Volume 259
Issue 4
21 August 2009
ISSN 0022-5193
ELSEVIER
Journal of Theoretical Biology

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

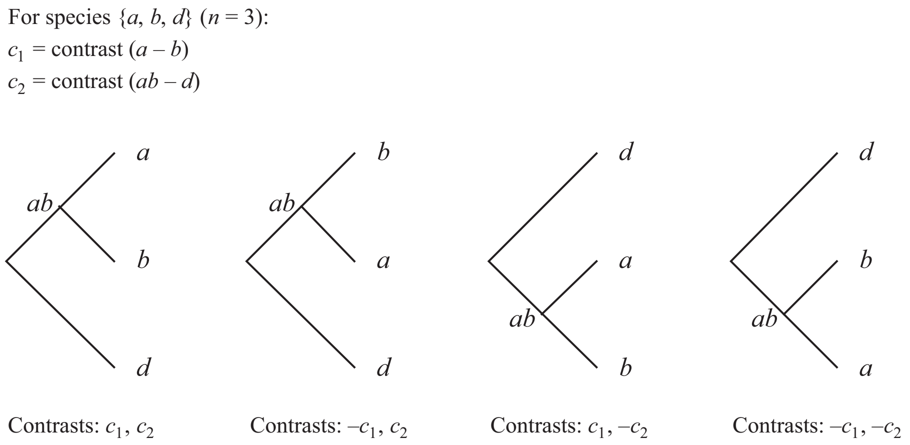

Fig. 1. Three-species example showing the contrasts observed on all $2^{(n-1)} = 4$ possible flipped-branch trees.

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 obtained than to another set that could be calculated on any other tree obtained by swapping branches at any or all nodes.

We can combine the sets of contrasts from all possible swapped trees. For *nsp* species and $n=(nsp-1)$ contrasts in a fully resolved binary tree, $2^{(nsp-1)}$ different trees can be drawn, since branches can be swapped at every node. When swapping branches at a given node, the contrasts calculated at that node change signs. Each contrast appears $2^{(nsp-2)}$ times in the data set combining the contrasts from all possible swapped trees. In Fig. 1 for instance, where $nsp=3$, there are $2^{(nsp-1)}=4$ possible trees containing 2 contrasts each. Each contrast appears $2^{(nsp-2)}=2$ times in the combined data set.

Analysis of the relationship between the contrasts of the dependent ($\mathbf{y}$) and independent ($\mathbf{x}_1$, $\mathbf{x}_2$, $\ldots$, $\mathbf{x}_m$) variables can be done by simple or multiple ordinary least-squares (OLS) linear regression, using the combined contrasts from all possible trees. What we are seeking is the expected regression parameter(s) for the set of all possible flipped trees.

The exact same regression line can be obtained by using a smaller data set containing each signed contrast only once. This data set, hereafter called the *doubled data set*, contains $2n=2(nsp-1)$ values for each variable; each contrast is represented once by a value having a positive sign (c) and another time by a value having a negative sign ($-c$). Any one set of calculated contrasts contains half this number of values, in any particular case, since we are computing $\Delta x=x_a-x_b$, for example, and not $\Delta x=x_b-x_a$.

An equivalent calculation is to use regression through the origin on the original set of $n=(nsp-1)$ contrast values; this is illustrated by the example in the next section. The equivalence between regression through the origin and simple linear regression on a doubled data set will be used below to design a permutation procedure for the tests of significance in regression through the origin. The slope of the regression line obtained is the same using regression through the origin or by simple linear regression on the doubled data set. Only the slope parameter(s) have to be estimated, not the intercept which is fixed at 0 by construction.

The demonstration in Garland et al. (1992) is clear about the number of degrees of freedom that should be used to test the significance of the regression parameters. The number of contrasts is $n=(nsp-1)$ for *nsp* species, whereas the number of estimated parameters is equal to the number of variables (m), not (m+1) since the intercept is fixed at zero by construct and, thus, does not have to be estimated from the data. These concepts are illustrated by the following example.

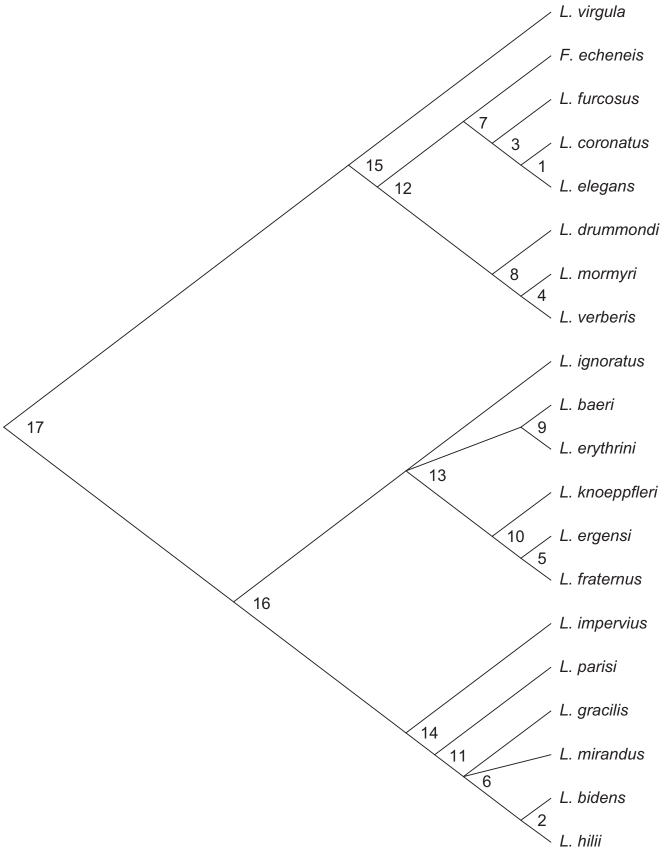

Fig. 2. Phylogenetic trees of the Mediterranean *Lamellodiscus* parasites estimated from 18S rDNA partial sequences. The labels identify the contrasts calculated at the nodes of the tree.

3. Example 1: specificity of *lamellodiscus* parasites, part 1

When a program for regression through the origin is not available, correct estimates of the regression parameters can be obtained as illustrated by the following example. The data are from Desdevises et al. (2002a, b) who studied the factors that affect parasite specificity (parasites of the genus *Lamellodiscus*: Monogenea, Diplectanidae) with respect to their teleostean hosts (Sparidae) in the Mediterranean. The response variable that we will consider is a non-specificity index (NSI). NSI is a semi-quantitative descriptor of specificity (Desdevises et al., 2002b) recorded as follows: (1) *specialists* using a single host; (2) *intermediate specialists* using two closely related hosts; (3) *intermediate generalists* using two or more hosts in the same terminal clade; and (4) *generalists* using two or more hosts across several clades. The lower NSI is, the higher is host specificity, hence its name. Contrasts computed from NSI all take different values; so, these contrasts will be treated as a quantitative variable in the present example. The explanatory variable is the maximum size of the host species. Standardized contrasts were computed using the program CAIC version 2.6.9 (Purvis and Rambaut, 1995), based upon a maximum likelihood reconstruction of the phylogeny of the Mediterranean *Lamellodiscus* (Fig. 2; Desdevises et al., 2002a). After trying all combinations of contrasts computed from the original and the log-transformed NSI and maximum host size data, we found that the regression with both variables log-transformed before computation of the contrasts produced the highest *R*-square. That regression is used here to illustrate the method. The contrasts used in this illustrative example are shown in Table 1. Table 2 presents the parameters and statistics computed using ordinary least-squares regression through the origin for the original set of contrast data, and also using ordinary least-squares regression on a "doubled data table" in which each row of the data set of contrasts is doubled by adding a row with opposite signs. For example, the contrast vector (0.04152, 0.00405) is doubled by adding the vector (−0.04152, −0.00405) to the data table.

Fig. 3a shows that the regression line through the origin (slope = −1.30324) differs from the ordinary OLS regression line (slope = −0.99395). In Fig. 3b, the five leftmost points were moved to the right of the scattergram by changing their signs on both coordinates; this result corresponds to flipping the corresponding nodes of the *Lamellodiscus* tree. The regression line through the origin remains unchanged (slope = −1.30324), but the OLS regression line has changed (slope = −1.70849). In Fig. 3c, the doubled data table is used for regression: the OLS regression line is now identical to the regression line through the origin (slopes = −1.30324). A doubled data table has also been used by Ackerly and Donoghue (1998) to obtain principal component axes of contrast data that passed through the origin.

Table 2 shows that by using OLS regression on a "doubled data table", one obtains correct estimates of the intercept and slope parameters, and of the coefficient of determination (R^2). All statistics involved in tests of significance are incorrect because the correct number of degrees of freedom is 16, since the number of independent contrasts is 17 (not 34) and the number of parameters to be estimated is 1 (not 2). The F-statistic obtained for the "doubled data table" has a doubled value and twice the number of degrees of freedom in the denominator, but this does not lead to a correct probability estimate. Likewise, the standard error of the regression coefficient (b) and its t-value are incorrect, hence the test of significance for b is wrong.

Table 1
Contrasts computed from the log-transformed variables, computed on the phylogenetic tree of *Lamellodiscus* parasites (from Desdevises et al., 2002b): non-specificity index (NSI, dependent variable) and maximum host size (explanatory variable).

Contrasts	NSI	Maximum host size
1	0.04152	0.00405
2	0.00000	0.00000
3	0.01716	0.03023
4	0.00000	0.00000
5	0.16553	0.09463
6	0.45859	−0.20256
7	0.18470	−0.11613
8	0.00000	0.09224
9	0.00000	−0.03719
10	−0.08754	−0.08257
11	0.11719	−0.02669
12	0.16120	−0.00904
13	0.25614	−0.07076
14	0.12913	−0.08901
15	0.09254	−0.04756
16	0.11082	−0.00829
17 (Root)	0.01401	0.04786

The numbers in the left-hand column identify the contrasts in Fig. 2.

Any OLS regression line passes through the centroid of the data points. It is for this reason that the OLS regression on a "doubled data set" always passes through the origin. Table 2 has shown that OLS regression on a "doubled data table" did not provide correct tests of significance of the regression parameters. It may, however, provide hints as to how a permutation test of the coefficient of determination (R^2) and the partial regression coefficients should be constructed for regression through the origin.

4. Parametric and permutation tests

Regression through the origin is available in a number of statistical packages. When the contrasts are not normally distributed or contain extreme values, for the reasons described in the Introduction, the parameters of the regression equation should be tested using a permutational procedure. The simulations reported in the next two sections will show that the proposed permutation test is indeed insensitive to a lack of normality (even very strong asymmetry) in the data.

We will now describe a permutation procedure for regression through the origin and compare it to the parametric test using numerical simulations, in order to demonstrate its validity. Before we discovered the procedure described in the next paragraphs, we carried out simulations for two more simple forms of permutation tests: (1) permuting at random the values of the response contrasts $\mathbf{y}$ with respect to the explanatory contrasts $\mathbf{x}$ in the original (not doubled) data set and using regression through the origin, and (2) permuting at random the values of $\mathbf{y}$ with respect to $\mathbf{x}$ in the doubled data set and using simple linear regression. Results of these procedures are reported at the end of the simulation results: these simple permutation methods did not have correct rates of type I error or failed in power. The double-permutation procedure that will now be described was developed to correct these problems.

