Which Variable Should Be Dependent in Phylogenetic Generalized Least Squares Regression Analysis

Zheng-Lin Chen[1,2], Hong-Ji Guo[1,2], and Deng-Ke Niu[1,2,*]

[1]*MOE Key Laboratory for Biodiversity Science and Ecological Engineering, College of Life Sciences, Beijing Normal University, Beijing, 100875, China*

[2]*Beijing Key Laboratory of Gene Resource and Molecular Development, College of Life Sciences, Beijing Normal University, Beijing, 100875, China*

[*]*Correspondence to be sent to: College of Life Sciences, Beijing Normal University, Beijing, 100875, China; E-mail: dkniu@bnu.edu.cn, dengkeniu@hotmail.com.*

ABSTRACT

Phylogenetic generalized least squares (PGLS) regression is one of the most commonly used methods in examining evolutionary correlations between two traits. Unlike the conventional correlation methods like Pearson and Spearman's rank tests, the two analyzed traits are in different positions when correcting the phylogenetic non-independence in PGLS regression. In examining the correlations of CRISPR-Cas and prophage contents with optimal growth temperature and minimal doubling time, we noticed that a remarkable frequency (26.3%) of conflicting results appears after swapping the independent and dependent variables. Then, we generated 12000 simulations of the evolution of two traits (X_1 and X_2) along a binary tree containing 100 terminal nodes with different models and variances. In this simulated dataset, swapping the dependent and independent variables gave conflicting results at a frequency of 17.2%. By conventional correlation analysis of the trait changes along the phylogenetic branches (ΔX_1 and ΔX_2), we established a golden standard for whether X_1 and X_2 correlate in each simulation. With this golden standard, we compared six potential criteria for dependent variable selection, log-likelihood, Akaike information criterion, R^2, p-value, Pagel's λ, and the estimated λ in Pagel's λ model. The last two criteria were found to be equivalent in their performance of dependent variable selection and superior to the other four criteria. Because Pagel's λ values, as indicators of phylogenetic signals, are generally calculated at the beginning of phylogenetic comparative studies, for practical convenience, we recommend the trait with a higher λ value to be used as the dependent variable in future PGLS regressions. Logical analysis of cause and effect should be done after establishing a significant correlation by PGLS regression rather than providing an indicator for the choice of the dependent variable.

Keywords

Evolutionary correlation; simulation; phylogenetic signal; phylogenetic generalized least squares (PGLS); regression; dependent variable; Akaike information criterion (AIC).

Due to mutations, genetic drift, or natural selection, some biological traits tend to evolve together over evolutionary time (Felsenstein 1985, Revell and Collar 2009, Goswami et al. 2014, Caetano and Harmon 2019, Revell et al. 2022). Correlation analysis between traits could quantify the magnitude and direction of changes in one trait given knowledge of evolutionary changes in another, providing evidence or counterexamples for hypotheses (Pearman et al. 2014), prompting deep thinking about biological processes, and contributing to better understanding and reconstruction of the events that occurred during evolutionary history (Bartomeus et al. 2018, Bawa et al. 2019, Suarez-Castro et al. 2020).

The correlation between two variables is usually assessed by computing the Pearson, Spearman's rank, or Kendall rank correlation coefficient. However, the values of evolutionary traits generally violate a basic assumption of these standard statistical methods. They are not independent of each other but related through the evolutionary relationships of the analyzed species (Felsenstein 1985). Ignoring the phylogenetic dependences would distort the trait correlation (Revell 2010, Whitney and Garland 2010). A series of methods have been developed to address the problem of phylogenetic non-independence (Felsenstein 1985, Grafen 1989, Lynch 1991, Garamszegi 2014). Among them, the phylogenetic generalized least squares (PGLS), initially formulated by Grafen (1989) and subsequently developed by Martins and Hansen (1997), Pagel (1997, 1999), and Rohlf (2001), became the most commonly employed method to determine the relationship between two or more traits. When interpreting the correlation between two traits using PGLS regression, the regressions' significant positive and negative slopes correspond to significant positive and negative correlations, respectively.

The two traits are at the same position in standard correlation analyses like Pearson and Spearman's rank tests. However, in a regression model, one trait is designated as the independent variable and the other as the dependent variable. When exploring trait

correlations using regression models, there is a common misconception that the positions of the independent and dependent variables are not equivalent. A causal relationship is often assumed in the regression analysis (Waugh 1943). For example, when analyzing the relationship between GC content and ecological factors, we tend to choose the ecological factors as the independent variable and the GC content as the dependent variable (Travnicek et al. 2019, Hu et al. 2022), with the assumption that the environmental factors might have driven the evolution of GC content. However, from the statistical principle, the regression analysis examines whether there is a significant positive or negative correlation between these two traits. In standard regression analysis, the independent and dependent variables are exchangeable; exchanging the independent and dependent variables will not affect the sign (positive or negative) of the regression coefficient or the significance test results ($p <$ or ≥ 0.05). Therefore, we could arbitrarily select one trait as the independent variable and designate the other as the dependent variable in standard regression analysis. However, at this point, the PGLS regression analysis is different from the standard regression analysis. Swapping the independent and dependent variables in the PGLS analysis will result in inconsistent estimates of the parameters, like Pagel's λ (Freckleton et al. 2002), and so inconsistent phylogenetic correction for the phylogenetic non-independence of the trait values. In principle, this inconsistency might lead to conflicting results.

