J. Statist. Comput. Simul., 1999, Vol. 62, pp. 271–303
Reprints available directly from the publisher
Photocopying permitted by license only

AN EMPIRICAL COMPARISON OF PERMUTATION METHODS FOR TESTS OF PARTIAL REGRESSION COEFFICIENTS IN A LINEAR MODEL

MARTI J. ANDERSON[a,]* and PIERRE LEGENDRE[b,†]

[a] *Centre for Research on Ecological Impacts of Coastal Cities and School of Biological Sciences, Marine Ecology Laboratories, A11 University of Sydney, Sydney, NSW 2006, Australia;*
[b] *Département de sciences biologiques, Université de Montréal, C.P. 6128, succ. Centre-ville, Montréal, Québec, Canada H3C 3J7*

(Received 24 October 1998; In final form 25 September 1998)

This study compared empirical type I error and power of different permutation techniques for the test of significance of a single partial regression coefficient in a multiple regression model, using simulations. The methods compared were permutation of raw data values, two alternative methods proposed for permutation of residuals under the reduced model, and permutation of residuals under the full model. The normal-theory *t*-test was also included in simulations. We investigated effects of (1) the sample size, (2) the degree of collinearity between the predictor variables, (3) the size of the covariable's parameter, (4) the distribution of the added random error and (5) the presence of an outlier in the covariable on these methods. We found that two methods that had been identified as equivalent formulations of permutation under the reduced model were actually quite different. One of these methods resulted in consistently inflated type I error. In addition, when the covariable contained an extreme outlier, permutation of raw data resulted in unstable (often inflated) type I error. There were no significant differences in power among the three permutation methods (raw data permutation, reduced-model permutation and full-model permutation), but all had greater power than the normal-theory *t*-test when errors were non-normal. The reduced model permutation method had the most consistent and reliable results of the methods investigated here for the test of a partial regression coefficient. However, reasonably extreme situations needed to be simulated in order to distinguish methods from the normal-theory *t*-test and from one another. Permutation of raw data, permutation under the reduced model, and permutation under the full model are generally asymptotically equivalent.

*Corresponding author. Tel.: (61) (2) 9351-4931, Fax: (61) (2) 9351-6713, e-mail: mjanders@bio.usyd.edu.au

†Tel.: (514) 343-7591, Fax: (514) 343-2293, e-mail: legendre@ere.umontreal.ca

Keywords: Permutation of residuals; multiple regression; linear model; type I error; power; partial regression; randomization

1. INTRODUCTION

Various permutational strategies have been proposed for a test of significance of a single predictor variable in multiple regression. It is well known that such a test may be influenced by the presence of one or more other variables in the linear model. This is true in the case of continuous or discrete predictor variables, as well as when the predictor variables represent orthogonal treatments in a linear analysis-of-variance model. The proposed permutation methods for such tests have different bases in terms of their philosophies and have been proposed in different contexts, (*e.g.*, Freedman and Lane, 1983; Smouse *et al.*, 1986; Oja, 1987; ter Braak, 1992; Kennedy, 1995; Manly, 1997). Properties of some of these approaches have been reviewed in part by Kennedy (1995) and Kennedy and Cade (1996). Expected and desirable qualities of the various techniques with regard to type I error and power have been suggested from a theoretical perspective, but few if any empirical simulations supporting these claims have been provided.

An important application of such permutation tests is their use in canonical analysis of multivariate data (*e.g.*, in ecological, biological and agricultural applications: ter Braak, 1987, 1990; Legendre and Legendre, 1998), where the data generally do not fulfill assumptions required by traditional parametric testing procedures. The present study only examines the behavior of the methods with respect to a single dependent variable, but the results will apply equally to multivariate situations. Any method found to be inappropriate in this simple case, however, will also be inappropriate for the analogous permutation tests for multivariate responses.

We present results of simulations designed to compare empirically certain methods of permutation under conditions of known changes to particular factors. Consider the following linear equation for multiple regression:

$$Y = \mu + \beta_{1\cdot2}X + \beta_{2\cdot1}Z + \varepsilon \qquad (1)$$

where Y is the response variable, X and Z are each predictor variables, $\beta_{1\cdot2}$ and $\beta_{2\cdot1}$ are the partial regression coefficients of the least-squares multiple regression of Y on X and Z, respectively, and ε is the added random error term. The notation $\beta_{2\cdot1}$ is used to indicate that this is the partial regression coefficient for the relationship between Y and variable 2 (Z) while controlling for the effect of variable 1 (X), similarly for $\beta_{1\cdot2}$. The null hypothesis of interest is: $\beta_{2\cdot1} = 0$, that is, there is no significant effect of variable Z in the multiple linear regression. The familiar t-statistic for this null hypothesis takes the following form:

$$t = \frac{(b_{2\cdot1} - 0)}{\mathrm{se}(b_{2\cdot1})} \qquad (2)$$

where $b_{2\cdot1}$ is the least-squares estimate of $\beta_{2\cdot1}$ and $\mathrm{se}(b_{2\cdot1})$ is the estimated standard error of the partial regression coefficient.

We investigated type I error and power of different permutational procedures with regard to changes in the following factors: (a) sample size; (b) degree of collinearity between the predictor variables X and Z; (c) size of the covariable's parameter $\beta_{1\cdot2}$; and (d) distribution of the added random error ε. Each of these factors has been discussed to some extent by proponents of the various permutation methods, without having been investigated empirically to any great length or detail. Our simulations are much more extensive than any presented so far for comparative analysis of these permutation methods (*e.g.*, Kennedy and Cade, 1996; Manly, 1997).