4.1. Permutation test: double-permutation procedure

The permutational method that we will describe for testing the significance of R^2 as well as the regression coefficients is a

Table 2
Comparison of regression parameters and statistics for non-specificity index as a function of maximum host size.

	OLS regression through the origin	OLS regression on doubled data table	Estimates
n	17	34	
Intercept	Forced to 0	0.00000	Correct
Slope (b)	−1.30324	−1.30324	Correct
Standard regression coefficient	−0.75937	−0.63078	Incorrect
r	−0.63078	−0.63078	Correct
R^2	0.39788	0.39788	Correct
R^2_{adj}	0.36025	0.37907	Incorrect
F	10.57295	21.14589	Double
Degrees of freedom	$n-1=16$	$n-2=32$	Double
Standard error of b	0.40080	0.28341	Incorrect
t-value for b	−3.25161	−4.59847	Incorrect
p-value for R^2			
Parametric	0.00500	0.00006	Incorrect
Permutational (99,999 perm.)	0.00763	0.00012	Incorrect
p-value for t			
Parametric, one-tailed	0.00250	0.00003	Incorrect
Parametric, two-tailed	0.00500	0.00006	Incorrect
Permutational, one-tailed (99,999 perm.)	0.00380	0.00005	Incorrect
Permutational, two-tailed (99,999 perm.)	0.00763	0.00012	Incorrect

Center-left: regression through the origin. Center-right: regression on the "doubled data table". The right-hand column indicates whether the estimates by regression on the doubled data table are correct or incorrect.

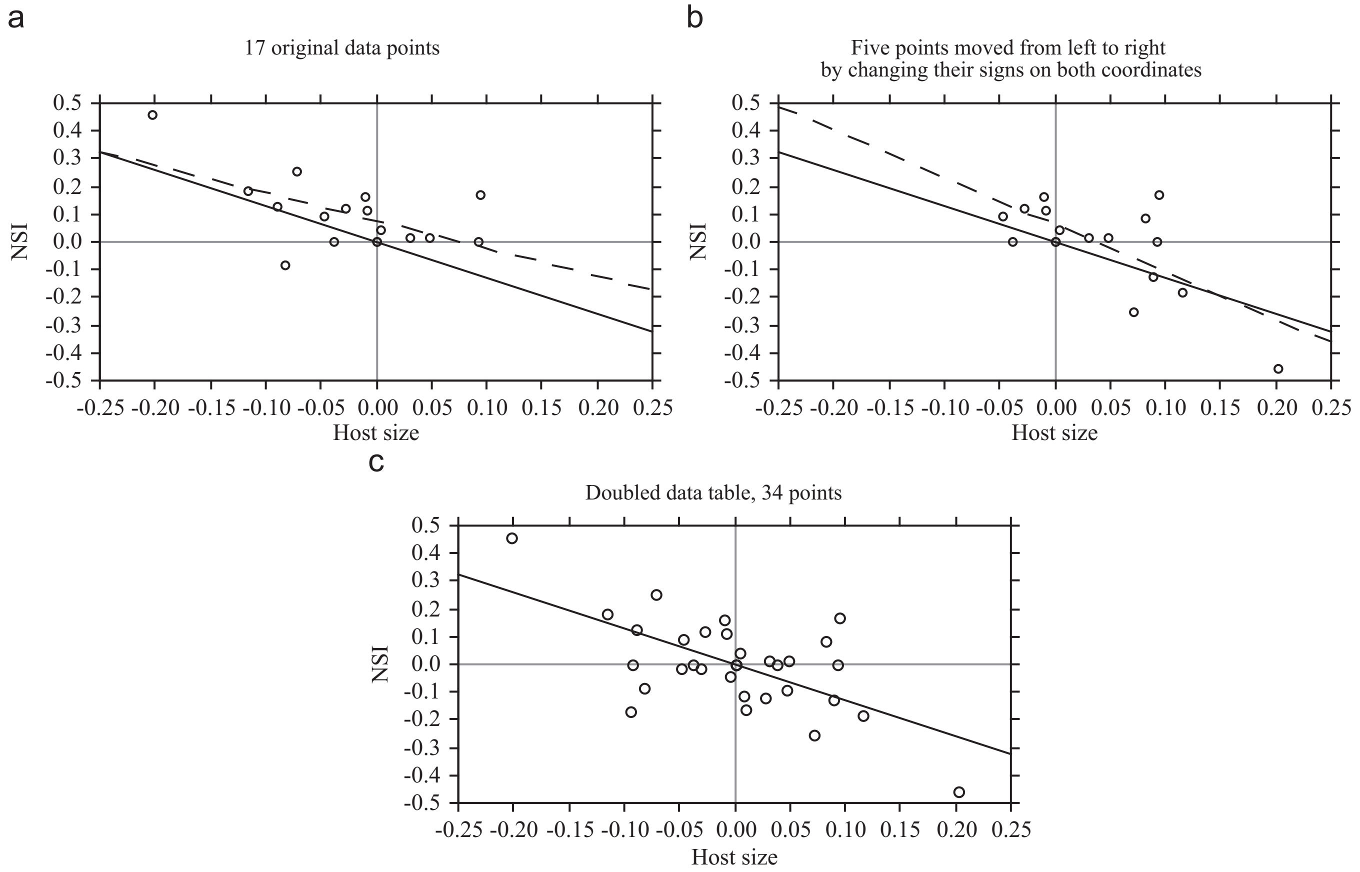

Fig. 3. Regression of contrasts of NSI on contrasts of maximum host size: (a) regression line through the origin (full line) and OLS regression line (dashed) for the original data table containing 17 pairs of contrasts; (b) same, after moving five points from the leftmost portion of the scattergram to the right by changing their signs on both variables; (c) when using the doubled data table, the two regression lines are identical.

double-permutation procedure involving n independent contrasts, or in general n observed values y_i of the dependent variable $\mathbf{y}$. The method is based on the concept of doubling the data set, used in the previous sections to explain why regression through the origin can be used in the case of independent contrasts. The objective is to obtain, under the null hypothesis (H_0) of the test, repeated randomized scatters of points of the doubled data set that have slopes near zero. The slope estimated for the real data set can then be compared to the distribution of slopes obtained under H_0 in order to test the hypothesis that the actual slope is not different from 0. The procedure is the following:

(1) Compute the (multiple) regression through the origin of $\mathbf{y}$ on $\mathbf{x}$ (or matrix $\mathbf{X}$) using the n observed sets of values (y_i, x_i) with y_i representing the values of the response and x_i the values of the single explanatory variable, or (y_i, x_{i1}, x_{i2},..., x_{ij}, ..., x_{im}) in the case of a multiple regression with y_i the values of the response and (x_{i1}, x_{i2}, ..., x_{ij}, ..., x_{im}) the values of all the explanatory variables. Calculate R^2, the F-statistic associated with R^2, the vector of Gaussian multipliers located on the diagonal of matrix $[\mathbf{X}'\mathbf{X}]^{-1}$, the regression coefficients b_j for the explanatory variables $\mathbf{x}_j$, the standard errors of the regression coefficients SE(b_j), and the associated t_j-statistics; see Appendix A.

(2) Consider the vector $\mathbf{y} = [y_1\ y_2\ \ldots\ y_i\ \ldots\ y_n]$ and the vector of doublets $\mathbf{x}_d^t = [(x_1, -x_1)(x_2, -x_2)\ldots(x_i, -x_i)\ldots(x_n, -x_n)]$ where $\mathbf{x}^t$ denotes the transposed of vector $\mathbf{x}$. For a multiple regression involving m regressors, consider the doubled matrix with n rows and $2m$ columns:

$$\mathbf{x}_d = \begin{bmatrix} [x_{11}\ldots x_{1m}] & [-x_{11}\ldots - x_{1m}] \\ [x_{21}\ldots x_{2m}] & [-x_{21}\ldots - x_{2m}] \\ \ldots & \ldots \\ [x_{i1}\ldots x_{im}] & [-x_{i1}\ldots - x_{im}] \\ \ldots & \ldots \\ [x_{n1}\ldots x_{nm}] & [-x_{n1}\ldots - x_{nm}] \end{bmatrix} \tag{1}$$

(3) Permute vector $\mathbf{y}$ at random to obtain a vector of permuted values $\mathbf{y}^* = [y_1^*\ y_2^*\ldots y_i^*\ldots y_n^*]$. This is the first level of permutation, noted by a single asterisk, involved in this procedure.

(4) Create a vector of doublets of the y values:

$$\mathbf{y}_d^* = [(y_1^*, -y_1^*)(y_2^*, -y_2^*)\ldots(y_i^*, -y_i^*)\ldots(y_n^*, -y_n^*)] \tag{2}$$

(5) For each doublet y_i, draw a number u_i at random from a uniform distribution $U(0,1)$. If <0.5, leave the corresponding pair unmodified. If ≥ 0.5, change the order of the elements in the pair. One might obtain, for instance:

$$\mathbf{y}_d^{**} = [(-y_1^*, y_1^*)(y_2^*, -y_2^*)\ldots(y_i^*, -y_i^*)\ldots(-y_n^*, y_n^*)]$$

This is the second level of permutation, noted by two asterisks, in this procedure.

(6) Create the doubled response vector $\mathbf{y}_d^{**}$ and explanatory matrix $\mathbf{X}_d$ that will be used in the regression procedure (the original and doubled portions are separated by dashes):

$$\mathbf{y}_d^{**} = \begin{bmatrix} y_1^{**} \\ \cdots \\ y_i^{**} \\ \cdots \\ y_n^{**} \\ -- \\ -y_1^{**} \\ \cdots \\ -y_i^{**} \\ \cdots \\ -y_n^{**} \end{bmatrix}$$

for example:

$$\mathbf{y}_d^{**} = \begin{bmatrix} -y_1^{*} \\ \cdots \\ y_i^{*} \\ \cdots \\ -y_n^{*} \\ -- \\ y_1^{*} \\ \cdots \\ -y_i^{*} \\ \cdots \\ y_n^{*} \end{bmatrix} \quad \text{and} \quad \mathbf{x}_d = \begin{bmatrix} x_{11} & \cdots & x_{1m} \\ \cdots & \cdots & \cdots \\ x_{i1} & \cdots & x_{im} \\ \cdots & \cdots & \cdots \\ x_{n1} & \cdots & x_{nm} \\ -- & -- & -- \\ -x_{11} & \cdots & -x_{1m} \\ \cdots & \cdots & \cdots \\ -x_{i1} & \cdots & -x_{im} \\ \cdots & \cdots & \cdots \\ -x_{n1} & \cdots & -x_{nm} \end{bmatrix} \tag{3}$$

The size of the permutation set (i.e., the number of possible, different permutations) is $n!$ for the first permutation (step 3) and 2^n for the second permutation (step 5). So, the permutation set for the double-permutation procedure is of size $(n!)(2^n)$.

(7) Compute the (multiple) regression of $\mathbf{y}_d^{**}$ against $\mathbf{x}_d$ (or $\mathbf{X}_d$). One can apply: (1) an OLS (multiple) regression procedure to the doubled data sets created during step 6, or (2) a procedure for (multiple) regression through the origin using the doubled data sets, or else (3) regression through the origin using either the upper or the lower half of the doubled data sets as presented in step 6. Compute the R^2 statistic and the vector of regression coefficients $\mathbf{b} = [b_j\ ]$: the obtained values are the same for all three regression methods.