When scrutinizing the correlations in a recent publication of our laboratory (Liu et al. 2023), we noticed that, in some cases, swapping the independent and dependent variables affects the results and conclusions of the correlation analysis qualitatively (see the Results section). That is, a significant correlation between two traits ($p < 0.05$) might disappear ($p \geq 0.05$) after swapping the independent and dependent variables. How do we know, in these cases, which regression should we accept?

In the present study, using simulated data, we first evaluated the prevalence of conflicting

results resulting from swapping the independent and dependent variables, then compared potential criteria to designate the independent and dependent variables in PGLS analysis appropriately

MATERIALS AND METHODS

Phylogenetic Generalized Least Squares

Phylogenetic generalized least squares, as proposed by Martins and Hansen (1997), rely on the generalized least squares model and can be written as

$$\boldsymbol{y} = \boldsymbol{X\alpha} + \boldsymbol{\varepsilon},\ \boldsymbol{\varepsilon} \sim MVN(0, \sigma^2\boldsymbol{\Sigma}), \tag{1}$$

where $\boldsymbol{y}$ is an n-dimensional vector of values of trait y, considered as a response variable in the regression model, n is the sample size. $\boldsymbol{X}$ is an $n \times (1+p)$ matrix consisting of a column of ones and p columns of explanatory variables; the first column of ones can be interpreted as the intercept. $\boldsymbol{\alpha}$ is a column vector of regression coefficients. $\boldsymbol{\varepsilon}$ is a column vector of errors, under the multivariate normal distribution, with a mean of 0 and a variance-covariance matrix of $\sigma^2\boldsymbol{\Sigma}$. $\boldsymbol{\Sigma}$ is a matrix describing the phylogenetic relationships (topology and branch length). Therefore $\boldsymbol{\varepsilon}$ will have a multinormal probability density given by

$$p(\boldsymbol{\varepsilon}) = \frac{1}{(2\pi\sigma^2)^{n/2}} |\boldsymbol{\Sigma}|^{-1/2} \times exp\left[-\frac{1}{2\sigma^2}\boldsymbol{\varepsilon}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\varepsilon}\right]. \tag{2}$$

According to equation (1), equation (2) can be expressed by

$$p(\boldsymbol{\varepsilon}) = \frac{1}{(2\pi\sigma^2)^{n/2}} |\boldsymbol{\Sigma}|^{-1/2} \times exp\left[-\frac{1}{2\sigma^2}(\boldsymbol{y} - \boldsymbol{X\alpha})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{y} - \boldsymbol{X\alpha})\right], \tag{3}$$

and the log-likelihood function can be given by

$$L(\widehat{\boldsymbol{\alpha}}, \hat{\sigma}^2) = -\frac{(\boldsymbol{y}-\boldsymbol{X}\widehat{\boldsymbol{\alpha}})^T(\hat{\sigma}^2\boldsymbol{\Sigma})^{-1}(\boldsymbol{y}-\boldsymbol{X}\widehat{\boldsymbol{\alpha}})}{2} - \frac{\log(|\hat{\sigma}^2\boldsymbol{\Sigma}|)}{2} - \frac{n\log(2\pi)}{2}, \tag{4}$$

The parameters $\boldsymbol{\alpha}$ and σ^2 can be estimated by maximizing the value of $L(\widehat{\boldsymbol{\alpha}}, \hat{\sigma}^2)$ via a maximum likelihood approach.

All the PGLS regressions in this study were performed using the R (version 4.0.2)

package phylolm (version 2.6.2) (Ho and Ane 2014).

Evolutionary Models

The Brownian motion (BM) model on a phylogeny is like a "random walk" model, in which the trait value changes with a constant probability σ^2 in any unit of time. In this model, $\boldsymbol{\Sigma}_{ij} = \boldsymbol{C}_{ij}$, where $\boldsymbol{C}_{ij}$ is the distance from the root node to the most recent common ancestor of tip i and j. If $i = j$, $\boldsymbol{C}_{ij}$ will be the distance from the root node to the tip i. Pagel (1997) introduced the parameter λ to represent different evolution rates on branches and to do length transformations. In Pagel's λ model, the off-diagonal elements of the variance-covariance matrix are multiplied by λ. We noted the variance-covariance matrix modified by λ as $\boldsymbol{\Sigma}(\lambda)$. The multinormal probability density can be written as:

