Journal of Theoretical Biology 259 (2009) 727–743

Contents lists available at ScienceDirect

Journal of Theoretical Biology

journal homepage: www.elsevier.com/locate/yjtbi

Independent contrasts and regression through the origin

Pierre Legendre [a], Yves Desdevises [b,c,*]

[a] Département de Sciences Biologiques, Université de Montréal, C.P. 6128, succursale Centre-ville, Montréal, Québec, Canada H3C 3J7
[b] UPMC Univ Paris 06, UMR 7628, Modèles en Biologie Cellulaire et Évolutive, Observatoire Océanologique, F-66651, Banyuls/Mer, France
[c] CNRS, UMR 7628, Modèles en Biologie Cellulaire et Évolutive, Observatoire Océanologique, F-66651, Banyuls/Mer, France

ARTICLE INFO

Article history:
Received 2 November 2008
Received in revised form
11 March 2009
Accepted 24 April 2009
Available online 3 May 2009

Keywords:
Comparative analysis
Permutation test
Power
Simulations
Type I error

ABSTRACT

Following the pioneering work of Felsenstein and Garland, phylogeneticists have been using regression through the origin to analyze comparative data using independent contrasts. The reason why regression through the origin must be used with such data was revisited. The demonstration led to the formulation of a permutation test for the coefficient of determination and the regression coefficient estimates in regression through the origin. Simulations were carried out to measure type I error and power of the parametric and permutation tests under two models of data generation: regression models I and II (correlation model). Although regression through the origin assumes model I data, in independent contrast data error is present in the explanatory as well as the response variables. Two forms of permutations were investigated to test the regression coefficients: permutation of the values of the response variable **y**, and permutation of the residuals of the regression model. The simulations showed that the parametric tests or any of the permutation tests can be used when the error is normal, which is the usual assumption in independent contrast studies; only the test by permutation of **y** should be used when the error is highly asymmetric; and the parametric tests should be used when extreme values are present in covariables. Two examples are presented. The first one concerns non-specificity in fish parasites of the genus *Lamellodiscus*, the second the richness in parasites in 78 species of mammals.

1. Introduction

Biologists generally agree that when looking for correlations between phenotypic traits across species, or between traits and environmental factors, one must take the phylogenetic relatedness of the species into account; see Harvey and Pagel (1991) or Martins et al. (2002) for reviews. The reason is that species cannot be considered to be independent observations; they are related to one another through their phylogeny and share inherited attributes. The phylogeny acts as a confounding variable and must be controlled for. The many approaches developed to control for the phylogeny (e.g., Stearns, 1983; Cheverud et al., 1985; Felsenstein, 1985, 2008; Grafen, 1989; Lynch, 1991; Diniz-Filho et al., 1998; Houseworth et al., 2004) are grouped under the designation "comparative analyses" or "comparative methods". The first of these techniques, which is still widely used (e.g., Laurin, 2004; Fjerdingstad and Crozier, 2006; Kolm et al., 2007; Kohlsdorf et al., 2008; Xiang et al., 2008; Poorter et al., 2008), is the method of phylogenetically independent contrasts proposed by Felsenstein (1985).

In a classical paper, Garland et al. (1992) showed how to carry out the analysis of comparative data using phylogenetically independent contrasts. This type of analysis is important, in particular, when relating phenotypic traits of species to one another, or to environmental or ecological factors, using simple or multiple regression. In summary: (1) for each variable, independent contrasts are computed for each bifurcation of the phylogenetic tree by subtracting one observed value of the variable from the other; for a fully resolved tree, there are $(n-1)$ contrasts for n objects; (2) before using them in statistical analyses, contrasts must be standardized by dividing each one by its standard error, computed as the square root of the sum of the branch lengths for this variable on the tree. Branch lengths represent evolutionary time since divergence and the variance of the character under study is proportional to time. Note that branch lengths can be transformed to meet the method's assumptions. After standardization, the branch lengths are expressed in units of expected standard deviation of change; and (3) the contrasts are analyzed using regression through the origin.

The method of independent contrasts has been developed under the Brownian motion model, which gives support to the assumption that the contrasts should be normally distributed. This applies to the evolutionary process underlying the data, but it is no guarantee that the contrasts computed from observed variables will actually be normally distributed. There are three

* Corresponding author. Tel.: +33 04 68 88 73 13; fax: +33 04 68 88 16 99.
E-mail addresses: pierre.legendre@umontreal.ca (P. Legendre), yves.desdevises@obs-banyuls.fr (Y. Desdevises).