(8) Calculation of the degrees of freedom is based upon the n original observations (contrasts or other types), not the $2n$ values of the doubled data sets. One degree of freedom is lost for each of the estimated regression parameters, but none is lost for the intercept, even if OLS regression is used: the intercept does not have to be estimated since it is 0 by construct.

(9) The permutational test of significance for R^2 can be based either upon the R^2 statistic itself, or on the derived F-statistic. In simple or multiple regression, R^2 is a statistic equivalent to F for permutation testing because F is a monotonic function of R^2 for any constant value of n and m (Manly, 1997; Legendre and Legendre, 1998). R^2 can be compared to the distribution of R^{2^*} values obtained under permutation, or F can be compared to the distribution of F^* values obtained under permutation. Under permutation, one must be careful to use the correct numbers of degrees of freedom in the calculation of F^*:

$$F^* = \frac{R^{2^*}/m}{(1 - R^{2^*})/(n-m)} \tag{4}$$

Otherwise, the values F^* will not be comparable to the F-statistic computed for the unpermuted data. The probability $P(F)$ associated with R^2 will be estimated by comparing the value F to the distribution of the F^* obtained under permutation. The reference value F is added to the distribution of F^* values before computing the probability (Hope, 1968) to insure that the test is valid (Edgington, 1995).

(10) In multiple regression, through the origin or not, the t_j-statistic associated with regression coefficient b_j is used for testing. Since t_j is a pivotal statistic, it is expected to produce correct type I error and is thus appropriate for permutation testing. This is not the case for b_j: under permutation, the values b_j^* are not monotonic to the corresponding values t_j^* because the standard error of the partial regression coefficient, $\mathrm{SE}(b_j)$, changes from one permutation to the next (Legendre and Legendre, 1998). For permutation testing, the two statistics are only equivalent in the case of simple linear regression. One must be careful to compute the t_j^*-statistics correctly under permutation. The standard error of b_j^* is computed as

$$SE(b_j^*) = \left[\frac{\sum_{i=1}^{n} (\mathrm{residual}_i)^2}{(n-m)} \times \text{Gaussian multiplier}_j \right]^{1/2} \tag{5}$$

One can either use the Gaussian multipliers obtained during step 1, or recompute them during each permutation using the first n points only of the double data set of step 6. The latter procedure would be a waste of computer time since the column vectors of matrix $\mathbf{X}$ are not permuted with respect to one another during the permutations, in accordance with the principle of ancillarity (which means *relatedness*; Welch, 1990; ter Braak, 1992). Hence $[\mathbf{X}'\mathbf{X}]^{-1}$ remains unchanged through the permutations. The values t_j^* can now be computed as

$$t_j^* = b_j^*/\mathrm{SE}(b_j^*) \tag{6}$$

The probability $P(t_j)$ associated with b_j will be estimated by comparing the value t_j to the distribution of the t_j^* obtained under permutation. The reference value t_j is added to the distribution of t_j^* values before computing the probability (Hope, 1968).

For permutation testing, one could use a pseudo-F statistic computed without degrees of freedom instead of the classical F-statistic. The degrees of freedom form a multiplicative constant which has the same effect on the unpermuted value F and on all permuted values F^* of the statistic. Hence, the permutational probability calculated using F or pseudo-F would be the same. One must exert caution and make sure that F is compared to the distribution of F^*, or pseudo-F to the distribution of pseudo-F^*. Likewise, one can compute a pseudo-SE(b_j) and pseudo-t statistic without degrees of freedom and then compare pseudo-t to the distribution of pseudo-t^*, instead of comparing t to the distribution of t^*.

Permutations will be carried out in two different ways: (1) by permuting the values of $\mathbf{y}$, as described above, or (2) permuting the residuals of the full regression model, where the permuted elements are the residuals of the regression of $\mathbf{y}$ on $\mathbf{X}$. A third method, which consists in permuting the residuals of a null model, will not be used here because, in multiple regression, it requires a new set of permutations to test each regression coefficient. The three methods are described in Legendre and Legendre (1998), ter Braak and Smilauer (1998), Anderson and Legendre (1999), and Legendre (2000). They have been compared by Anderson and Legendre (1999), using numerical simulations, in tests of partial regression coefficients, and by Legendre (2000) in tests of partial correlation coefficients. Permutation of the values of **y** is appropriate for testing R^2. In previous studies, permuting the residuals has been found in some situations to be better than permuting the values of **y**, so it must be investigated here for regression through the origin:

- The simulations of Anderson and Legendre (1999) showed that the normal-theory *t*-tests for partial regression coefficients had incorrect level of type I error, and less power than any of the permutation methods, when the error in the data departed strongly from normality. All methods of permutation gave asymptotically equivalent results in most situations and had good power. Permutation of the values of **y** had destabilized type I error when the covariable contained extreme values; the two methods of permutation of residuals were more appropriate in that situation.
- In partial correlation analysis (Legendre, 2000), with highly skewed data, the normal-theory *t*-test had again inflated type I error rates; for very small sample sizes ($n<20$) and in the absence of extreme values, permutation of the values of **y** was not affected by non-normal error whereas the two methods of permutation of residuals had slightly inflated type I error rates. With normal error, when extreme values were present in the covariable, permutation of the values of **y** had inflated type I error rates whereas the tests by permutation of residuals were not adversely affected. In combinations involving highly skewed data and extreme values, the two methods of permutation of residuals were less affected than the normal-theory *t*-test or the permutation of the values of **y**.

Practical aspects: Permutation of the values of **y** and permutation of the residuals of the full regression model are appropriate to test the significance of partial regression coefficients because all coefficients can be tested using a single series of permutations. In permutation of the residuals of the full regression model, the permuted elements are the residuals of the regression of **y** on **X**, as mentioned above. A program shortcut, described by Anderson and Legendre (1999), is to regress these permuted residuals directly on **X** to obtain the vector of regression coefficients $\mathbf{b}^*$ under permutation and the associated t_j^* statistics. In the programs mentioned at the end of the Discussion, permutation of the values of **y** is used when there is a single predictor, because permutation of residuals is only potentially useful in the presence of extreme values in a covariable. With two predictors or more, the user is offered the choice between permutation of the values of **y** (method 1) or permutation of the residuals of the full regression model (method 2) for the tests of the partial regression coefficients. The permutational test of R^2 is always done by permutation of the values of **y**. The programs also compute the parametric (normal-theory) tests of significance. Likewise, in the simulations described in the next section, only the permutation of the values of **y** will be used with a single predictor ($m=1$); in that case, the test of the regression coefficient is equivalent to the test of the coefficient of determination (R^2). Both types of permutation tests will be used in tests of partial regression coefficients when $m=2$.

5. Numerical simulations: methods

Simulations were performed to check the type I error and power of the permutational test of significance of the regression coefficients in regression through the origin. A program was written in FORTRAN77 to carry out the simulations. Data were generated according to two different models:

- *Regression model* I, which is assumed for the parametric tests of significance in regression through the origin, requires that the predictors have fixed values. Data were generated that followed this model by selecting fixed values of one ($\mathbf{x}_1$) or two predictors ($\mathbf{x}_1$ and $\mathbf{x}_2$). Replicate values of the response variable **y** were generated using the model $y_i=\beta_0+\beta_1x_{i1}+\beta_2x_{i2}+\varepsilon_i$ where the errors ε_i were drawn at random from one of three distributions, described below. Rejection rates in one-tailed (upper and lower tails) and two-tailed tests were obtained. Permutation tests were carried out using permutation of the values of **y** and permutation of the residuals.
- *Regression model* II, also called the correlation model (Fig. 4), assumes that the predictors as well as the response are random variables. It was important to do the simulation study for regression model II data because independent contrast data belong to that type: error is present in the explanatory as well as the response variables. These data do not strictly verify the conditions of application of the parametric tests of significance. For two predictors and a response variable, three vectors $\mathbf{z}_1$, $\mathbf{z}_2$, and $\mathbf{z}_3$, of length n, were created by random draw from one of the error distributions described below and written to a matrix **Z** of size ($n\times3$). The deterministic components of the model consisted of three correlation coefficients $\rho(\mathbf{z}_1\mathbf{z}_2)$, $\rho(\mathbf{z}_1\mathbf{z}_3)$, and $\rho(\mathbf{z}_2\mathbf{z}_3)$, written into a correlation matrix **R**, which reflected the desired amounts of correlation structuring the statistical population from which the simulated points were drawn. Matrix **R** was decomposed using Cholesky factorization, $\mathbf{R}=\mathbf{L}'\mathbf{L}$, where **L** is a ($3\times3$) upper triangular matrix. Matrix **W** containing the correlated vectors was obtained by computing $\mathbf{W}=\mathbf{ZL}$. Its columns were called **y**, $\mathbf{x}_1$ and $\mathbf{x}_2$ in view of the regression. The rationale for this transformation is the following: if the column vectors forming **Z** are drawn at random from distributions with mean 0 and variance 1, then $[1/(n-1)]\mathbf{Z}'\mathbf{Z}=\mathbf{I}$ (expected value of a correlation matrix among random normal deviates) and the covariance between the columns of **W** reflect the original correlations assigned to matrix **R**. This statement is demonstrated as follows using the elements mentioned above (Legendre, 2000):

$$[1/(n-1)]\mathbf{W}'\mathbf{W}=[1/(n-1)]\mathbf{L}'\mathbf{Z}'\mathbf{Z}\mathbf{L}=\mathbf{L}'\mathbf{I}\mathbf{L}=\mathbf{L}'\mathbf{L}=\mathbf{R} \tag{7}$$

For both data generation models, error values ε_i were drawn at random from three types of distributions:

- Standard normal distribution with $\mu=0$ and $\sigma^2=1$.
- Exponential deviates: 3×10^6 deviates, enough to carry out 10,000 repeated simulations with three variables and 100 observations, were obtained from a standard exponential distribution with $\mu=1$ and $\sigma^2=1$. They were standardized to mean 0 and standard deviation 1.
- Cubed exponential deviates, i.e., standard exponential deviates to the power 3: this distribution was used to examine the behavior of the different types of tests in the presence of highly

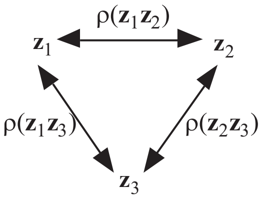

Fig. 4. Correlation model for generation of data.

skewed data, as in the simulations reported by Manly (1997, pp. 163–166), Anderson and Legendre (1999), and Legendre (2000). The 3×10^6 cubed exponential deviates actually had a mean near 6 and a standard deviation near 26. They were standardized to mean 0 and standard deviation 1, and stored in a file before being used in the simulations.

There were 10,000 repeated simulations for each situation; they allowed us to calculate the rejection rate of the null hypothesis for different significance levels α, as well as the 95% confidence interval of the rejection rate. For permutation testing, 999 random permutations were done. Simulation results were reported for the coefficient of determination (R^2) and the regression coefficient of the first explanatory variable ($\mathbf{x}_1$).

5.1. Type I error

Type I error occurs when the null hypothesis is rejected while the data conform to H_0. To be valid, a test of significance should have a rate of rejection of the null hypothesis no larger than the nominal (α) significance level of the test when H_0 is true (Edgington, 1995).