We restricted our attention to methods which do not ignore a potential relationship (collinearity) between the predictor variables (*i.e.*, methods which do not disobey the principle of ancillarity, which means *relatedness*: Welch, 1990; ter Braak, 1992). This limits the discussion to methods which permute either the raw data values (Y) or residuals of some kind, as opposed to permuting predictor variable(s) (such as described by Oja, 1987). Neither do we consider methods involving restricted randomization, which will generally be applicable only to cases involving several replicates at each individual set of paired values of X and Z (*e.g.*, Brown and Maritz, 1982).

We begin with a description of the permutation methods to be investigated, along with the predictions offered in the literature with regard to their relative type I error or power. We then present the methodology used and the results of simulations designed to test these specific predictions, with a view to understanding the general behavior of these permutational strategies.

2. DESCRIPTION OF THE PERMUTATION METHODS

2.1. Permutation of Raw Data

Permutation of raw data for multiple regression was described by Manly (1991, 1997). The following procedure is employed to test the null hypothesis $\beta_{2\cdot 1} = 0$.

1. The variable Y is regressed on X and Z together (using least squares) to obtain an estimate $b_{2\cdot 1}$ of $\beta_{2\cdot 1}$ and a value of the usual t-statistic, t_{ref}, for testing $\beta_{2\cdot 1} = 0$ for the real data. We hereafter refer to this as the reference value of t.
2. The Y values are permuted randomly to obtain permuted values Y^*.
3. The Y^* values are regressed on X and Z (unpermuted) together to obtain an estimate $b^*_{2\cdot 1}$ of $\beta_{2\cdot 1}$ and a value t^* for the permuted data.
4. Steps 2–3 are repeated a large number of times, yielding a distribution of values of t^* under permutation.
5. The absolute value of the reference value t_{ref} is placed in the distribution of absolute values of t^* obtained under permutation (for a two-tailed t-test). The probability is calculated as the proportion of values in this distribution greater than or equal, in absolute value, to the absolute value of t_{ref} (Hope, 1968).

Permutation of the raw data preserves the covariance between Z and X as well as covariances among the X variables, if there are more than one, across all permutations.

The test of significance of a partial regression coefficient is the same as that of the corresponding partial correlation coefficient. Thus, as an alternative to calculating the estimate for the slope parameter and its standard error to obtain the t-statistic, in the case of a single covariable one can calculate the partial regression coefficient easily using the well-known formula: $r_{YZ.X} = (r_{YZ} - r_{YX}r_{ZX})/$

$\sqrt{(1 - r_{YX}^2)(1 - r_{ZX}^2)}$; the t-statistic is then $t = r_{YZ.X}\sqrt{(n-3)}/\sqrt{1 - r_{YZ.X}^2}$.

Manly (1991) initially introduced this permutation method using a non-pivotal test-statistic for the test. The importance of using a pivotal (or asymptotically pivotal) test-statistic, such as t, as opposed to derivatives (such as sums of squares), has been shown for permutation tests of single terms in complex models (Stapel and ter Braak, 1994; Kennedy and Cade, 1995) and for the related technique of bootstrapping (Hall and Titterington, 1989; Fisher and Hall, 1990). By a pivotal statistic, we mean the following: in general, to find a confidence interval for a parameter θ, based on an estimator $T(\underline{x})$, where $\underline{x}$ denotes a vector of n independent random variables, a pivotal statistic τ has the following properties: (i) τ is a function of $T(\underline{x})$ and θ, (ii) τ is monotonic in θ, and (iii) τ has a known sampling distribution that does not depend on θ or on any other unknown parameters. In the present context, $\theta = \beta_{2\cdot 1}$, $T(\underline{x}) = b_{2\cdot 1}$ and $\tau = t$ of Eq. (2) (or alternatively $\tau = r_{YZ.X}^2$, the squared partial correlation coefficient, also fulfills these requirements), where the "known" sampling distribution is created by permutation. Note that $b_{2\cdot 1}$ is not pivotal because it does not fulfill item (iii) above, but depends on the value of $\beta_{2\cdot 1}$. In all of the simulations done in this paper, we only consider permutation tests carried out using the t-statistic, shown in Eq. (2), as in Manly (1997).

The rationale for this method is that the permutable units for the test are the original Y values, independent of any model (linear or otherwise) which might be imposed: that is, any value of Y could have been observed associated with any combination of paired values (X, Z). Thus, the error associated with each Y value "travels with it" because it is considered to be a part of that individual replicate. This view is expressed clearly in the work of Edgington (1995) concerning randomization tests and experimental design. Edgington states that tests involving the implicit assumption of "unit additivity" (as is the case for permutation of residuals from a linear model, see Sections 2.2 and 2.3 below) should not be regarded as distribution-free or non-parametric tests (p. 122). In the context of analysis of variance or with other models involving categorical variables, restricted permutations may be applied as a means of controlling for one factor (X) while testing for the effects of another (Z) (*e.g.*, Brown and Maritz, 1982; Edgington, 1995).

Kennedy (1995) and Kennedy and Cade (1996) stated that the method of permuting Y is only justified when the covariable's parameter $\beta_{1 \cdot 2}$ is zero. Their argument is essentially that the permutation of raw data ignores the covariable's parameter, which often may not be justified. In particular, Kennedy and Cade (1996) suggested that the method of permuting raw data for a test of $\beta_{2 \cdot 1}$ would give biased results if the errors ε and the Y values had radically different distributions in the presence of a non-zero $\beta_{1 \cdot 2}$. In limited simulations, they found that permutation of the raw data (Y) in multiple regression resulted in inflated type I error when outliers were included in X and $\beta_{1 \cdot 2} \neq 0$ (Kennedy and Cade, 1996).