0022-5193/$ - see front matter © 2009 Elsevier Ltd. All rights reserved.
doi:10.1016/j.jtbi.2009.04.022

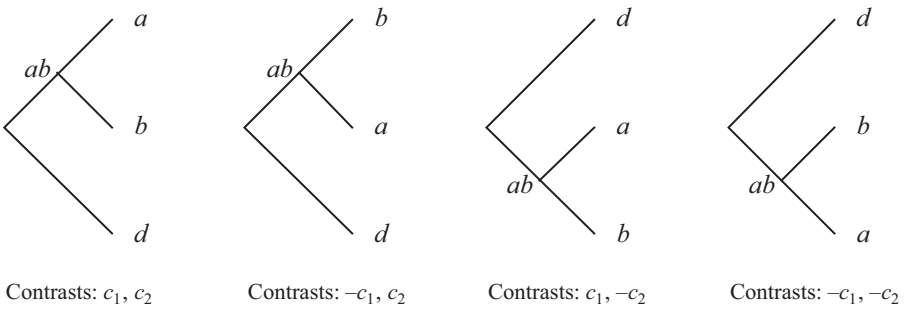

Fig. 1. Three-species example showing the contrasts observed on all $2^{(n-1)} = 4$ possible flipped-branch trees.

main reasons for this: (1) we measure variables on physical scales that often make them, as well as the contrasts calculated from them, non-normal. This is true of many of the ecological variables that are analyzed using independent contrasts. Examples are: basal metabolic rate (27.1–18,943 ml O_2/g h), mammal density (0.02–7500 ind/ha), body mass (3–65,320 g) in the study of Morand and Harvey (2000); host geographical range (32,690–505,000 km^2), longevity (12–60 months), parasite species richness (4–28 species) in Feliu et al. (1997). Users of independent contrasts often find it useful to transform the data to approach normality before computing contrasts, but also to solve problems of allometry (e.g., Diaz-Uriarte and Garland, 1996, 1998); (2) there are cases where we can be rather confident that the evolution of the trait under study can be modelled by Brownian motion (see Felsenstein, 1985, 1988; Hansen and Martins, 1996; Houseworth et al., 2004), but the contrasts are not normally distributed because the data (e.g., molecular sequences) and/or the method used to construct the tree did not produce an unbiased estimate of the true tree. In particular, branch lengths, which are in units of expected evolutionary change, may not accurately represent time, which is a strong assumption of the independent contrasts method; and (3) the clade under study may not be entirely or randomly sampled; this may result in highly asymmetric distributions, including the presence of extreme values (outliers). In these situations, it can often be extremely difficult to find a transformation that will effectively normalize the data and prevent extreme contrast values from exerting high leverage in regression models. These limitations have sometimes precluded the use of independent contrasts in previous studies (e.g., Pouydebat et al., 2008). Parametric tests in regression through the origin cannot be used to identify relationships between sets of computed contrasts in such cases, because of the lack of normality of the contrasts, but permutation tests can. However, the independent contrasts method always relies on the assumption of a Brownian motion model of phenotypic evolution, regardless of the testing procedure used to study the relationships between contrasts. These situations define the domain of application of the permutation test described in this paper.

In an appendix to their paper, Garland et al. (1992) gave algebraic reasons why regression through the origin should be used, but they did not provide an intuitive geometric interpretation. Users of the method may be wondering whether the algebraic reasons given are sufficient, or whether estimation should not allow for departures from the ideal model. Doubts are nourished by the observation that, in many instances of contrasts, the regression line does not seem to willingly go through the origin. Kvålseth (1985) and Neter et al. (1996) commented that regression through the origin has to be used with caution. If the regression model has an intercept near zero, there is no harm in estimating it; if it does not, the regression-through-the-origin model is probably inadequate for the data at hand. What about independent contrast data which, in most instances, do not seem to obey a linear model going through the origin?

The present paper recalls the statistical reasons why regression through the origin should be used in this type of analysis, and supports the recommendation of Garland et al. (1992) through additional geometric reasons. The geometric line of reasoning leads to the formulation of a permutation test for regression through the origin. This type of test can be used when the data are not normally distributed.

2. Regression through the origin

Regression through the origin can alternatively be described as a form of linear regression based upon a doubled data set. This property will be used as the basis for a double-permutation procedure, described in this paper for testing the significance of R^2 and the regression coefficients. Consider an explanatory variable **x** whose values complement the *nsp* species names labelling the leaves of the tree. A contrast is noted $\Delta x = x_a - x_b$, for any two sister species a and b; likewise for the internal nodes found at the various bifurcation points of the tree. When computing contrasts, one makes the arbitrary decision that a, for instance, is the 'upper' species or node (for a tree drawn sideways) and b is the 'lower' one, or the opposite.

There are $n = (nsp-1)$ contrasts in any bifurcating tree of *nsp* species. A given tree leads to the calculation of particular values for each contrast, $c = \Delta x = x_a - x_b$, for variable **x**. Depending on the way the tree happens to be drawn, either c or $-c$ can be obtained at each node. Actually, branches can be swapped at any node of a tree without changing the phylogeny that it represents. Since the order (upper or lower) of the branches at any node is arbitrary, we are just as likely to observe $\Delta x = x_b - x_a$ as we are to obtain $\Delta x = x_a - x_b$. Likewise for any contrast $\Delta y = y_a - y_b$ of a response variable **y**. The only constraint is that the direction of the subtraction must be the same for all variables. Hence, the particular set of contrasts observed on a tree has signs that could very well have been partly or entirely different, had the tree be drawn in some other equivalent way. There is no reason to give more importance to the set of contrast values that has been