It is not easy to generate random data that conform to the null hypothesis of regression through the origin. In multiple regression, data of that sort are easily produced by generating a response variable **y** that is linearly independent of the explanatory variables **X**; how to generate such data was described, for instance, by Manly (1997) and Anderson and Legendre (1999). Regression through the origin may, however, produce a significant slope for such data unless the values of **y** are centered on the abscissa. Data conforming to the null hypothesis were obtained by setting the parameters of the models as follows:

- *For regression model* I *data*: β_0, β_1, and β_2 were set to 0. Simulations with a single explanatory variable were carried out using $n = \{10, 20, \ldots, 80, 90\}$. Simulations with two explanatory variables were done using $n = \{25, 50, 75, 100\}$. In these simulations, the explanatory variables had fixed values of $\{-1.0, -0.5, 0.0, 0.5, 1.0\}$; hence n had to be multiples of 5 for $m = 1$, or multiples of 25 for $m = 2$. Error was normal, exponential, or cubed exponential.
- *For regression model* II *data*: $\rho(\mathbf{z}_1\mathbf{z}_2)$, $\rho(\mathbf{z}_1\mathbf{z}_3)$, and $\rho(\mathbf{z}_2\mathbf{z}_3)$ were set to 0. Simulations with one and two explanatory variables were carried out using $n = \{10, 20, \ldots, 90, 100\}$.

The rate of rejection of the null hypothesis, after 10,000 repeated simulations, was calculated for tests carried out using significance levels $\alpha = \{0.01, 0.02, 0.03, 0.04, 0.05\}$.

5.2. Effect of an extreme value in $\mathbf{x_2}$

Because previous studies had shown that permutation of the residuals was more appropriate than permutation of the values of **y** for tests of the regression coefficients in the presence of extreme values in the covariable, we had to verify if that conclusion held in the case of regression through the origin. Following Anderson and Legendre (1999), extreme values in the covariable $\mathbf{x}_2$ were generated as follows: the first $(n-1)$ values of $\mathbf{x}_2$ were drawn at random from a uniform distribution on the interval (0, 3) and the nth value was set equal to 55. These simulations represented random variables in which the error in **X** is null or, at any rate, much smaller than the error in **y**.

Simulations with two explanatory variables were carried out for model I data only, for $n = \{5, 10, 25, 50, 100\}$. Using the model $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i$, data conforming to the null hypothesis for β_1 were obtained by setting all parameters β of the model to 0, or by setting $\beta_1 = 0$ and $\beta_2 = \{5, 10, 15, 20\}$ to allow for an effect of the covariable containing the outlier on **y**. Error was either normal or cubed exponential. The simulations were done without and with collinearity among the explanatory variables. Collinearity was introduced by computing $x'_{i1} = x_{i1} + x_{i2}$.

5.3. Power

A test of significance should be able to reject the null hypothesis in most instances when H_0 is false. The ability to reject H_0 in these circumstances is referred to as the power of a test. In the present simulation study, power is defined as the rate of rejection of the null hypothesis when H_0 is false by construct. Power was studied using the same type of simulations as described above, except that this time the alternative hypothesis (H_1) was made to be true. Two types of simulations were done:

- *For regression model* I *data*: β_0 and β_2 were set to 0 whereas β_1 was set to 0.5; the regression coefficient of explanatory variable $\mathbf{x}_1$ was studied. This value of β_1 was selected because it produced rates of rejection of the null hypothesis that were higher than 0 and smaller than 1 in all simulations. Simulations with a single explanatory variable were carried out with $n = \{10, 25, 50, 100\}$, whereas simulations involving two explanatory variables were done using $n = \{25, 50, 75, 100\}$. The explanatory variables had fixed values of $\{-1.0, -0.5, 0.0, 0.5, 1.0\}$; hence n had to be multiples of 5 (for $m = 1$) or 25 (for $m = 2$). Error was either normal or cubed exponential.
- *For regression model* II *data*: The correlations were set in such a way that the partial correlation $\rho(\mathbf{z}_1\mathbf{z}_2.\mathbf{z}_3)$ always had the same value; $\rho(\mathbf{z}_1\mathbf{z}_2.\mathbf{z}_3) = 0.2$ was chosen as an adequate value for reporting the results. $\rho(\mathbf{z}_1\mathbf{z}_3)$ was set to 0 in order to have no effect of $\mathbf{x}_2$ on **y**, and the collinearity $\rho(\mathbf{z}_2\mathbf{z}_3)$ was set to $\{0.0, 0.1, 0.5, 0.9\}$. The value of $\rho(\mathbf{z}_1\mathbf{z}_2)$ allowing us to keep $\rho(\mathbf{z}_1\mathbf{z}_2.\mathbf{z}_3)$ constant is found using the following equation:

$$\rho(\mathbf{z}_1\mathbf{z}_2) = \rho(\mathbf{z}_1\mathbf{z}_2.\mathbf{z}_3)\sqrt{1 - \rho(\mathbf{z}_2\mathbf{z}_3)^2} \tag{8}$$

This equation gives the following pairs of values for $\rho(\mathbf{z}_1\mathbf{z}_2.\mathbf{z}_3) = 0.2$: $\rho(\mathbf{z}_2\mathbf{z}_3) = 0.0$, $\rho(\mathbf{z}_1\mathbf{z}_2) = 0.20000$; $\rho(\mathbf{z}_2\mathbf{z}_3) = 0.1$, $\rho(\mathbf{z}_1\mathbf{z}_2) = 0.19900$; $\rho(\mathbf{z}_2\mathbf{z}_3) = 0.5$, $\rho(\mathbf{z}_1\mathbf{z}_2) = 0.17321$; $\rho(\mathbf{z}_2\mathbf{z}_3) = 0.9$, $\rho(\mathbf{z}_1\mathbf{z}_2) = 0.08718$. Simulations with one and two explanatory variable were carried out using $n = \{10, 50, 100\}$. Error was either normal or cubed exponential. We checked that the values $\rho(\mathbf{z}_1\mathbf{z}_2) = 0.0$, $\rho(\mathbf{z}_1\mathbf{z}_3) = 0.0$, $0.0 \geq \rho(\mathbf{z}_2\mathbf{z}_3) \geq 1.0$ did produce realizations of the null hypothesis, as expected, in regression through the origin.

Additional simulations for power involved data generated without structure, as in the type I error study. An effect was produced by moving the centroid away from the origin. The centroid was located at coordinates (2, 2) or (10, 10). For normal error ($N(0, 1)$), regression through the origin should find the slope of the regression line significant. These simulations go beyond the data configurations expected for independent contrast data. They were carried out to eliminate a procedure which is inadequate in more general cases of regression through the origin; see the last paragraph of the section "Numerical simulations: results".

The rate of rejection of the null hypothesis, after 10,000 repeated simulations, was calculated for tests at significance levels $\alpha = \{0.01, 0.02, 0.03, 0.04, 0.05\}$. Only the results for $\alpha = 0.05$ will be reported in detail.

6. Numerical simulations: results

We will examine the behavior, under simulations, of the method described in the "Permutation test: double-permutation procedure" section. Table 3 summarizes the results. Using normal data, these results show first that the two permutation tests work correctly: the *F*-test of the coefficient of determination and the *t*-test of individual regression coefficients both have correct levels of type I error, and the same power as the parametric form. A test has correct type I error if the rejection rate is approximately equal to the significance level of the test.

6.1. Type I error

With normal error, the parametric and permutation tests had the same behavior (Fig. 5) and were thus equivalent. This was true for data generated under the regression model I (Fig. 5a and c) or II (correlation model: Fig. 5b and d).

With highly asymmetric error (cubed exponential deviates), the permutation tests behaved better than the parametric forms in both the global test of R^2 and the *t*-test of a regression coefficient. To be valid, a test of significance should have a rate of rejection of the null hypothesis no larger than the nominal α

Table 3
Simulation results.

	m	*F*-test of R^2		*t*-test of regression coefficient		
Testing procedure ⇒		Parametric	Permute **y**	Parametric	Permute **y**	Permute res
Type I error (summary of 132 × 10^4 simulations)						
Normal error (model I or II data)	1	OK all *n*	OK all *n*	OK all *n*	OK all *n*	
Exp error						
Model I data	1	$n>10$	$n>40$	$n>10$	$n>40$	
Model II data	1	$n>10$	$n>10$	$n>10$	$n>10$	
Exp^3 error						
Model I data	1	Rate $<\alpha$	OK all *n*	Rate $<\alpha$	OK all *n*	
Model II data	1	Invalid[a]	$n>10$[b]	Invalid[a]	$n>10$[b]	
Normal error (model I or II) Fig. 5	2	OK all *n*	OK all *n*	OK all *n*	OK all *n*	OK all *n*
Exp error						
Model I data	2	OK all *n*	$n>10$[c]	OK all *n*	OK all *n*	OK all *n*
Model II data	2	$n>40$	$n>20$	$n>30$	$n>10$	$n>10$
Exp^3 error						
Model I data Fig. 6a and c	2	Rate $<\alpha$	OK all *n*	Rate $<\alpha$	OK all *n*	OK all *n*[d]
Model II data Fig. 6b and d	2	Invalid	$n>40$	Invalid[e]	$n>20$	$n>40$
Type I error with outlier in covariable (summary of 200 × 10^4 simulations)						
Normal error (model I data)						
No effect of covariable on **y**	2	OK all *n*	OK all *n*	OK all *n*	OK all *n*	OK all *n*[f]
Effect of covariable on **y**	2	([g])	([g])	OK all *n*	Invalid	OK all *n*[f]
Exp^3 error (model I data)						
No effect of covariable on **y**	2	$n>25$[h]	$n>10$[i]	$n>5$[j]	$n>5$[j]	$n>5$[j]
Effect of covariable on **y**	2	([g])	([g])	$n>5$[j]	$n>5$[j]	$n>5$[j]
Power (summary of 82 × 10^4 simulations)						
Normal error (model I or II data)	1	*P*(param)	= *P*(permy)	P(param)	= *P*(permy)	
Exp^3 error						
Model I data	1	*P*(param)	< *P*(permy)[k]	P(param)	< *P*(permy)[k]	
Model II data	1	*P*(param)	= *P*(permy)	P(param)	= *P*(permy)	
Normal error (model I or II) Fig. 7a and c	2	*P*(param)	= *P*(permy)	P(param)	= *P*(permy)	= *P*(perres)
Ex^3 error						
Model I data Fig. 7b and d	2	*P*(param)	< *P*(permy)	P(param)	< *P*(permy)	= *P*(perres)
Model II data	2	Invalid[l]	valid[m]	Invalid[l]	*P*(permy)[n]	= *P*(perres)[n]

"*m*" = number of explanatory variables. "Parametric": parametric *F* or *t*-test. "Permute **y**": test by permutation of the values of **y**. "Permute res": test by permutation of the residuals. "OK all *n*": the test has correct type I error for all sample sizes (*n*) investigated in the simulations. "$n>10$": the test has correct type I error for $n>10$. "rate $<\alpha$": the test is valid but too conservative. "invalid": the test has inflated type I error for all *n* and is thus invalid. Blank: no simulation was done. In the power section, "*P*(param)": power of the parametric test; "*P*(permy)": power of the test by permutation of **y**; "*P*(perres)": power of the test by permutation of the residuals.

[a] Test valid for $n \geq 50$ at $\alpha = 5\%$; test invalid for all *n* at $\alpha = 4$–1%.