$$p(\boldsymbol{\varepsilon}) = \frac{1}{(2\pi\sigma^2)^{n/2}} |\boldsymbol{\Sigma}(\lambda)|^{-1/2} \times exp\left[-\frac{1}{2\sigma^2}(\boldsymbol{y} - \boldsymbol{X\alpha})^T \boldsymbol{\Sigma}(\lambda)^{-1}(\boldsymbol{y} - \boldsymbol{X\alpha})\right], \quad (5)$$

and the log-likelihood function can be given by

$$L(\lambda, \hat{\boldsymbol{\alpha}}, \hat{\sigma}^2) = -\frac{(\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\alpha}})^T(\hat{\sigma}^2\boldsymbol{\Sigma}(\lambda))^{-1}(\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\alpha}})}{2} - \frac{\log(|\hat{\sigma}^2\boldsymbol{\Sigma}(\lambda)|)}{2} - \frac{n\log(2\pi)}{2}. \quad (6)$$

The parameter λ, lying between 0 and 1, is estimated by a search procedure to maximize the likelihood of equation (6) (Freckleton et al. 2002). The higher the λ value is, the stronger phylogenetic dependence is. If $\lambda = 0$, there is no phylogenetic dependence between the residuals. If $\lambda = 1$, $\boldsymbol{\Sigma}(\lambda) = \boldsymbol{C}$, Pagel's λ model and Brownian motion model are equivalent.

The phylogenetic signals (λ) were estimated using the R (version 4.0.2) package phytools (version 1.0-3) (Revell 2012).

Empirical Datasets

A dataset including the empirical minimal doubling times, the CRISPR spacer numbers, the optimal growth temperature, and the number of prophages of 262 bacteria was

extracted from the Supplemental Material Table S1 of Liu et al. (2023). The phylogenetic tree of these 262 bacteria was retrieved from the Genome Taxonomy Database (GTDB; accessed 8 April 2022) (Parks et al. 2022).

Simulation Data

First, we generated a binary tree containing 100 terminal nodes using the package ape in R (Paradis and Schliep 2019). The trait X_1 evolved under a Brownian motion model (using the packages ape in R) with a variance rate ${\sigma^2}_{BM} = 4$ along the tree, and then the trait X_2 was simulated based on

$$X_2 = X_1 + \varepsilon, \tag{7}$$

where ε was a normally distributed random noise term with a mean of 0 and a variance varying from 1, 4, 16, 64, 256, to 1024. The term ε introduced noise to the dependent variable X_2 and a gradient variance of this noise term (from 1 to 1024) changed the correlation between X_1 and X_2, from strong to weak. We named this case "BM & BM + Norm."

Then, we simulated the trait X_1 under a normal distribution with a mean of 0 and a variance of 4. The trait X_2 was simulated based on equation (7), where ε is simulated under a Brownian motion model with ${\sigma^2}_{BM}$ varying from 1, 4, 16, 64, 256, to 1024. We named this case "Norm & Norm + BM."

Each parameter condition was simulated 1000 times, and totally we performed 12000 simulations.

RESULTS

Swapping the Dependent and Independent Variables in PGLS Analyses of Empirical Data Sometimes Gave Conflicting Results

In a recent publication of our laboratory (Liu et al. 2023), a dataset containing a series of traits of 262 bacteria was deposited as Supplemental Material Table S1. We scrutinized the correlations of empirical minimal doubling time and optimal growth temperature with the genomic characters. In total, 38 pairs of traits were re-analyzed by PGLS using Pagel's λ model. Swapping the dependent and independent variables in ten pairs gave conflicting results and led to different conclusions (Table 1). For example, when choosing the average prophage number as the dependent variable and the optimal growth temperature as the independent variable, the PGLS analysis showed a significant negative correlation between the two traits ($p = 9 \times 10^{-4}$, Table 1). However, no statistically significant correlation was observed when the dependent and independent variables were swapped ($p = 0.242$, Table 1). Between each pair of conflicting results, we have to find the correct one and make the conclusion based on it. However, how do we know which is correct?

Table 1. Swapping the dependent and independent variables in PGLS might give conflicting results[a]

Dependent Variable ~ Independent Variable	Slope	p	R^2
EMDT ~ min_CRISPR_spacer	0.010	0.049	0.015
min_CRISPR_spacer ~ EMDT	1.376	0.053	0.014
EMDT ~ mean_prophage_number	0.565	0.013	0.024
mean_prophage_number ~ EMDT	0.008	0.558	0.001
EMDT ~ min_prophage_number	0.622	0.007	0.028
min_prophage_number ~ EMDT	0.011	0.427	0.002
EMDT ~ max_prophage_number	0.448	0.037	0.017
max_prophage_number ~ EMDT	0.005	0.722	5×10^{-4}