The results of Kennedy and Cade (1996) were not supported by further simulations published by Manly (1997). Although Manly (1997) suggested that a more extensive set of simulations was needed on the topic, his results seemed to show that the method of permuting Y for tests of partial regression coefficients was not necessarily flawed in the way that Kennedy and Cade (1996) had claimed.

We did empirical simulations to test whether permutation of raw data would have biased type I error compared with the other methods (described below), particularly in the presence of a non-zero value for the covariable's parameter. We also examined in more depth the idea of Kennedy and Cade (1996), that a problem with this method would be particularly emphasized if X contained an outlier. In an attempt to resolve this issue, we replicated and expanded the simulation studies done by them and by Manly (1997).

2.2. Permutation of Residuals under the Reduced Model

In contrast to Manly's method of permuting raw data and to the view of Edgington (1995), concerning randomization methods, are those techniques which use the residuals of a linear (or other) model as the permutable units for the test. This approach can generally be referred to as model-based permutation. Here, the stochastic element is considered to be the error ε', disassociated from each particular value of Y by the application of a model to produce residuals, as opposed to the original Y values themselves (*e.g.*, Kempthorne, 1952).

Consider a model where the null hypothesis $\beta_{2 \cdot 1} = 0$ is true, which we can call the "reduced model", as follows:

$$Y = \beta_0 + \beta_1 X + \varepsilon' \tag{3}$$

The rationale for permutation of residuals under the reduced model is that, given some estimate of the relationship between Y and X (even if it is zero), there is no further variation in Y which can be explained by Z. There are two different methods of permutation which have been described to provide a test having this rationale. First, there is the approach of Freedman and Lane (1983); second, there is the approach outlined by Smouse *et al.* (1986) in the context of comparisons among multivariate distance matrices and articulated for ordinary multiple regression by Kennedy (1995).

Freedman and Lane (1983) proposed the following permutational procedure:

1. The variable Y is regressed on X and Z together to obtain an estimate $b_{2\cdot 1}$ of $\beta_{2\cdot 1}$ and a reference value t_{ref} for the real data.
2. The variable Y is regressed on X alone according to the model in Eq. (3), providing estimates b_0 of β_0, b_1 of β_1 and residuals $R_{Y|X}$.
3. The residuals from the regression in 2 are permuted randomly, producing $R^*_{Y|X}$.
4. New values for Y^* are calculated by adding the permuted residuals to the fitted values as follows: $Y^* = b_0 + b_1 X + R^*_{Y|X}$.
5. Y^* are regressed on X and Z together according to the model $E(Y^*) = \beta^*_0 + \beta^*_{1\cdot 2} X + \beta^*_{2\cdot 1} Z$ to obtain an estimate $b^*_{2\cdot 1}$ of $\beta^*_{2\cdot 1}$ and a value t^* for the permuted data. Here, $t^* = b^*_{2\cdot 1}/\text{se}(b^*_{2\cdot 1})$.
6. Steps 3 – 5 are repeated a large number of times, yielding a distribution of values of t^* under permutation.
7. The absolute value of the reference value t_{ref} is placed in the distribution of absolute values of t^* obtained under permutation (for a two-tailed t-test). The probability is calculated as the proportion of values in this distribution greater than or equal, in absolute value, to the absolute value of t_{ref}. Permutation of the residuals under the reduced model preserves across all permutations the covariances between Y and X, Z and X, and among the X variables if there are more than one – but not between Y and Z.

Although Freedman and Lane (1983) wrote that their method was a "nonstochastic" approach, referring to the proportion of the values $t^* \geq t_{\text{ref}}$ as a "descriptive statistic" instead of a probability, the rationale for the test is effectively that of a model-based approach. The goal is to isolate the test of Y on Z alone, while taking X into account through the use of the linear regression equation and permutation of residuals.

We note that equivalent regression models under permutation for Step 5 above are $E(Y^*) = \beta_0^* + \beta_{1\cdot2}^* X + \beta_{2\cdot1}^* R_{Z|X}$, or $E(R^*_{Y|X}) = \beta_0^* + \beta_{1\cdot2}^* X + \beta_{2\cdot1}^* Z$, or $E(R^*_{Y|X}) = \beta_0^* + \beta_{1\cdot2}^* X + \beta_{2\cdot1}^* R_{Z|X}$, with $R_{Z|X}$ defined as in the Kennedy method (below). The values of $b_{2\cdot1}^*$ and t^* obtained under permutation are exactly the same for these models as for the model in Step 5 above.

Freedman and Lane (1983) emphasized two conditions for the use of their method of permutation: that the data should not contain extreme outliers and that X and Z should not be highly collinear. Also, the sample size, n, should be relatively large. Due to the permutation of residuals, the test is not exact in the randomization sense (*e.g.*, Edgington, 1995), but has asymptotically exact significance levels.

Kennedy (1995) presented a method of permutation which he stated (on p. 90) was identical to the Freedman and Lane (1983) method. The rationale for Kennedy's method is the same as that for the Freedman and Lane approach, but computationally, it proceeds by the following steps:

1–3. The same first three steps are done as in the Freedman and Lane method.