[b] Test at $\alpha = 5\%$ valid for $n>10$; at $\alpha = 4\%$, 3% for $n>20$; at $\alpha = 2\%$, 1% for $n>40$.

[c] For the test of R^2, the rejection rate is slightly $>\alpha$ (all α-levels) for $n = 25$; rate slightly $>\alpha$ ($\alpha = 5\%$, 4%) for $n = 50$.

[d] Test valid but slightly conservative for $n = 25$.

[e] Test valid for $n \geq 50$ at $\alpha = 5\%$ only. Test always invalid for the other values of α.

[f] Rejection rate slightly larger than α for $n = 5$.

[g] H_0 is false in these simulations because there is an effect of the covariable on **y**. Hence the type I error rate of the *F*-test cannot be estimated.

[h] Testing at $\alpha = 5\%$: rejection rate $>\alpha$ for $n = 5, 10, 25$; rate $<\alpha$ (test valid) for $n = 50, 100$. Testing at $\alpha = 4\%$ or 3%: rejection rate $<\alpha$ (test valid) for $n = 100$. Test always invalid when testing at $\alpha = 2\%$ or 1%.

[i] Testing at $\alpha = 5\%$ or 4%: rejection rate $>\alpha$ for $n = 5, 10$; rate $<\alpha$ (test valid) for $n = 25, 50, 100$. Testing at $\alpha = 3\%, 2\%, 1\%$: rejection rate $>\alpha$ for $n = 5, 10, 25$; rate $<\alpha$ (test valid) for $n = 50, 100$.

[j] Testing at $\alpha = 5\%$ or 4%: rejection rate $<\alpha$ (test valid) for $n \geq 10$. Testing at $\alpha = 3\%$ or 2%: rejection rate $<\alpha$ (test valid) for $n \geq 25$. Testing at $\alpha = 1\%$: rejection rate $<\alpha$ (test valid) for $n \geq 50$.

[k] The confidence intervals of the rejection rates overlapped partly for $n = 10$ and 25, but not for $n = 50$ and 100 where the power of the permutation test was clearly greater.

[l] There was no point in examining power of the parametric test and comparing it to that of the permutation test since the parametric test is invalid; see section on type I error.

[m] The permutation test of R^2 is valid for $n>40$; see section on type I error.

[n] Power of the test by permutation of the values of **y** is greater than the power of the test by permutation of the residuals for $n = 10$ only.

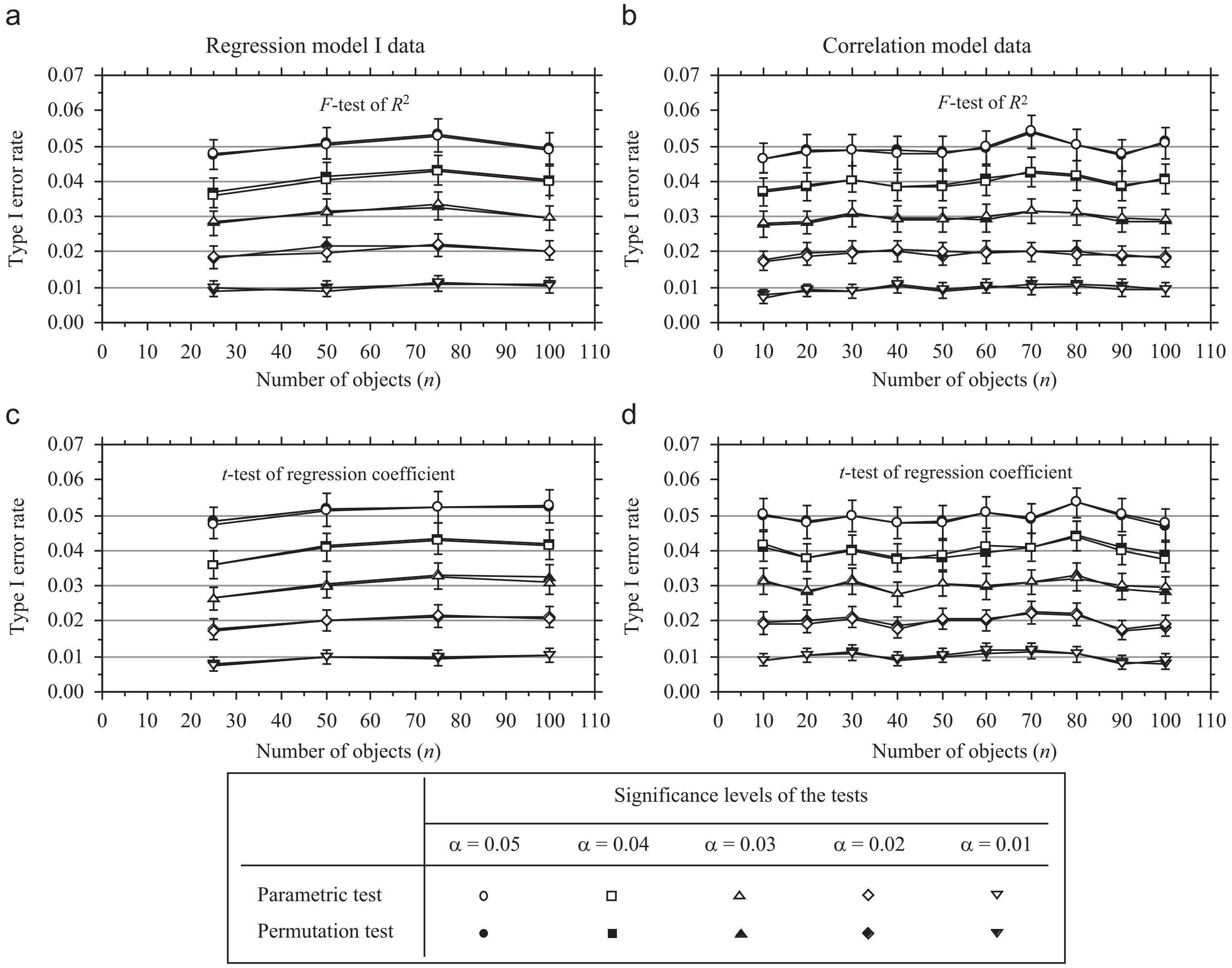

Fig. 5. Mean and 95% confidence intervals of the empirical rates of type I error of the *F*-test of the coefficient of determination (R^2) and the two-tailed *t*-test of the first regression coefficient, for different significance levels ($\alpha = 5\%$, 4%, 3%, 2%, and 1%, materialized by horizontal lines), with increasing sample sizes (*n*). There were two explanatory variables in these simulations. *Left*: data generated under the regression model I. *Right*: data generated under the regression model II (correlation model). The population parameters for the simulations were chosen in such a way that H_0 was true; the error terms were random standard normal deviates. Open symbols: parametric test; black symbols: test by permutation of the values of **y**; black symbols are often hidden by the corresponding open symbols. Overlapping confidence intervals are drawn as their union for clarity.

significance level of the test when H_0 is true (Edgington, 1995). With regression model I data, the parametric tests were too conservative (Fig. 6a and c), the rate of rejection of the null hypothesis being systematically far too low. Tests that are too conservative are not invalid, but conservatism will affect the power of the parametric tests. With regression model II data (correlation model: Fig. 6b and d), the parametric tests were invalid. Thus parametric tests should be avoided with this type of data.

With highly asymmetric data, the rejection rates for parametric and permutational one-tailed *t*-tests of individual regression coefficients in the upper tail (results not shown) were too high, making these tests invalid; tests in the lower tail almost never rejected the null hypothesis. So, with highly asymmetric data, one should avoid one-tailed tests.

Additional simulations for type I error were performed, in which the mean value of the explanatory variables was 10 instead of 0: (a) when the error was normal, all forms of tests had correct type I error; (b) with a single explanatory variables and highly asymmetric error, all forms of tests of the coefficient of determination (R^2) were far too conservative, having rejection rates close to 0. Note that a conservative test is still a valid test. The conservative type I error will simply translate in reduced power when an effect is present in the data. The *t*-tests of the regression coefficients had the same behavior as the tests of R^2 since the two tests are equivalent; and (c) with two explanatory variables and highly asymmetric error, the permutation test of the coefficient of determination (R^2) was too conservative, but not as strongly as the parametric *F*-test; hence the permutation test will have higher power than the parametric test when an effect is present in the data. For the *t*-test of a regression coefficient, both forms of permutational tests had correct type I error whereas the parametric *t*-test had error rates well below α. Again, this will translate in the permutation tests having higher power to detect an effect when present in the data.

6.2. *Type I error: effect of extreme values in the covariable*

For symmetric model I data (normal error), permutation of the values of **y** was invalid when there was an effect of the covariable ($\mathbf{x}_2$) on **y**. The test by permutation of the residuals generally

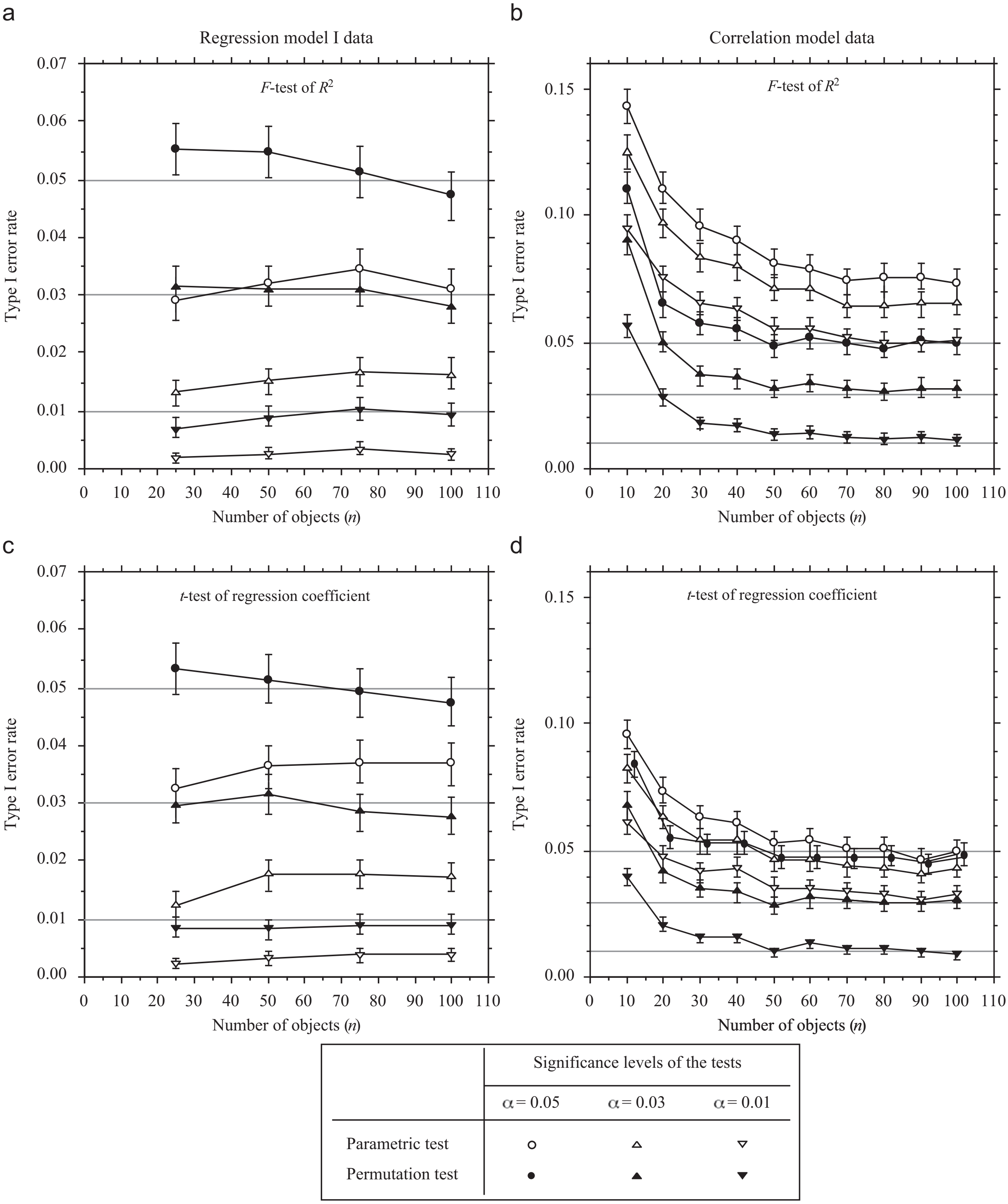

Fig. 6. Mean and 95% confidence intervals of the empirical rates of type I error of the F-test of the coefficient of determination (R^2) and the t-test of the first regression coefficient, for different significance levels ($\alpha = 5\%$, 3%, and 1%, materialized by horizontal lines), with increasing sample sizes. As in Fig. 3, except that the error terms are random standardized cubed exponential deviates. In (d) the black dots and their confidence intervals have been moved sideways for clarity.