OPT ~ min_CRISPR_array	0.308	0.061	0.014
min_CRISPR_array ~ OPT	0.053	8×10^{-5}	0.058
OPT ~ mean_prophage_number	–0.197	0.242	0.005
mean_prophage_number ~ OPT	–0.031	9×10^{-4}	0.042
OPT ~ min_prophage_number	–0.154	0.374	0.003
min_prophage_number ~ OPT	–0.026	0.004	0.031
OPT ~ max_prophage_number	–0.183	0.214	0.006
max_prophage_number ~ OPT	–0.036	4×10^{-4}	0.047
OPT ~ prophage_existence	–1.046	0.182	0.007
prophage_existence ~ OPT	–0.009	8×10^{-4}	0.043
OPT ~ Average_Cas_gene_number	0.197	0.610	0.002
Average_Cas_gene_number ~ OPT	0.023	0.012	0.043

[a]Pagel's λ model was used in these analyses. The dataset containing 262 bacteria was extracted from the Supplemental Material Table S1 of Liu et al. (2023), and the phylogenetic tree used in PGLS analysis was retrieved from the Genome Taxonomy Database (Parks et al. 2022). EMDT, empirical_minimal_doubling_times; OPT, mean_optimal_growth_temperature.

Prevalence of Conflicting Resulting from Swapping the Dependent and Independent Variables

In the above survey of correlations in the empirical data, we found that, in 26.3% of cases, swapping the dependent and independent variables would lead to conflicting results. It seems not a rare case. To assess the prevalence of conflicting results caused by swapping the dependent and independent variables, we simulated the evolution of two traits along a binary tree containing 100 terminal nodes. Different distributions of the features were simulated. In the case of "BM & BM + Norm", we constructed trait X_1 under the Brownian motion model

("BM") and trait X_2, as "BM + Norm", equals to X_1 plus a noise term "Norm." In the case of "Norm & Norm + BM," the noise term is "BM." To account for varying levels of correlation, we set the gradient variance of the noise term, with the variance ${\sigma^2}_{Norm}$ varying from 1, 4, 16, 64, 256, to 1024. For each case, we simulated 1000 times.

For the data of each simulation, we performed two rounds of PGLS analysis, $X_1 \sim X_2$ and $X_2 \sim X_1$. The results are deposited in Supplementary Table S1 and summarized in Table 2. In 3768 simulations, neither $X_1 \sim X_2$ nor $X_2 \sim X_1$.gave significant correlations ($p \geq 0.05$ for all of them). In 6150 simulations, both $X_1 \sim X_2$ and $X_2 \sim X_1$.gave significant correlations ($p < 0.05$). In each of these 6150 cases, the regression coefficients of $X_1 \sim X_2$ and $X_2 \sim X_1$ have the same sign. In the other 2064 simulations, only one of $X_1 \sim X_2$ and $X_2 \sim X_1$.gave significant correlations ($p < 0.05$ for one and $p \geq 0.05$ for the other). That is, the frequency of conflicting results caused by swapping the dependent and independent variables in our simulated dataset is 17.2%.

From Table 2, we can see a relationship between the variance in the simulations and the frequency of conflicting results caused by swapping the dependent and independent variables. When the variance of the noise term is slight (like 1 and 4), *i.e.*, there is a strong correlation between X_1 and X_2, swapping the independent and dependent variables gives almost the same results. As the variance of the noise term increases, i.e., when the correlation between X_1 and X_2 becomes weak, swapping the independent and dependent variables leads to many conflicting results. In cases where the variance equals 16 in "BM~BM+Norm" and 64 in "Norm~Norm+BM," there are even close to 50% of cases with conflicting results. However, as the variance becomes much more prominent and the correlation between X_1 and X_2 becomes weaker, the frequency of conflicting results caused by swapping the dependent and independent variables diminishes (Table 2).

Table 2. Frequency of conflicting results caused by swapping the dependent and independent variables in PGLS analyses

X_1[a]	BM						Norm					
X_2[b]	BM+Norm						Norm+BM					
Noise Term Variance	1	4	16	64	256	1024	1	4	16	64	256	1024
Simulation times	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000
Times of $X_1 \sim X_2$ giving $p < 0.05$	1000	982	538	187	68	53	1000	1000	925	386	133	74
Times of $X_2 \sim X_1$ giving $p < 0.05$	1000	1000	954	525	170	80	1000	1000	1000	844	317	128
Both $X_1 \sim X_2$ and $X_2 \sim X_1$ giving $p < 0.05$	1000	982	529	166	34	14	1000	1000	925	374	87	39
Both $X_1 \sim X_2$ and $X_2 \sim X_1$ giving $p \geq 0.05$	0	0	37	454	796	881	0	0	0	144	637	837
Times of conflicting results between $X_1 \sim X_2$ and $X_2 \sim X_1$[c]	0	18	434	380	170	105	0	0	75	482	276	124

[a]In 6000 simulations, the trait X_1 evolved under a Brownian (BM) motion model, and in the other 6000 simulations, the trait X_1 evolved under a normal distribution. For details, please see the MATERIALS AND METHODS section.