4. The variable Z is regressed on X alone according to the model $E(Z) = \lambda_0 + \lambda_1 X$, providing residuals $R_{Z|X}$.
5. The permuted residuals $R^*_{Y|X}$ from Step 3 are regressed on $R_{Z|X}$ according to the model $E(R^*_{Y|X}) = \beta_0^* + \beta_2^* R_{Z|X}$ to obtain an estimate b_2^* of β_2^* and a value t^* for the permuted data. Here, $t^* = b_2^*/\mathrm{se}(b_2^*)$ and t^* is calculated with $(n-3)$ degrees of freedom, as in the Freedman and Lane method.
6. Step 3 followed by Step 5 is repeated a large number of times, yielding a distribution of values of t^* under permutation. (The residuals $R_{Z|X}$ calculated in Step 4 are not recalculated with each permutation; they remain constant).
7. The absolute value of the reference value t_{ref} is placed in the distribution of absolute values of t^* obtained under permutation (for a two-tailed t-test). The probability is calculated as the proportion of values greater than or equal, in absolute value, to the absolute value of t_{ref}.

An equivalent regression model for calculating the reference value for the t-statistic is $E(R_{Y|X}) = \beta_0 + \beta_{2 \cdot 1} R_{Z|X}$. This alternative formula is a standard computational method for obtaining partial regression coefficients and gives the same estimated partial regression coefficient $b_{2 \cdot 1}$. The t-statistic is calculated with $(n-3)$ degrees of freedom, to account for the fact that variable X has been used to obtain the residuals $R_{Y|X}$ and $R_{Z|X}$.

This permutation technique has also been used by earlier workers in the context of tests of partial correlation coefficients for distance matrices (*e.g.*, Smouse *et al.*, 1986; Legendre and Fortin, 1989). Although the Kennedy method will give the same value as the Freedman and Lane method for the estimate of the slope coefficient $b_{2 \cdot 1}$, the value of the t-statistic under permutation is different for the two methods.

The reason for the discrepancy between the methods is a rather subtle point, but has important consequences (*e.g.*, see Section 4.1). Kennedy's method ostensibly removes the effect of the variable which is not of interest for the test by an initial regression of Y on X (Step 2). Thus, the parameter associated with X in the underlying multiple regression model remains fixed throughout the ensuing permutations. In the Freedman and Lane method, however, this parameter does not stay fixed. The permuted residuals $R^*_{Y|X}$ are added back onto the fitted values to obtain Y^*. These are then regressed on X and Z together, so the parameter for X in the multiple regression model changes with each permutation. If the true value of the parameter of the regression of Y on X alone were known (*i.e.*, β_1 in Step 2 above), there would be no difference between these two methods. Although there is no relationship between $R_{Y|X}$ and X, some small relationship is reintroduced between $R^*_{Y|X}$ and X, simply by the permutation of these residuals. The method of Freedman and Lane takes this into account by maintaining the conditioning on X throughout the permutations, whereas that of Kennedy does not.

The two methods described above have generally been called permutation "under the null model" by ter Braak (1990), or "under the reduced model" by Cade and Richards (1996). We tested the prediction that these two techniques would give similar results, in various circumstances (see Section 3).

2.3. Permutation of Residuals under the Full Model

This approach was developed by ter Braak (1990, 1992), who introduced it as a permutational analog (resampling without replacement) to the bootstrapping method (resampling with replacement) which had been proposed by Hall and Titterington (1989). It is referred to as permutation "under the full model" by ter Braak (1990) and it uses the residuals from the full regression model as the permutable units for the test. The rationale for the method is that it uses the estimate of $\beta_{1 \cdot 2}$ as part of the test (akin to the Freedman and Lane method), but also uses the original estimate of $\beta_{2 \cdot 1}$ as part of the permutational procedure. This should have the effect of reducing the variance of the parameter of interest under permutation for purposes of the test, thus increasing power (ter Braak, 1992). This permutation method proceeds as follows:

1. The variable Y is regressed on X and Z together to obtain estimates b_0 of μ, $b_{1 \cdot 2}$ of $\beta_{1 \cdot 2}$, $b_{2 \cdot 1}$ of $\beta_{2 \cdot 1}$ and residuals $R_{Y|XZ}$, as well as a reference value t_{ref} for the original data.
2. The residuals $R_{Y|XZ}$ are permuted randomly, producing $R^*_{Y|XZ}$.
3. New values are calculated from the permuted residuals as follows:

$$Y^* = b_0 + b_{1 \cdot 2}X + b_{2 \cdot 1}Z + R^*_{Y|XZ}.$$

4. The new values Y^* are regressed on X and Z to obtain an estimate $b^*_{2 \cdot 1}$ and a value t^* under permutation. Here, $t^* = (b^*_{2 \cdot 1} - b_{2 \cdot 1})/\text{se}(b^*_{2 \cdot 1})$.
5. Steps 2–4 are repeated a large number of times, yielding a distribution of values of t^* under permutation.
6. The absolute value of the reference value t_{ref} is placed in the distribution of absolute values of t^* obtained under permutation (for a two-tailed t-test). The probability is calculated as the proportion of values in this distribution greater than or equal, in absolute value, to the absolute value of t_{ref}. Permutation of the residuals under the full model preserves all covariances across the permutations, *i.e.*, among Y, X and Z, as well as among the X variables if there are more than one. Note that, as was the case for the Freedman and Lane method, this permutation test under the full model also has asymptotically exact significance levels.