performed well, but it never outperformed the parametric t-test which always had correct type I error. When the error was strongly asymmetric, all tests were too conservative and, thus, remained valid. No form of test did better than the other forms. Nearly identical results were obtained with or without collinearity between the two explanatory variables.

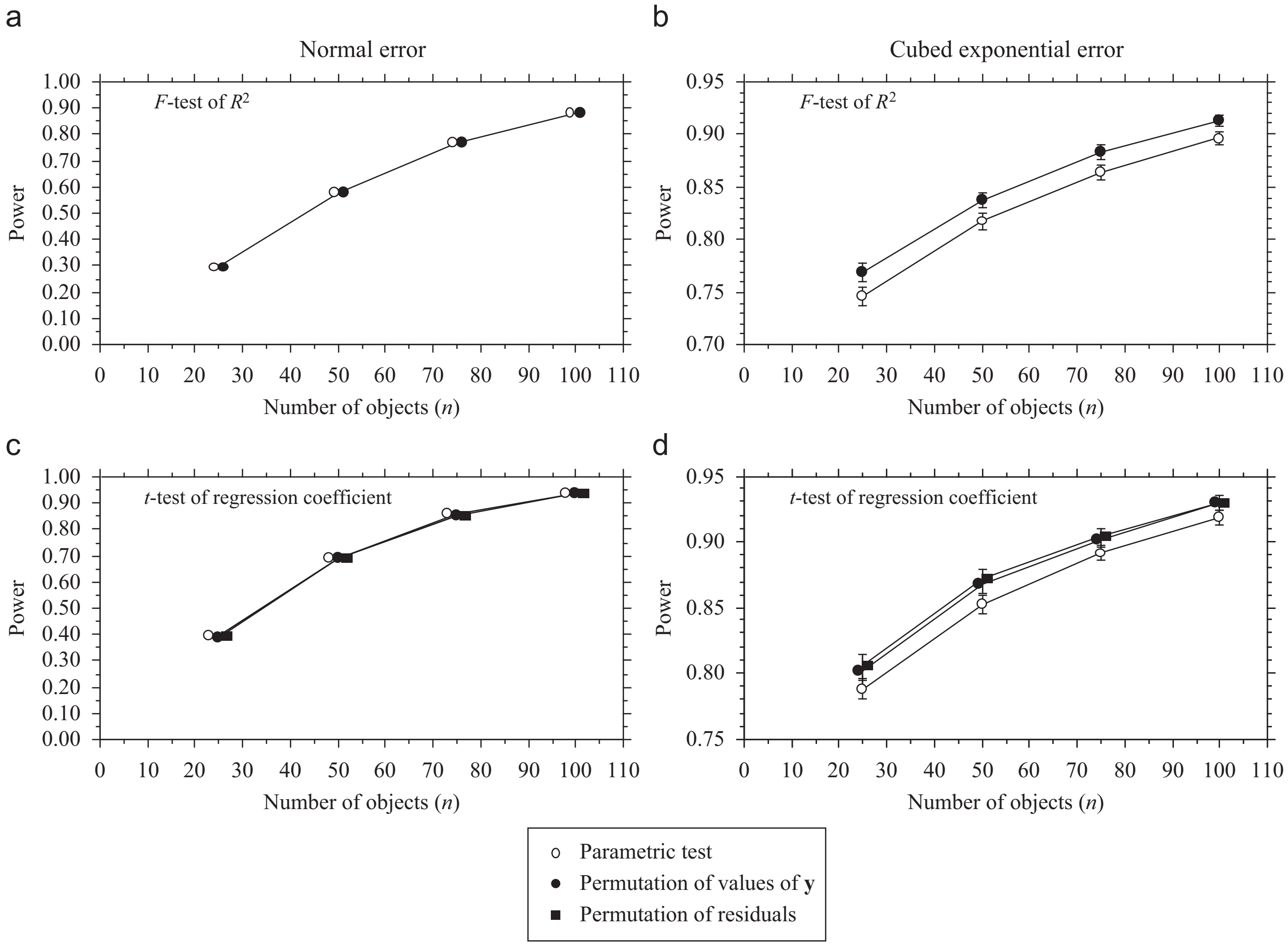

Fig. 7. Mean and 95% confidence intervals of empirical measures of power (at $\alpha = 5\%$) of the F-test of the coefficient of determination (R^2) (a, b) and the t-test of the first regression coefficient (c, d), with increasing sample sizes, for the parametric and permutation tests applied to data generated under the regression model I. Two explanatory variables were used in these simulations. When necessary, symbols have been moved sideways for clarity. With normal error, the confidence intervals are so small that their limits are hidden by the symbols.

Permutation of the residuals had been found by Anderson and Legendre (1999) to be useful for testing a regression coefficient when there were extreme values in one of the covariables; for normal error, the parametric t-test and the test by permutation of the residuals both had correct type I error in their simulations, whereas the test by permutation of the values of **y** had erratic type I error rates, the rate depending of the value of the covariable's parameter; for highly asymmetric data, permutation of the residuals was the only form of test having correct type I error. This appears not to be the case in regression through the origin: for symmetric data (normal error), the parametric t-test maintained correct type I error. For highly asymmetric data, all forms of tests were valid but too conservative for $n \geq 25$. So there is no need to resort to permutation of the residuals, in regression through the origin, in the presence of extreme values in the covariable.

6.3. Power

With normal error, all forms of test had equal power (Fig. 7a and c). With cubed exponential error and regression model I data, the power of the permutation test was higher than that of the parametric test (Fig. 7b) by the same amount (about 2%) as the degree by which the parametric test was too conservative in simulations for type I error (Fig. 6a). With regression model II data (correlation model), there was no point in examining the power of the parametric F-test since it was invalid in all cases (Fig. 6b). The permutation test was valid when $n > 20$ to 40, depending on the severity of asymmetry in the error term.

With cubed exponential error and regression model I data, the power of the two types of permutation tests was higher than that of the parametric test (Fig. 7d) by the same amount as the degree by which the parametric test was too conservative in simulations for type I error (Fig. 6c). With regression model II data (correlation model), there was no point in examining the power of the parametric t-test since it was invalid. The permutation tests were valid when $n > 10$–40, depending on the type of test and the severity of asymmetry in the error term. For regression model I data, the advantage found for the two forms of permutation tests over the parametric t-test, in simulations involving highly skewed error, was similar to that found by Anderson and Legendre (1999) in ordinary multiple regression, for data generated in the same way.

For data with $N(0,1)$ error, when the centroid of the data set was moved into the first quadrant, the permutation test detected a

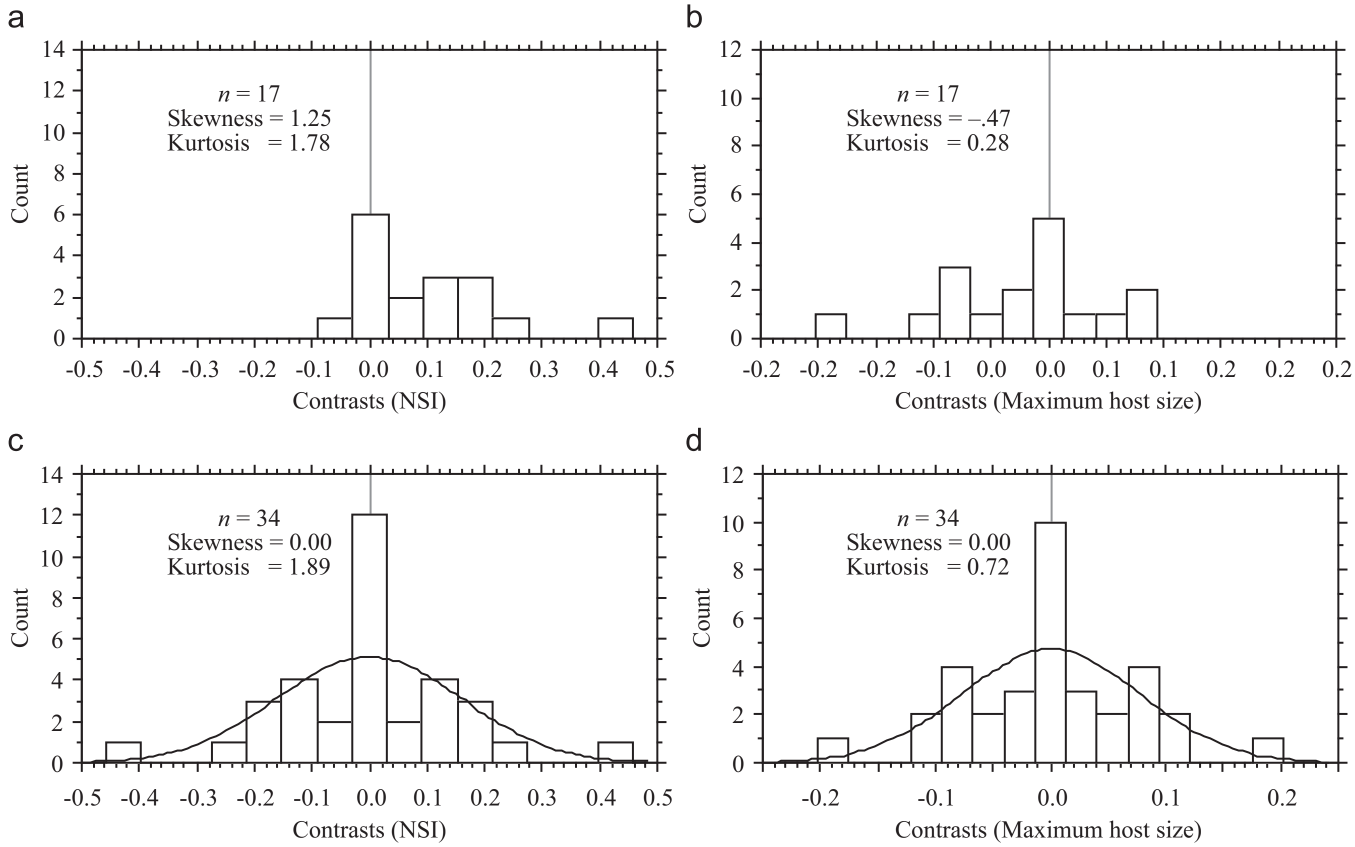

Fig. 8. Frequency histograms of the 17 original contrasts of Example 1: (a) NSI and (b) maximum host size. Histograms of the doubled data sets ($n = 34$): (c) NSI and (d) maximum host size; a normal curve is also shown for comparison.

significant regression-through-the-origin slope with the same power as the parametric test. This was true for models I and II data as well.