[b]The trait X_2 was simulated by adding a noise term to X_1. The noise term was under a normal distribution in the first 6000 simulations and under a Brownian (BM) motion model in the other 6000 simulations. For details, please see the MATERIALS AND METHODS section.

[c]When one of $X_1 \sim X_2$ and $X_2 \sim X_1$ gives a $p < 0.05$ result, and the other gives a $p \geq 0.05$ result, we define these results as conflicting results.

Establish a Golden Standards to Evaluate the Correlations Between Simulated Traits

In the PGLS analysis of both the empirical and the simulated datasets, swapping the dependent and independent variables produces a significant frequency of conflicting results. Therefore, we should not arbitrarily select one trait as the dependent variable in the PGLS analysis of two traits. Taking advantage of the simulated data, we will try to find a criterion for selecting a better dependent variable.

In empirical phylogenetic data, we only know the trait values for the terminal nodes, and potential correlations among the traits could be estimated by the methods like PGLS. By contrast, in simulated phylogenetic data, we also know the trait values of the internal nodes. As the changes along different phylogenetic branches are independent, we can measure the evolutionary correlation between two traits by analyzing their changes along the phylogenetic branches using standard statistical methods. First, we calculated the changes in the trait X_1 and the trait X_2 along evolutionary branches per unit time, $\Delta X_1/L$ and $\Delta X_2/L$, where L is the branch length. Then we did the Shapiro-Wilk test on $\Delta X_1/L$ and $\Delta X_2/L$ to determine whether they have normal distributions. If both of them satisfy normality, we used the Pearson correlation to detect the correlation between $\Delta X_1/L$ and $\Delta X_2/L$. Otherwise, we used Spearman's rank correlation. These analyses revealed significant correlations between the two traits, X_1 and X_2, in 7902 simulations (7099 positive and 803 negative, $p < 0.05$ for all these cases) but not in the other 4098 simulations ($p \geq 0.05$ for all these cases) (Supplementary Table S1). These results provide "golden standards" to judge whether PGLS analyses of trait values on the terminal nodes (X_1 and X_2) give correct results.

PGLS analyses of $X_1 \sim X_2$ show that there are significant positive correlations in 6317 simulations ($p < 0.05$ for all these cases), significant negative correlations in 29 simulations ($p < 0.05$ for all these cases), and no significant correlations in 5654 simulations ($p \geq 0.05$ for all these cases) (Supplementary Table S1). The same analyses of $X_2 \sim X_1$ show that there are

significant positive correlations in 7999 simulations ($p < 0.05$ for all these cases), significant negative correlations in 19 simulations ($p < 0.05$ for all these cases), and no significant correlations in 3982 simulations ($p \geq 0.05$ for all these cases) (Supplementary Table S1). By comparing with the golden standards, we found that PGLS analyses gave correct results for both $X_1 \sim X_2$ and $X_2 \sim X_1$ in 7475 simulations (62.29%) and gave incorrect results for both $X_1 \sim X_2$ and $X_2 \sim X_1$ in 2461 simulations (20.51%). In 2064 simulations (17.20%), only one of the two competing models, $X_1 \sim X_2$ or $X_2 \sim X_1$, gave correct results. Therefore, limited by the performance of PGLS regression analysis, we could at most get an accuracy of 79.49% in analyzing the data of our 12000 simulations by PGLS regressions. However, if we arbitrarily select one trait (X_1 or X_2) as the independent variable with the most bad fortune, we could get an accuracy of only 62.29%. In the following attempt to find an accurate criterion for dependent variable selection, we hope to perceive more correct cases from the 2064 simulations where $X_1 \sim X_2$ and $X_2 \sim X_1$ gave conflicting results.

Looking for an Accurate Criterion for Dependent Variable Selection

Referring to the golden standards, we evaluated six potential criteria for their performance in selecting a better model from $X_1 \sim X_2$ and $X_2 \sim X_1$.

In statistics, the goodness of fit of two competing statistical models is often assessed by calculating each model's log-likelihood (LLK) values. We first examined whether a higher (or lower) LLK could give an accurate prediction of the better model between $X_1 \sim X_2$ and $X_2 \sim X_1$. By calculating the LLKs of the two models for the 2064 simulations (Supplementary Table S1), we found that the models selected from $X_1 \sim X_2$ and $X_2 \sim X_1$.by a lower LLK (denoted as $Model_{LLK,lower}$) have more correct results than the alternate model (denoted as $Model_{LLK,higher}$), 1079 *vs.* 985. A χ^2 test showed that the difference is statistically significant ($p = 0.004$, Table 3).

Akaike information criterion (AIC) is a widely used estimator of the quality of statistical models for a given dataset (Akaike 1974). It balances the goodness of fit of the model and the model's complexity. By calculating the AIC values of the two competing models for the 2064 simulations (Table 1), we found that the models selected by a higher AIC value (denoted as $Model_{AIC,higher}$) have significantly more correct results than the alternate model (denoted as $Model_{AIC,lower}$), 1079 *vs.* 985, $p = 0.004$ (Table 3).