There are two other important points to note about this approach. The first is that the t-statistic is calculated under permutation according to the hypothesis that $b^*_{2\cdot1} = b_{2\cdot1}$, that is, that the values of $b^*_{2\cdot1}$ obtained under permutation are close to the original estimated value of $b_{2\cdot1}$. For this reason, this approach has also been called permutation "under the alternative hypothesis" by ter Braak (1990). A diagrammatic interpretation of the conceptual difference between this approach and the approach of permutations under the reduced model is shown in Figure 1.

The second important aspect of this method is that it requires a pivotal statistic (such as the t-statistic) for the test. The method works because the variability of the permuted values $b^*_{2\cdot1}$ around the estimated value $b_{2\cdot1}$ mimics the variability of $b_{2\cdot1}$ around the true value $\beta_{2\cdot1}$, as described by ter Braak (1992) and shown in Figure 1. Thus, the value of a pivotal statistic for the test of $\beta_{2\cdot1} = 0$ with the real data, as a ratio, can be validly compared with the distribution of that statistic for the test of $b^*_{2\cdot1} = b_{2\cdot1}$.

We note that an equivalent model for Step 4 would be to regress $R^*_{Y|XZ}$ on X and Z to obtain an estimate $b^*_{2\cdot1}$ (different from that obtained in Step 4 above) and a value $t^* = b^*_{2\cdot1}/\text{se}(b^*_{2\cdot1})$ under permutation which has the same value as in Step 4 and may be compared directly to the original t value (*e.g.*, Manly, 1997).

In his exposition of this approach, ter Braak (1992) suggested that, by reference to the theory put forth by Hall and Titterington (1989) in the context of bootstrapping, permutations under the full model should have greater power than permutations under the reduced model. Empirical simulation results demonstrating this effect have not been provided, however. We tested ter Braak's prediction in the present study.

Thus, the five methods we compared were: (1) permutation of raw data (Manly, 1997); (2) the Freedman and Lane method (1983); (3) the Kennedy method (1995); (4) the ter Braak method (1992) and (5) the normal-theory t-test.

3. METHODS FOR SIMULATIONS

We restricted our attention to the case of a linear model using ordinary least-squares regression for two-tailed t-tests. Systematic changes to

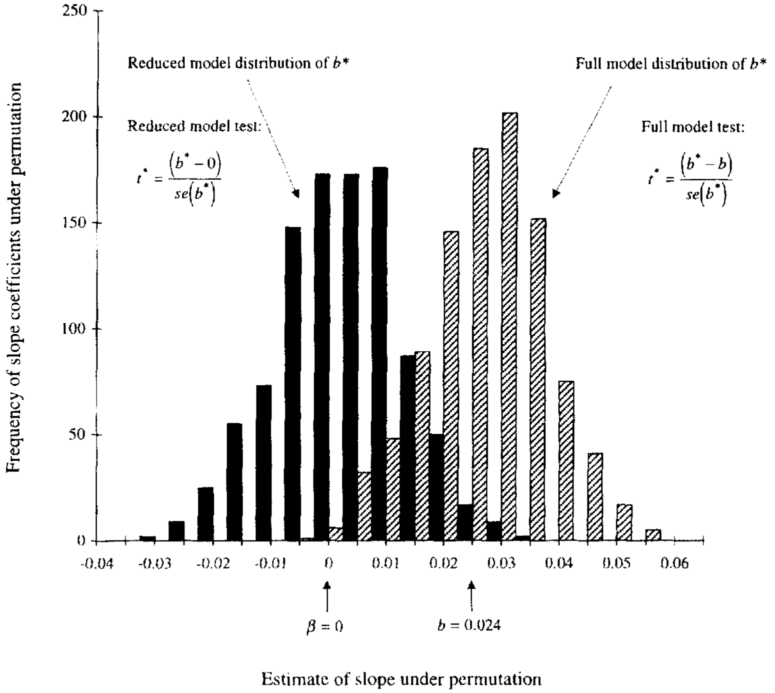

FIGURE 1 Frequencies of occurrence of 1000 values of the slope coefficient under each of two different methods of permutation: permutation under the reduced model (dark bars) and permutation under the full model (striped bars), where the null hypothesis was false and the slope parameter of interest was $\beta = 0.04$ (errors were $N(0, 1)$ and the covariable's parameter was equal to zero). The familiar method of permutation under the reduced model produces a distribution of values of the slope coefficient around the value of zero, in keeping with the null hypothesis that the slope parameter $\beta = 0$. The method of permutation under the full model, or "alternative hypothesis," produces a distribution of values of the slope coefficient around the estimate obtained from the original data, b. For the present data, the slope was estimated at $b = 0.024$. The distribution of b^* under the full model mimics its distribution under the reduced model, but is centered on b instead of zero. For the test, the reference value t_{ref} is the same for both methods (see text).

particular factors of interest were made within this framework. Replicate simulations were done across all levels of all factors under consideration in a completely crossed experimental design. Computer programs for simulations and calculations were written and compiled in Fortran 77. Random values were generated using functions snorm for standard normal errors, sexpo for standard exponential errors, and genunf for uniform values, from the "Ranlib" library of subroutines (l'Ecuyer and Côté, 1991). Permutations of the simulated data were done using a uniform random generation algorithm (Furnas, 1984).

