was divided into the direct determination and the indirect determination according to  $R^2 \approx tr(B)$ . This division can clearly show the complex regulatory mechanisms among variables. Still further, the generalized decision coefficient  $R_{y(j)}$  was constructed by synthesizing all the items related to  $x_j$ , which was used to express the comprehensive decision-making ability of  $x_j$  to  $Y = (Y_1, Y_2, \dots, Y_p)^T$ . In fact, the direct and indirect determinations all were products of corresponding path coefficients. The quantitative expression of the regulation among variables is helpful for decision makers to make more reasonable and optimized decision suggestions for target variables. The analysis results of the wheat data in arid areas strongly confirm this. It is worth mentioning that the path analysis of any closed system can be made according to the multiple-to-multiple path analysis central theorem. However, the application of multiple-to-multiple path analysis model still has some limitations. Firstly, the model is only applicable to the causal relationship analysis among multiple dependent variables and independent variables with correlation. Secondly, the difference between the generalized determination  $R^2$  and  $tr(B)$  is relatively large when the correlation among variables is very strong in multiple-to-multiple linear regression analysis, that is, the value of the correlation coefficient in correlation matrix is almost 1. Here, the division of the generalized determination coefficient  $R^2$  based on  $R^2 \approx tr(B)$  is very different from the actual result. Therefore, other division methods need to be further considered.

## 5 Conclusion

In the multiple-to-multiple path analysis model, the correlation among dependent variables caused by common independent variable is considered, besides the correlation among independent variables. Taking into account more correlation information analysis makes the results more practical and instructive.

## Author Contributions

**Conceptualization:** Junli Du, Zhifa Yuan.

**Funding acquisition:** Junli Du.

**Methodology:** Junli Du, Zhifa Yuan.

**Project administration:** Junli Du, Zhifa Yuan.

**Resources:** Junli Du.

**Software:** Yujie Du, Xi Liu.

**Supervision:** Junli Du, Zhifa Yuan.

**Writing – original draft:** Yujie Du, Xi Liu.

**Writing – review & editing:** Junli Du, Zhifa Yuan.

## References

1. Naes T, Romano R, Tomic O, et al. Sequential and orthogonalized PLS (SO-PLS) regression for path analysis: Order of blocks and relations between effects. *J Chemom.* 2020; e3243.
2. Wright S. On the nature of size factors. *Genetics.* 1918; 3(4): 367–374. PMID: [17245910](https://pubmed.ncbi.nlm.nih.gov/17245910/)
3. Wright S. Correlation and causation. *J Agric Res.* 1921; 20: 557–585.
4. Bollen KA. *Structural Equations with Latent Variables.* NY: Wiley. 1989.
5. Duncan OD. Path analysis: sociological examples. *Am J Sociol.* 1966; 72(1): 1–16.



6. Finney JM. Indirect effects in path analysis. *Socio Meth Res.* 1972; 1(2): 175–186.
7. Greene VL. An algorithm for total and indirect causal effects. *Polit Anal.* 1977; 369–381.
8. Berg PVD, Arentze T, Timmermans H. A path analysis of social networks, telecommunication and social activity-travel patterns. *Transp Res Part C. Emerg Technol.* 2013; 26: 256–268.
9. Kang DH. An path analysis of the elderly's deprivation experience on the thinking of suicide. *J Soc Sci.* 2019; 58: 197–245.
10. Hwang HJ, Chun HY, Ok KH. The path analysis of parental divorce on children's emotional and behavioural problems: Through child-rearing behaviours and children's self-esteem. *J Korean Home Econ Assoc.* 2010; 48(7): 99–110.
11. Diao ZJ, Chen B. Correlation and path coefficient analysis between thermal extraction yield and coal properties. *Energy Sources Part A. Recovery Util Environ Eff.* 2016; 38(22): 3412–3416.
12. Cankaya S, Abaci SH. Path analysis for determination of relationships between some body measurements and live weight of German fawn x hair crossbred kids. *Kafkas Univ Vet Fak Derg.* 2012; 18 (5): 769–773. <https://doi.org/10.9775/kvfd.2012.6376>
13. Norris D, Brown D, Moela AK, et al. Path coefficient and path analysis of body weight and biometric traits in indigenous goats. *Indian J Anim Res.* 2015; 49 (5): 573–578.
14. Marjanović-Jeromela A, Marinković R, Mijić A, et al. Correlation and path analysis of quantitative traits in winter rapeseed (*brassica napus* L.). *Agric Conspec Sci.* 2008; 73: 13–18.
15. Barbosa RP, Alcantara-Neto F, Gravina LM, et al. Early selection of sugarcane using path analysis. *Genet Mol Res.* 2017; 16(1): gmr16019038. <https://doi.org/10.4238/gmr16019038> PMID: 28198498
16. Grace JB, Pugesek BH. On the use of path analysis and related procedures for the investigation of ecological problems. *Am Nat.* 1998; 152 (1):151–159. <https://doi.org/10.1086/286156> PMID: 18811408
17. Kunanithaworn N, Wongpakaran T, Wongpakaran N, et al. Factors associated with motivation in medical education: a path analysis. *BMC Med Educ.* 2018; 18: 140. <https://doi.org/10.1186/s12909-018-1256-5> PMID: 29914462
18. Costello RM. Premorbid social competence construct generalizability across ethnic groups: Path analyses with two premorbid social competence components. *J Consult Clin Psychol.* 1978; 46(5): 1164–1165. <https://doi.org/10.1037//0022-006x.46.5.1164> PMID: 701557
19. Jöreskog KG. Structural analysis of covariance and correlation matrices. *Psychometrika.* 1978; 43(4): 443–477.
20. Graff J, Schmidt P. A general model for decomposition of effects. North–Holl Publ Co. 1982; 131–148. Netherlands.
21. Yuan ZF, Zhou JY, Guo MC, et al. Decision coefficients-decision indicators in path analysis. *J Northwest A&F Univ (Nat Sci Ed).* 2001; 29(5): 131–133. China.
22. Xie XL, Yuan ZF. Statistical test of decision coefficient and its application in breeding. *J Northwest A&F Univ (Nat Sci Ed).* 2013; 41(3): 111–114. China.
23. Mei Y, Guo W, Fan S, et al. Analysis of decision-making coefficients of the lint yield of upland cotton (*Gossypium hirsutum* L.). *Euphytica.* 2014; 196(1): 95–104.
24. Du JL, Li ML, Yuan ZF, et al. A decision analysis model for KEGG pathway analysis. *BMC Bioinform.* 2016; 17(1): 407. <https://doi.org/10.1186/s12859-016-1285-1> PMID: 27716040
25. Du JL, Yuan ZF, Ma ZW, et al. KEGG-PATH: Kyoto encyclopedia of genes and genomes-based pathway analysis using a path analysis model. *Mol Biosyst.* 2014; 10(9): 2441–2447. <https://doi.org/10.1039/c4mb00287c> PMID: 24994036
26. Xie XL, Du JL, Xie XZ, et al. Generalized complex correlation coefficient and its application in wheat breeding. *J Triticeae Crop.* 2017; 37(1): 87–93. China.
27. Duleba AJ, Olive DL. Regression analysis and multivariate analysis. *Semin Reprod Endocrinol.* 1996; 14(2): 139–153. <https://doi.org/10.1055/s-2007-1016322> PMID: 8796937
28. Zhang Z, Wang D. Wheat drought-resistant ecological breeding. Xi'an: Shaanxi People's Education Press; 1992. China.