The R^2 describes the proportion of the total variation in the dependent variable that is explained by the independent variables in the regression model, and the p-value of the regression coefficient is the probability of observing the test statistic value under the assumption that the null hypothesis is true, where the regression coefficient equals 0. These two parameters are widely used indicators of the goodness of fit of regression models. The models selected by a higher R^2 (denoted as $Model_{R^2,higher}$) and a lower p-value (denoted as $Model_{p,lower}$) have significantly more correct results than the alternate models ($p = 4 \times 10^{-4}$ for both cases, Table 3).

The phylogenetic signal is a measure of the extent to which the phylogenetic structure influences species trait values. Pagel's λ is the most commonly used indicator of the phylogenetic signal (Pagel 1999). The estimated λ ($\hat{\lambda}$) is a parameter in Pagel's λ model (Pagel 1997) that measures the relatedness of the regression residuals with the phylogenetic structure. We found that the models that use the trait with a higher λ value as the dependent variable (denoted as $Model_{\lambda_y,higher}$) have significantly more correct results than the alternate models (denoted as $Model_{\lambda_y,lower}$) ($p < 2.2 \times 10^{-16}$, Table 3). And the model selected by a higher $\hat{\lambda}$ (denoted as $Model_{\hat{\lambda},higher}$) also have significantly more correct results than the alternate models (denoted as $Model_{\hat{\lambda},lower}$) ($p < 2.2 \times 10^{-16}$, Table 3).

Table 3. Differences in the performance of models selected by six criteria.

Criteria	Models	Correct	Incorrect	χ^2	p^a
LLK	$Model_{LLK,higher}$	985	1079	8.381	0.004
	$Model_{LLK,lower}$	1079	985		
AIC	$Model_{AIC,higher}$	1079	985	8.381	0.004
	$Model_{AIC,lower}$	985	1079		
R^2	$Model_{R^2,higher}$	1089	975	12.37	4×10^{-4}
	$Model_{R^2,lower}$	975	1089		
p	$Model_{p,higher}$	975	1089	12.37	4×10^{-4}
	$Model_{p,lower}$	1089	975		
λ_y	$Model_{\lambda_y,higher}$	1736	328	1918.3	$< 2.2 \times 10^{-16}$
	$Model_{\lambda_y,lower}$	328	1736		
$\hat{\lambda}$	$Model_{\hat{\lambda},higher}$	1736	328	1918.3	$< 2.2 \times 10^{-16}$
	$Model_{\hat{\lambda},lower}$	328	1736		

[a]The minimal significance value that the R environment (version 4.1.0) gives is 2.2×10^{-16}. Llk, log-likelihood; AIC, Akaike information criterion. When the χ^2 tests were extended to all the 12000 simulations, λ_y and $\hat{\lambda}$ are also the best criteria (Supplementary Table S2).

Furthermore, we defined a virtual criterion where one model was randomly chosen from the two competing models, $X_1 \sim X_2$ and $X_2 \sim X_1$. The results of the models selected by this virtual criterion ($Model_{rc}$) were compared with the better models selected by the above six criteria. Although the better models selected by the above six criteria consistently have more correct results than $Model_{rc}$ (Table 4), the differences of $Model_{LLK,lower}$ and $Model_{AIC,higher}$ with $Model_{rc}$ are not statistically significant ($p = 0.093$ for both cases),

the differences of $Model_{R^2,higher}$ and $Model_{p,lower}$ with $Model_{rc}$ are marginally significant ($p = 0.046$ for both cases), but the differences of $Model_{\hat{\lambda},higher}$ and $Model_{\lambda_y,higher}$ are highly significant ($p < 2.2 \times 10^{-16}$ for both cases).

Table 4. Comparing the performance of six criteria with random choosing.

Criteria	Models	Correct	Incorrect	χ^2	p[a]
$\hat{\lambda}$	$Model_{\hat{\lambda},higher}$	1736	328	552.7	$< 2.2 \times 10^{-16}$
	$Model_{rc}$	1024	1040		
λ_y	$Model_{\lambda_y,higher}$	1736	328	552.7	$< 2.2 \times 10^{-16}$
	$Model_{rc}$	1024	1040		
LLK	$Model_{LLK,lower}$	1079	985	2.827	0.093
	$Model_{rc}$	1024	1040		
AIC	$Model_{AIC,higher}$	1079	985	2.827	0.093
	$Model_{rc}$	1024	1040		
R^2	$Model_{R^2,higher}$	1089	975	3.971	0.046
	$Model_{rc}$	1024	1040		
p	$Model_{p,lower}$	1089	975	3.971	0.046
	$Model_{rc}$	1024	1040		

[a]The minimal significance value that the R environment (version 4.1.0) gives is 2.2×10^{-16}.