3.1. Examination of Type I Error

Empirical probabilities of type I error were studied for the four permutation methods with regard to the following four factors:

1. The size of the sample, $n = \{9, 18, 36, 54, 72, 90\}$.
2. The degree of collinearity between X and Z, $\rho = \{0.0, 0.5, 0.9\}$.
3. The size of the covariable's parameter, $\beta_{1\cdot 2} = \{0.0, 0.5, 1.0, 2.41\}$.
4. The distribution of added random errors, $\varepsilon = \{$standard normal, exponential, or exponential$^3\}$.

X and Z were set uniformly at values of $\{1, 2, 3\}$ in a crossed 3×3 design, in accordance with model I regression. Y was generated as a vector using the model $Y = \mathbf{W}\boldsymbol{\beta} + \varepsilon$ where matrix $\mathbf{W}$ contained the standardized vectors X and Z as columns, vector $\boldsymbol{\beta}$ contained chosen values of the parameters $\beta_{1\cdot 2}$ and $\beta_{2\cdot 1}$ and vector ε contained errors drawn randomly from a standard normal (0, 1) or an exponential (1) distribution (factor 4 above). Radically non-normal errors were created as in Manly (1997) by a third distribution: random drawing of an exponential variate which was then cubed (denoted in factor 4 above as exponential^3). For type I error, the parameter under test $\beta_{2\cdot 1}$ was set at 0. The smallest sample size consisted of one replicate at each combination of the 3×3 values for (X, Z): (*i.e.*, $n = 9$). Sample size (n) was increased in multiples of 9 by increasing the number of replicates $(1, 2, 4, 6, 8$ or $10)$ drawn within each combination of (X, Z) (corresponding to $n = 9, 18, 36, 54, 72$ or 90, respectively).

Collinearity was introduced between X and Z through the use of the square root of a correlation matrix, $\mathbf{R}$, reflecting the desired correlation between the two variables. We computed

$$\mathbf{W}_R = \mathbf{W}\mathbf{R}^{1/2} \tag{4}$$

where $\mathbf{W}_R$ is a matrix of two new variables X_R and Z_R, correlated according to the correlation coefficient (ρ) in the correlation matrix $\mathbf{R}$, which could then be used to compute simulated Y, using the model $Y = \boldsymbol{W_R\beta} + \varepsilon$.

Data were simulated in all combinations of the chosen levels for each of the above four factors. Empirical type I errors and 95% confidence intervals were calculated from 10,000 simulated data sets for each combination of factors for each permutation method. There were 999 permutations done according to each of the four methods of permutation for each set of simulated data. The probability associated with the normal-theory t-test was also calculated for each data set. The significance level for the tests was set at $\alpha = 0.05$. For type I error to match the significance level, the number of significant p-values out of the 10,000 simulations was expected to be 500 for each of the methods.

3.2. Examination of Power

The empirical probability of type I error for the Kennedy method was clearly above 0.05, especially with small samples (see Results), so this method was not considered further and comparisons of power were only done for the three other permutation methods (*i.e.*, Manly, Freedman and Lane and ter Braak) and the normal-theory t-test. Power was defined as the proportion of rejections, out of 10,000 simulations, of the null hypothesis when it was false. Power was examined using the same design (with the same four factors) as was done for comparisons of type I error.

Data were generated in the same manner as for type I error, except that the null hypothesis was false ($\beta_{2\cdot 1} \neq 0$). Preliminary trials were done to determine appropriate values for $\beta_{2\cdot 1}$ under different conditions of collinearity and distributions of errors. We chose values of $\beta_{2\cdot 1}$ for different sample sizes that would allow measurement and comparison of power (*i.e.*, $\beta_{2\cdot 1}$ was chosen to be large enough to have the null hypothesis rejected relatively frequently, but not so large as to generate a 100% rejection rate). Naturally, $\beta_{2\cdot 1}$ needed to be smaller for larger sample sizes. For normal or exponential error distributions, the following parameters were used: $\beta_{2\cdot 1} = 0.75$ (for $n = 9$, 18), $\beta_{2\cdot 1} = 0.4$ (for $n = 36$, 54), and $\beta_{2\cdot 1} = 0.3$ (for $n = 72$, 90). There were 10,000 simulated data sets for each combination of the above factors and 999 permutations done on each set of simulated data.

Theoretically, increasing values of the parameter under test should result in greater power for the full-model method under permutation, compared to the reduced-model method (ter Braak, pers. comm.). This specific prediction was examined by producing results on power for the Freedman and Lane method and the ter Braak method alone while increasing the value of $\beta_{2\cdot 1}$. Data were simulated with $\beta_{1\cdot 2} = 0$ and $\rho = 0$ for each of exponential errors and exponential errors cubed for sample sizes of $n = 9$, 18, 36, 54, 72 and 90. There were 10,000 simulations, each tested with 999 permutations.

3.3. Examination of Effects of an Outlier in X

Kennedy and Cade (1996) and Manly (1997) drew different conclusions concerning effects of an outlier in X on the permutation of raw data *versus* permutation of residuals under the reduced model. We therefore simulated data in the same general manner as described by Manly (1997, see pp. 162–166) and Kennedy and Cade (1996). Our study differs from theirs by providing more extensive sets of simulations, yielding a stronger basis of comparison.