6.4. Alternative strategies for permutation tests

Before we imagined the double-permutation procedure described in the "Permutation test: double-permutation procedure" section, we carried out simulations to test the behavior of two simple permutation strategies: (1) in the first strategy, the data were not doubled. The values of **y** were permuted at random with respect to **x** and regression through the origin was computed. Simulations for type I error and power were carried out as described in the previous section, using normal error. The most important effect was that, when the centroid of the data set was moved into the first quadrant, the tests of R^2 and the regression coefficient had no power above the α significance level, whereas the parametric test correctly detected a significant regression-through-the-origin slope in all cases. This was true for models I and II data; (2) in the second strategy, the values of **y** in the doubled data set were permuted at random with respect to **x** and regression through the origin was recomputed. All simulations showed greatly inflated type I error rates for the tests of R^2 and the regression coefficient. These two forms of permutation test are thus incorrect.

7. Example 1, part 2

In the example presented above, we regressed the contrasts computed from a measure of non-specificity (NSI) of a group of parasites on the contrasts computed from the maximum size of their fish host species. The two variables were measured with error, hence their contrasts are also with error; this is thus a case of model II regression. Since the data are of the correlation (or model II) type, we are interested in testing the "correlation through the origin" between NSI and maximum host size; we can proceed using regression through the origin since the test of a simple linear regression coefficient is the same as that of a coefficient of linear correlation. Desdevises et al. (2002b) had hypothesized NSI to be lower for animals that use larger hosts, so the test of significance should be one-tailed in the lower tail.

Frequency histograms are presented in Fig. 8a and b for the two independent contrast variables. From casual examination of the histograms, it is hard to decide what type of error is present in the data. Indeed, a contrast data set is but one of the many realizations that could have been obtained by drawing the phylogenetic tree in different ways, as in Fig. 1; each way of drawing the tree would have led to a different contrast data set and a different histogram. To assess the degree of asymmetry, we will: (1) draw frequency histograms of the doubled data sets (a "doubled data set" has been defined in the Rationale section as one in which the doubled data points have reversed signs) and (2) look at the kurtosis of the distributions: doubled normal contrast data will have kurtosis near zero whereas doubled asymmetric contrast variables will have leptokurtic (i.e., pointed) distributions. The skewness parameter is useless since the frequency distribution of any doubled variable is symmetric by construct (Fig. 8c and d).

The frequency histograms of the two doubled contrast data sets (Fig. 8c and d) display some amount of kurtosis, especially the NSI contrasts, showing that the data are certainly not normal; one

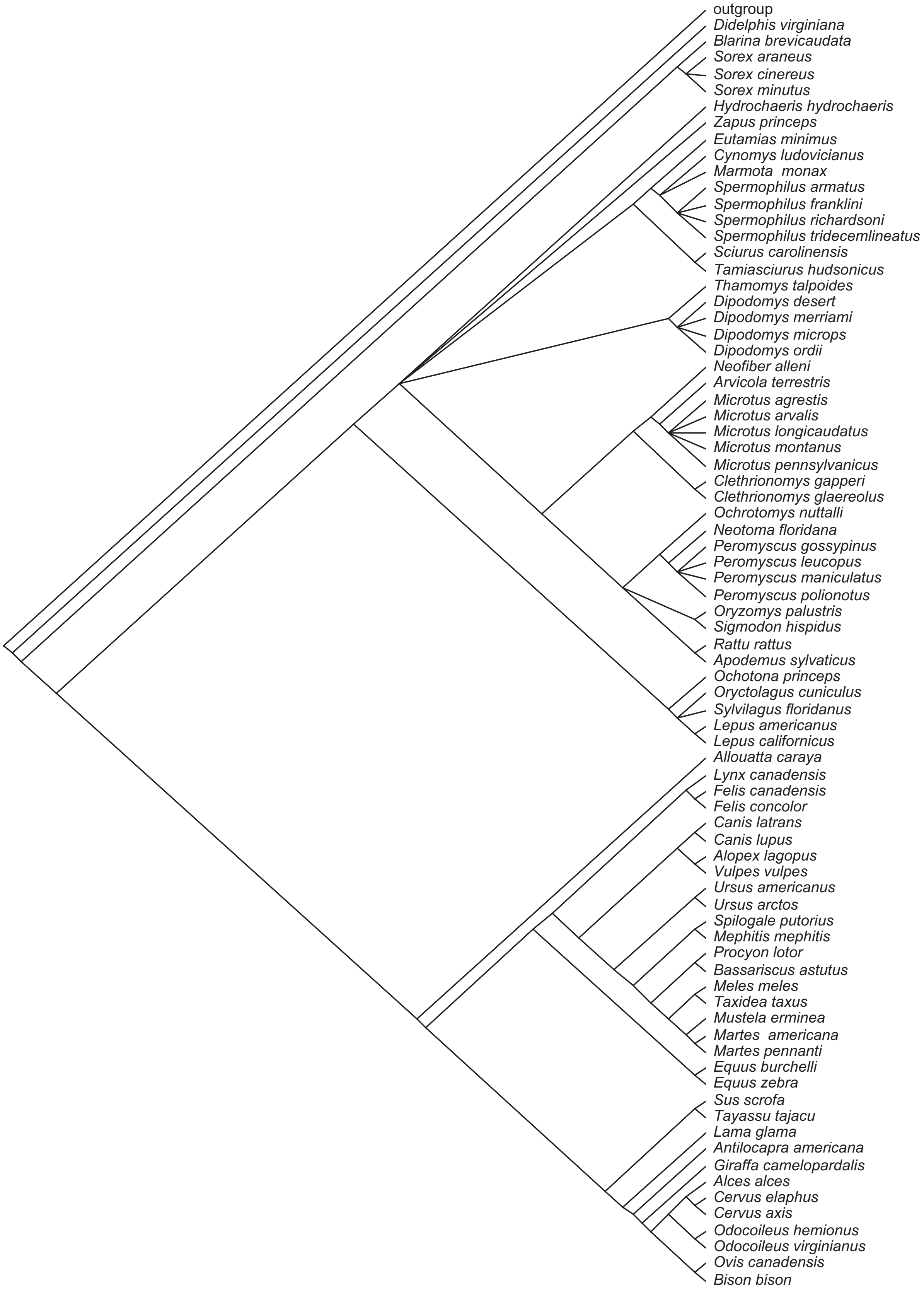

Fig. 9. Phylogenetic trees of the 78 mammal species plus outgroup. From Morand and Poulin (1998).

may feel safer in using the permutation test. Kurtosis is not near what would be expected for highly asymmetric data of the type that were used in the simulations, so that the parametric test can also be used in that case. The two one-tailed tests show the relationship to be negative and highly significant (Table 2).

8. Example 2

Independent contrasts were computed on the phylogeny (Fig. 9) of 78 species of mammals. That tree, derived from various sources, was published by Morand and Poulin (1998). The response variable **y** was richness in parasites (i.e., number of parasite species). The explanatory variables were: $\mathbf{x}_1 =$ spatial density of hosts (i.e., number of hosts per hectare) and $\mathbf{x}_2 =$ average mass of adult hosts (in kg). These data can be obtained at URL: http://www.pubs.roysoc.ac.uk. The 61 contrasts (not shown) were computed using the program CAIC version 2.6.8b (Purvis and Rambaut, 1995). Regression through the origin was used to determine if richness in parasites is related to and can be explained by host density and host body mass. Since we expect larger individuals, as well as those living in denser populations, to harbor and share more species of parasites (Morand and Poulin, 1998), the hypotheses lead to one-tailed tests of significance. Notice that the data are of the correlation (or model II) type. So, we are actually interested in testing the "partial correlations through the origin" of richness with the two explanatory variables; we can proceed using regression through the origin since the test of a partial regression coefficient is the same as that of a partial correlation coefficient.

The contrasts are more asymmetric than in the previous example: kurtoses of the doubled data sets are 3.66 for *Parasite richness*, 10.04 for *Spatial density*, and 9.60 for *Host mass*; they were 1.89 and 0.72 in Example 1. So we feel safer in using permutation tests. The results of regression through the origin are presented in Table 4. We first notice the low explanatory power ($R^2 = 0.04529$) and lack of significance of the regression model. Furthermore, the signs of the regression coefficients are opposite to the predictions of our hypotheses. Using the one-tailed probabilities provided by the program (they are computed in the direction of the signs of the regression coefficients), we can calculate the probabilities under our stated alternative hypotheses: the proportions of permuted values t^* as large as or larger than the observed t-statistics are 0.7415 for *Spatial density* and 0.9413 for *Host mass*. Collinearity between the two explanatory variables is low ($r = -0.03189$), so there is no point in attempting a backward elimination of the least significant variable, *Host mass*, followed by recomputation of the regression for the remaining variable, *Spatial density*. We conclude that our hypotheses are not supported by the data.

9. Discussion

Independent contrasts are computed under the assumption of Brownian motion, but in actual data contrasts may not be normally distributed for a variety of reasons: the physical scale of measurement may not lead to normally distributed data; the type of data and/or the method used to reconstruct the tree may produce biased estimates of the true tree; non-random selection of taxa may lead to asymmetric distributions. It can often be difficult to find a transformation that will effectively normalize the data and prevent extreme contrast values from exerting high leverage in regression models. In such cases, permutation tests may be more appropriate than parametric tests to identify significant relationships between contrast data by regression through the origin. This permutational procedure can also be applied to the extension of the independent contrasts method proposed by Felsenstein (2008) to consider within-species variation.