LLK, log-likelihood; AIC, Akaike information criterion.

From Tables 3-4 and S2, it could be seen that each pair of criteria gave identical results, LLK and AIC, R^2 and p-value, Pagel's λ and $\hat{\lambda}$. Among the three pairs, Pagel's λ and $\hat{\lambda}$ seem

to be the best criterion for dependent variable selection. For a quantitative evaluation of these impressions, we performed the χ^2 tests to compare Pagel's λ with $\hat{\lambda}$, LLK, AIC, R^2, and p-value using the 2064 simulations where $X_1 \sim X_2$ and $X_2 \sim X_1$ gave conflicting results. As shown in Table 5, the equivalency between Pagel's λ and $\hat{\lambda}$ and the superiority of λ to LLK, AIC, R^2, and p-value have been statistically confirmed.

Table 5. Comparing the performance of λ_y with other criteria for dependent variable selection

Criteria	Models	Correct	Incorrect	χ^2	p[a]
λ_y *vs.* $\hat{\lambda}$	$Model_{\lambda_y,higher}$	1736	328	0	1
	$Model_{\hat{\lambda},higher}$	1736	328		
λ_y *vs.* LLK	$Model_{\lambda_y,higher}$	1736	328	480.6	$< 2.2 \times 10^{-16}$
	$Model_{LLK,lower}$	1079	985		
λ_y *vs.* AIC	$Model_{\lambda_y,higher}$	1736	328	480.6	$< 2.2 \times 10^{-16}$
	$Model_{AIC,higher}$	1079	985		
λ_y *vs.* R^2	$Model_{\lambda_y,higher}$	1736	328	468	$< 2.2 \times 10^{-16}$
	$Model_{R^2,higher}$	1089	975		
λ_y *vs.* p	$Model_{\lambda_y,higher}$	1736	328	468	$< 2.2 \times 10^{-16}$
	$Model_{p,lower}$	1089	975		

[a]The minimal significance value that the R environment (version 4.1.0) gives is 2.2×10^{-16}.

LLK, log-likelihood; AIC, Akaike information criterion.

In summary, Pagel's λ and $\hat{\lambda}$ are the best criteria for dependent variable selection. Among the 2064 simulations where $X_1 \sim X_2$ and $X_2 \sim X_1$ gave conflicting results, these two criteria led to correct results in 1736 simulations. Combined with the 7475 simulations that both $X_1 \sim X_2$ and $X_2 \sim X_1$ gave correct results, PGLS analysis can achieve an accuracy of

76.8% when the trait with a stronger phylogenetic signal was selected as the dependent variable. As this accuracy is still 2.7% lower than the upper limit of PGLS analysis (79.49%), a much better criterion might be found in the future. Moreover, the PGLS regression analysis itself should also be improved.

DISCUSSION

In the PGLS correlation analysis of an empirical dataset (Liu et al. 2023), we recognized that swapping the dependent and independent variables could lead to a remarkable frequency of conflicting results. Then, we simulated the evolution of two traits (X_1 and X_2) along a binary tree containing 100 terminal nodes with different models and variances for 12000 times. PGLS analysis of these simulated datasets showed that swapping the dependent and independent variables gave conflicting results at a frequency of 17.2%. Taking advantage of simulation, we established a golden standard for whether X_1 and X_2 are correlated in each simulation by conventional correlation analysis of the changes of the two traits along the branches of the phylogenetic tree. With this golden standard, we can tell which model, $X_1 \sim X_2$ or $X_2 \sim X_1$, is correct. Six potential criteria for dependent variable selection, LLK, AIC, R^2, p-value, Pagel's λ, and $\hat{\lambda}$, have been compared. The last two criteria are equivalent in dependent variable selection and have exhibited their superiority to the other four criteria. The Pagel's λ values are generally calculated at the beginning of a phylogenetic comparative analysis, so they are already known before the PGLS analysis. If we can choose the trait with a higher λ value as the dependent variable, two rounds of PGLS analysis, like the $X_1 \sim X_2$ and $X_2 \sim X_1$, are not required. Considering the practical convenience, Pagel's λ is superior to $\hat{\lambda}$.

It should be highlighted that the terms independent variable and dependent variable are misleading in evolutionary correlation studies. They should not be taken literally. A PGLS

regression analysis does not provide a model that uses the independent variable to explain or predict changes in the dependent variable. It replaces conventional correlation methods, like Pearson and Spearman's rank tests, in phylogenetic comparative studies. The choice of the dependent variable in a PGLS regression analysis should not be based on a pre-assumption of the cause-and-effect relationship between the analyzed traits but should guarantee an accurate perception of the relationship, whether correlated or not.

SUPPLEMENTARY MATERIAL

Supplementary material is available on GitHub at https://github.com/BNU-Genome-Evolution/dependent-variable-selection.