First, $(n-1)$ values of X were chosen randomly from a uniform distribution on the interval $(0, 3)$. The nth value of X was set as an outlier equal to 55. Then, n values of Z were chosen randomly from a uniform distribution on the interval $(0, 3)$. We wished to investigate type 1 error, so $\beta_{2\cdot 1}$ was set equal to zero. We generated data Y using the model in Eq. (1) (with $\mu = 0$) for all combinations of the following parameters: $\beta_{1\cdot 2} = \{0, 2, 4, 5, 6, 8, 10, 20\}$, $n = \{10, 20, 100\}$, $\varepsilon = \{$normal, exponential$^3\}$. Note that although X and Z have been simulated without any explicit collinearity, they are not strictly uncorrelated here, as they were above for $\rho = 0.0$ (Sections 3.1, 3.2). Although X and Z were drawn randomly from uncorrelated uniform distributions ($\rho = 0.0$), they are finite samples from these distributions and thus have some measurable correlation $(i.e., r^2_{XZ} \geq 0.0)$. We generated 10,000 simulated data sets in this way, with each data set tested using 999 permutations under each of three different permutation methods: Freedman and Lane's (reduced model), ter Braak's (full model), and Manly's method (raw data permutation). The probability associated with the normal-theory t-test was also calculated for each data set.

Each simulated data set contained new values of X, Z and Y, as described above. This differs from the other sets of simulations, where X and Z were fixed for a given n. Measures of type I error for two-tailed tests were obtained as before for each of the methods.

As in Manly (1997), we investigated the further effect of collinearity in the situation where an outlier was present in X. Manly had done this by creating new Z values which were $Z\text{new} = 0.5X + Z\text{old}$. Although Manly stated that this creates collinearity between X and Z ($r_{XZ} = 0.972$), this is only the case in the presence of the outlier point, which is an extremely high leverage point. When the outlier point is removed, the collinearity is only $r_{XZ} = 0.286$ (see data in Tab. 8.8, Manly, 1997). Thus, instead, we introduced stronger collinearity by creating new Z values as $Z\text{new} = 1.0X + Z\text{old}$. By doing this, collinearity is apparent even when the high leverage point (the introduced outlier) is removed from the data ($r_{XZ} = 0.583$). We repeated the above experimental design (eight values of $\beta_{1\cdot 2}$ and three values of n for each of two error distributions) but where collinearity had been introduced. Again, 10,000 simulations were done.

4. RESULTS OF SIMULATIONS

4.1. Type I Error

The Kennedy method of permutation resulted in inflated type I error for data of virtually all kinds (*e.g.*, Fig. 2). This was especially apparent at small sample sizes, with the problem decreasing as the sample size increased (Fig. 2a). The presence of a non-zero parameter for the covariable (Fig. 2b, d) or the presence of collinearity between X and Z (Fig. 2c, d) had little influence on the consistently inflated type I error at small sample sizes for this method. For data generated with standard normal or standard exponential errors, type I error was significantly greater than $\alpha = 0.05$ for the Kennedy method in all situations for $n = 9$. In general, Kennedy's method did not consistently produce type 1 error at 0.05 until sample sizes reached $n = 54$.

Intuitively, one could consider using $(n-2)$ degrees of freedom for the Kennedy method under permutation only. The rationale for doing this is that the model in Step 5 of the Kennedy method, where $R^{*}_{Y|X}$ are