Examination of the logic underlying regression through the origin led to the formulation of a permutation test for the coefficient of determination and individual regression coefficients in this type of regression. A simulation study was conducted; it led to the following recommendations about the use of the parametric and permutation tests in regression through the origin:

- When the error is normal, the parametric and permutation tests can be used equally well in all situations: with regression model I (no error in the predictors) or regression model II data (error in the predictors; independent contrast data belong to that category), with any number of explanatory variables, and with all sample sizes.
- When the error is highly asymmetric and the predictors are without error (regression model I data), the parametric F-test of the coefficient of determination and the parametric t-test of individual regression coefficients are too conservative, having rejection rates well under the significance level when the null hypothesis is true, whereas the permutation tests have correct type I error. As a consequence, both forms of permutation tests have higher power than the parametric test for detecting an effect, and should thus be preferred.
- When the error is highly asymmetric and the predictors are measured with error (regression model II data, e.g., independent contrasts), the parametric F-test of the coefficient of determination and the parametric t-test of individual regression coefficients are invalid, having inflated type I error rates. Valid permutational two-tailed tests can be performed, whereas parametric tests should be avoided. No test is valid for very small sample sizes ($n \leq 10$ or 40, depending on the type of test and the degree of asymmetry of the error).
- In the presence of extreme values in the covariable, permutation of the values of **y** had inflated type I error rates for normal data when the covariable had an effect on **y**. Permutation of the residuals had correct type I error in most situations, but it did not outperform the parametric t-test. For highly asymmetric error, all tests remained valid but were too conservative. The best overall solution is thus to use the parametric t-test in the presence of extreme values.
- Except for extreme values in the covariable, permutation of the values of **y** can be used safely in all situations. It always has correct type I error and the same power as the parametric tests when the error is normal or moderately asymmetric, and it outperforms the parametric tests when the error is highly asymmetric.
- There is no situation where permutation of the residuals outperformed both the parametric t-test and the test by

Table 4
Regression through the origin for Example 2, with parametric and permutation tests (9999 permutations of the values of $\mathbf{y} =$ parasite richness).

Explanatory variables	Regression coefficients (b)	t	P (permutational)[a]	P (parametric)[a]
Spatial density	−0.00057	−0.61734	0.2585	0.26969
Host mass	−0.01835	−1.55731	0.0587	0.06237

$R^2 = 0.04529$, P (parametric) $= 0.25482$, P (permutational) $= 0.2402$.

[a] One-tailed tests in the direction of the sign of b.

permutation of values of **y**. Since this permutational method requires more computing time than permutation of the values of **y**, it is not necessary to include it in computer programs.

To summarize, the parametric tests or any of the permutation tests can be used to test the significance of the coefficient of determination (R^2) or individual regression coefficients when the error is normal. Only the test by permutation of the values of **y** can be used when the error is highly asymmetric. The parametric tests should be used in the presence of extreme values in the covariables. Asymmetric independent contrast data can be detected by examining the frequency histograms of doubled data sets: normal data have kurtosis near 0 whereas asymmetric data are leptokurtic (kurtosis >0).

Examples were presented where contrasts were highly non-normal; this was found by examining the kurtoses of the distributions of double contrast data tables. It is likely that many actual data sets analyzed by the method of independent contrasts violate one or several of the distributional assumptions of the parametric tests used in regression through the origin.

A Fortran program (Regression_test: source code, compiled versions for Macintosh and DOS, and program documentation) and the R-language library 'lmorigin' are available from the web page 〈http://www.bio.umontreal.ca/legendre/indexEn.html〉 to carry out multiple linear regression through the origin with parametric and permutation tests.

Acknowledgments

We are grateful to two anonymous reviewers for useful comments on the manuscript. This research was supported by NSERC Grant no. OGP0007738 to P. Legendre.

Appendix A

The computational particularities of regression through the origin are (Kvålseth, 1985; Neter et al., 1996):

(1) For matrix calculation, matrix **X** only contains the m explanatory variables. No column of 1's is added to **X** to estimate the intercept. The number of parameters estimated during the regression is thus m. Vector **b** containing the m partial regression coefficients is estimated in the usual way: $\mathbf{b} = [\mathbf{X}'\mathbf{X}]^{-1}\mathbf{X}'\mathbf{y}$ where **y** is the dependent variable.

(2) The t_j-statistic associated with each partial regression coefficient estimate, b_j, is calculated as usual: $t_j = b_j/$(standard error of b_j). The t_j-statistic is tested for significance with $(n-m)$ degrees of freedom instead of $(n-m-1)$. Table 2 shows the calculations for Example 1, which has $m = 1$ explanatory variable.

(3) The coefficient of determination is calculated as follows: $R^2 = \sum(\text{fitted values})^2/\sum(y^2)$. This formula produces the exact same value for R^2 as ordinary OLS regression on the doubled data set. An alternative formula for R^2 , not used in this paper, is $R^2 = 1 - \sum e_i^2/\sum(y_i - \bar{y})^2$; that formula may produce negative values for R^2 in regression through the origin.

(4) The F-statistic associated with R^2 is calculated as $F = (R^2/m)/[(1-R^2)/(n-m)]$ and is tested with $\nu_1 = m$ and $\nu_2 = (n-m)$ degrees of freedom. For $m = 1$, the value of the F-statistic of regression through the origin is exactly twice the value of the F-statistic of ordinary OLS regression on the doubled data set.

(5) The adjusted coefficient of determination is computed using the formula: $R^2_{adj} = 1 - (1 - R^2)(n/(n - m))$.

References

Ackerly, D.D., Donoghue, M.J., 1998. Leaf size, sapling allometry, and Corner's rules: phylogeny and correlated evolution in maples (*Acer*). Am. Nat. 152, 767–791.

Anderson, M.J., Legendre, P., 1999. An empirical comparison of permutation methods for tests of partial regression coefficients in a linear model. J. Stat. Comput. Simul. 62, 271–303.

Cheverud, J.M., Dow, M., Leutenegger, W., 1985. The quantitative assessment of phylogenetic constraints in comparative analyses: sexual dimorphism in body weight among primates. Evolution 39, 1335–1351.

Desdevises, Y., Morand, S., Jousson, O., Legendre, P., 2002a. Coevolution between *Lamellodiscus* (Monogenea: Diplectanidae) and Sparidae (Teleostei): the study of a complex host–parasite system. Evolution 56, 2459–2471.

Desdevises, Y., Morand, S., Legendre, P., 2002b. Evolution and determinants of host specificity in the genus *Lamellodiscus* (Monogenea). Biol. J. Linn. Soc. 77, 431–443.

Diaz-Uriarte, R., Garland Jr., T., 1996. Testing hypotheses of correlated evolution using phylogenetically independent contrasts: sensitivity to deviations from Brownian motion. Evolution 52, 1247–1262.

Diaz-Uriarte, R., Garland Jr., T., 1998. Effect of branch length errors on the performance of phylogenetically independent contrasts. Syst. Biol. 47, 654–672.

Diniz-Filho, J.A.F., de Sant'Ana, C.E.R., Bini, L.M., 1998. An eigenvector method for estimating phylogenetic inertia. Evolution 52, 1247–1262.

Edgington, E.S., 1995. Randomization Tests, third ed. Marcel Dekker, New York.

Feliu, C., Renaud, F., Catzeflis, F., Hugot, J.-P., Durand, P., Morand, S., 1997. A comparative analysis of species richness of Iberian rodents. Parasitology 115, 453–466.

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

Felsenstein, J., 1988. Phylogenies and quantitative characters. Annu. Rev. Ecol. Syst. 19, 445–471.

Felsenstein, J., 2008. Comparative methods with sampling error and within-species variation: contrasts revisited and revised. Am. Nat. 171, 713–725.

Fjerdingstad, E.J., Crozier, R.H., 2006. The evolution of worker caste diversity in social insects. Am. Nat. 167, 390–400.

Garland Jr., T., Harvey, P.H., Ives, A.R., 1992. Procedures for the analysis of comparative data using phylogenetically independent contrasts. Syst. Biol. 41, 18–32.

Grafen, A., 1989. The phylogenetic regression. Proc. Roy. Soc. London Ser. B Biol. B 205, 581–598.

Hansen, T.F., Martins, E.P., 1996. Translating between microevolutionary process and macroevolutionary patterns: the correlation structure of interspecific data. Evolution 50, 1404–1417.

Harvey, P.H., Pagel, M.D., 1991. The comparative method in evolutionary biology. Oxford University Press, Oxford.

Hope, A.C.A., 1968. A simplified Monte Carlo test procedure. J. Roy. Stat. Soc. Ser. B 50, 35–45.

Houseworth, E.A., Martins, E.P., Lynch, M., 2004. The phylogenetic mixed model. Am. Nat. 163, 84–96.

Kohlsdorf, T., Grizante, M.B., Navas, C.A., Herrel, A., 2008. Head shape evolution in Tropidurinae lizards: does locomotion constrain diet? J. Evol. Biol. 21, 781–790.

Kolm, N., Stein, R.W., Mooers, A.Ø., Verspoor, J.J., Cunningham, E.J., 2007. Can sexual selection drive female life histories? A comparative study on Galliform birds. J. Evol. Biol. 20, 627–638.

Kvålseth, T.O., 1985. Cautionary note about R^2. Am. Nat. 39, 279–285.

Laurin, M., 2004. The evolution of body size, Cope's rule and the origin of amniotes. Syst. Biol. 53, 594–622.

Legendre, P., 2000. Comparison of permutation methods for the partial correlation and partial Mantel tests. J. Stat. Comput. Simul. 67, 37–73.

Legendre, P., Legendre, L., 1998. Numerical Ecology, second English ed. Elsevier Science BV, Amsterdam.

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

Manly, B.J.F., 1997. Randomization, Bootstrap and Monte Carlo Methods in Biology, second ed. Chapman & Hall, London.

Martins, E.P., Diniz-Filho, J.A.F., Houseworth, E.A., 2002. Adaptive constraints and the phylogenetic comparative method: a computer simulation test. Evolution 56, 1–13.

Morand, S., Harvey, P.H., 2000. Mammalian metabolism, longevity and parasite species richness. Proc. Roy. Soc. London Ser. B Biol. B 267, 1999–2003.

Morand, S., Poulin, R., 1998. Density, body mass and parasite species richness of terrestrial mammals. Evol. Ecol. Res. 12, 717–727.

Neter, J., Kutner, M.H., Nachtsheim, C.J., Wasserman, W., 1996. Applied Linear Statistical Models, fourth ed. Irwin, Chicago.

Poorter, L., Wright, S.J., Paz, H., Ackerly, D.D., Condit, R., Ibarra-Manríquez, G., Harms, K.E., Licona, J.C., Martínez-Ramos, M., Mazer, S.J., Muller-Landau, H.C., Peña-Claros, M., Webb, C.O., Wright, I.J., 2008. Are functional traits good predictors of demographic rates? Evidence from five neotropical forests. Ecology 89, 1908–1920.

Pouydebat, E., Laurin, M., Gorce, P., Bels, V., 2008. Evolution of grasping among anthropoids. Journal of Evolutionary Biology 21, 1732–1743.

Purvis, A., Rambaut, A., 1995. Comparative analysis by independent contrasts (CAIC): an Apple Macintosh application for analysing comparative data. Comput. Appl. Biosci. 11, 247–251.

Stearns, S.C., 1983. The influence of size and phylogeny on patterns of covariation among life-history traits in mammals. Oikos 41, 173–187.

ter Braak, C.J.F., 1992. Permutation versus bootstrap significance tests in multiple regression and ANOVA. In: Rothe, K.-H.G., Sendler, W. (Eds.), Bootstrap and Related Techniques. Springer, Berlin, pp. 79–86.

ter Braak, C.J.F., Smilauer, P., 1998. CANOCO Reference Manual and User's Guide to Canoco for Windows—Software for Canonical Community Ordination (Version 4). Microcomputer Power, Ithaca.

Welch, W.J., 1990. Construction of permutation tests. J. Am. Stat. Assoc. 85, 693–698.

Xiang, Q.Y., Thorne, J.L., Seo, T.K., Zhang, W., Thomas, D.T., Ricklefs, R.E., 2008. Rates of nucleotide substitution in Cornaceae (Cornales)—pattern of variation and underlying causal factors. Mol. Phylogenet. Evol. 49, 327–342.