ACKNOWLEDGMENTS

This work was supported by the National Natural Science Foundation of China (Grant number 31671321).

REFERENCES

Akaike H. 1974. A new look at the statistical model identification. EEE Trans. Automat. Contr., 19:716-723.

Bartomeus I., Cariveau D.P., Harrison T., Winfree R. 2018. On the inconsistency of pollinator species traits for predicting either response to land-use change or functional contribution. Oikos, 127:306-315.

Bawa K.S., Ingty T., Revell L.J., Shivaprakash K.N. 2019. Correlated evolution of flower size and seed number in flowering plants (monocotyledons). Ann. Bot., 123:181-190.

Caetano D.S., Harmon L.J. 2019. Estimating correlated rates of trait evolution with uncertainty. Syst. Biol., 68:412-429.

Felsenstein J. 1985. Phylogenies and the comparative method. Am. Nat., 125:1-15.

Freckleton R.P., Harvey P.H., Pagel M. 2002. Phylogenetic analysis and comparative data: A test and review of evidence. Amer. Natur., 160:712-726.

Garamszegi L.Z. 2014. Modern Phylogenetic Comparative Methods and Their Application in Evolutionary Biology: Concepts and Practice. Berlin, Springer.

Goswami A., Smaers J.B., Soligo C., Polly P.D. 2014. The macroevolutionary consequences of phenotypic integration: from development to deep time. Philos. Trans. R. Soc. B-Biol. Sci., 369:20130254.

Grafen A. 1989. The phylogenetic regression. Philos. Trans. R. Soc. Lond. B Biol. Sci., 326:119-157.

Ho L.S.T., Ane C. 2014. A linear-time algorithm for Gaussian and non-Gaussian trait evolution models. Syst. Biol., 63:397-408.

Hu E.-Z., Lan X.-R., Liu Z.-L., Gao J., Niu D.-K. 2022. A positive correlation between GC content and growth temperature in prokaryotes. BMC Genomics, 23:110.

Liu Z.-L., Hu E.-Z., Niu D.-K. 2023. Investigating the relationship between CRISPR-Cas content and growth rate in bacteria. Microbiol. Spectr., 11:e03409-03422.

Lynch M. 1991. Methods for the analysis of comparative data in evolutionary biology. Evolution, 45:1065-1080.

Martins E.P., Hansen T.F. 1997. Phylogenies and the comparative method: A general approach to incorporating phylogenetic information into the analysis of interspecific data. Amer. Natur., 149:646-667.

Pagel M. 1997. Inferring evolutionary processes from phylogenies. Zool. Scr., 26:331-348.

Pagel M. 1999. Inferring the historical patterns of biological evolution. Nature, 401:877-884.

Paradis E., Schliep K. 2019. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. Bioinformatics, 35:526-528.

Parks D.H., Chuvochina M., Rinke C., Mussig A.J., Chaumeil P.-A., Hugenholtz P. 2022. GTDB: an ongoing census of bacterial and archaeal diversity through a phylogenetically consistent, rank normalized and complete genome-based taxonomy. Nucleic Acids Res., 50:D785-D794.

Pearman P.B., Lavergne S., Roquet C., Wüest R., Zimmermann N.E., Thuiller W. 2014. Phylogenetic patterns of climatic, habitat and trophic niches in a European avian assemblage. Glob. Ecol. Biogeogr., 23:414-424.

Revell L.J. 2010. Phylogenetic signal and linear regression on species data. Methods Ecol. Evol., 1:319-329.

Revell L.J. 2012. phytools: an R package for phylogenetic comparative biology (and other things). Methods Ecol. Evol., 3:217-223.

Revell L.J., Collar D.C. 2009. Phylogenetic analysis of the evolutionary correlation using likelihood. Evolution, 63:1090-1100.

Revell L.J., Toyama K.S., Mahler D.L. 2022. A simple hierarchical model for heterogeneity in the evolutionary correlation on a phylogenetic tree. PeerJ, 10:e13910.

Rohlf F.J. 2001. Comparative methods for the analysis of continuous variables: Geometric interpretations. Evolution, 55:2143-2160.

Suarez-Castro A.F., Mayfield M.M., Mitchell M.G.E., Cattarino L., Maron M., Rhodes J.R. 2020. Correlations and variance among species traits explain contrasting impacts of fragmentation and habitat loss on functional diversity. Landscape Ecol, 35:2239-2253.

Travnicek P., Certner M., Ponert J., Chumova Z., Jersakova J., Suda J. 2019. Diversity in genome size and GC content shows adaptive potential in orchids and is closely linked to partial endoreplication, plant life-history traits and climatic conditions. New Phytol., 224:1642-1656.

Waugh F.V. 1943. Choice of the dependent variable in regression analysis. J Am Stat Assoc, 38:210-214.

Whitney K.D., Garland T., Jr. 2010. Did genetic drift drive increases in genome complexity? PLoS Genet., 6:e1001080.