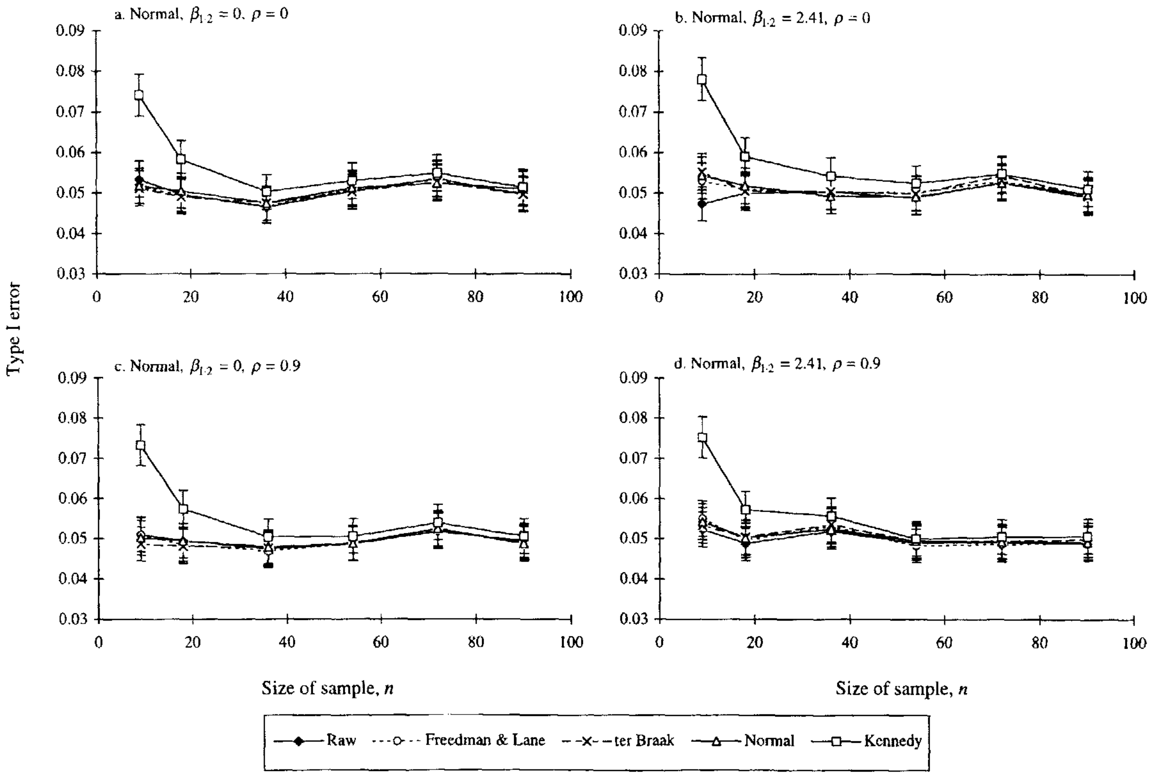

FIGURE 2 Type 1 error (± 95% C.I.) with increasing sample size for four methods of permutation and the normal-theory t-test obtained from 10,000 simulations where errors are standard normal and (a) $\beta_{1\cdot2} = 0$, $\rho = 0.0$; (b) $\beta_{1\cdot2} = 2.41$, $\rho = 0.0$; (c) $\beta_{1\cdot2} = 0$, $\rho = 0.9$; (d) $\beta_{1\cdot2} = 2.41$, $\rho = 0.9$.

regressed on $R_{Z|X}$, only contains 2 parameters, rather than the three parameters of the original model. That is, we might consider $R^{*}_{Y|X}$ and $R_{Z|X}$ to be two new variables under permutation. Some simulations were done to determine the extent to which the problems of inflated type I error with the Kennedy method could be remedied by using $(n-2)$ degrees of freedom under permutation, even though $(n-3)$ degrees of freedom had been used to calculate t_{ref}.

The simulations showed that the problem of inflated type I error with the Kennedy method was indeed reduced by using $(n-2)$ degrees of freedom under permutation (Tab. I, number of independent variables = 2). However, with increases in the number of covariables in the model (*i.e.*, when X becomes a matrix), the bias in the Kennedy method becomes larger and larger, such that even using $(n-2)$ degrees of freedom under permutation does not eliminate the problem for small n (Tab. I, number of independent variables = 5 or 10). Although Kennedy's method asymptotically approaches the Freedman and Lane method as n increases (Tab. I), we obtained a 60.2% rejection rate of a true null hypothesis for $n = 12$ with 9 covariables using Kennedy's method, which is clearly unacceptable for a valid testing procedure.

For normal or exponential errors, there were no significant differences among the other three methods (raw data permutation, Freedman and Lane's or full-model permutation); they matched the normal-theory t-test and had type 1 error which did not differ significantly from 0.05 in all sets of simulations. With radically non-normal errors, however, all methods had a tendency to become more conservative. The normal-theory t-test had type 1 error consistently below 0.05 for all sets of simulations with exponential cubed errors (Fig. 3). For smaller sample sizes ($n < 54$) the permutation methods (except the Kennedy method) also showed conservative type 1 error rates. When the covariable's parameter was equal to zero, raw data permutation maintained error rates at 0.05 for small samples, while the model-based tests were too conservative (Fig. 3a, c). In contrast, with increases in the covariable's parameter and with collinearity between X and Z, permutation under the reduced model had the best level-accuracy for smaller sample sizes (Fig. 3d). All permutation methods, however, converged asymptotically to appropriate type 1 error much more quickly than the normal-theory t-test in these situations of extremely non-normal error distributions.

TABLE I Type I error from 20,000 simulations, each with 999 permutations, for three different methods of permutation: the Freedman and Lane method (F and L), the Kennedy method where $(n-2)$ df were used for the test $(\mathrm{K}(n-2))$, and the Kennedy method where $(n-p)$ degrees of freedom were used for the test $(\mathrm{K}(n-p))$, where p is the number of independent variables plus 1 (*i.e.*, the total number of parameters in the full model including an intercept). Note that the different degrees of freedom used for the Kennedy method were for permuted data only. All of these methods of permutation use $(n-p)$ df when calculating t_{ref}. The simulated data Y were generated under the following model: $Y = \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_{p-1} X_{p-1} + \varepsilon$, where each independent variable X_i was generated from a random uniform distribution on the interval (0, 3), the ε's were random deviates drawn from a standard exponential distribution, β_1 was set at zero and all other β_i's were set equal to 1.0. The t-test was then done with permutations to test the true null hypothesis: $\beta_1 = 0$

No. of variables $(p-1)$	*Sample size, n*	*F and L*	*K*$(n-2)$	*K*$(n-p)$
2	9	0.048	0.054	0.069
2	18	0.050	0.051	0.058
2	36	0.051	0.051	0.055
2	54	0.049	0.050	0.052
2	72	0.052	0.052	0.054
2	90	0.050	0.050	0.052
5	9	0.050	0.097	0.216
5	18	0.051	0.056	0.095
5	36	0.049	0.051	0.066
5	54	0.051	0.051	0.060
5	72	0.050	0.050	0.057
5	90	0.051	0.051	0.056
10	12	0.048	0.264	0.602
10	18	0.048	0.070	0.206
10	36	0.051	0.054	0.097
10	54	0.050	0.052	0.076
10	72	0.049	0.050	0.069
10	90	0.051	0.052	0.065

4.2. Power

The Kennedy method was not included in tests of power since it generally had inflated type I error rates. Not surprisingly, all methods showed decreases in power with increases in collinearity between X and Z (*e.g.*, see the differences in scales of the ordinates for Fig. 4a *vs.* b and Fig. 4c *vs.* d). Also, for all methods, there were increases in power with increases in sample size (*e.g.*, Fig. 5a *vs.* c and Fig. 5b *vs.* d).

For data generated with normal errors, there were no significant differences in power among any of the methods, including the normal-theory t-test (Fig. 4). This result was consistent for all sets of simulations for data generated with exponential errors as well (not

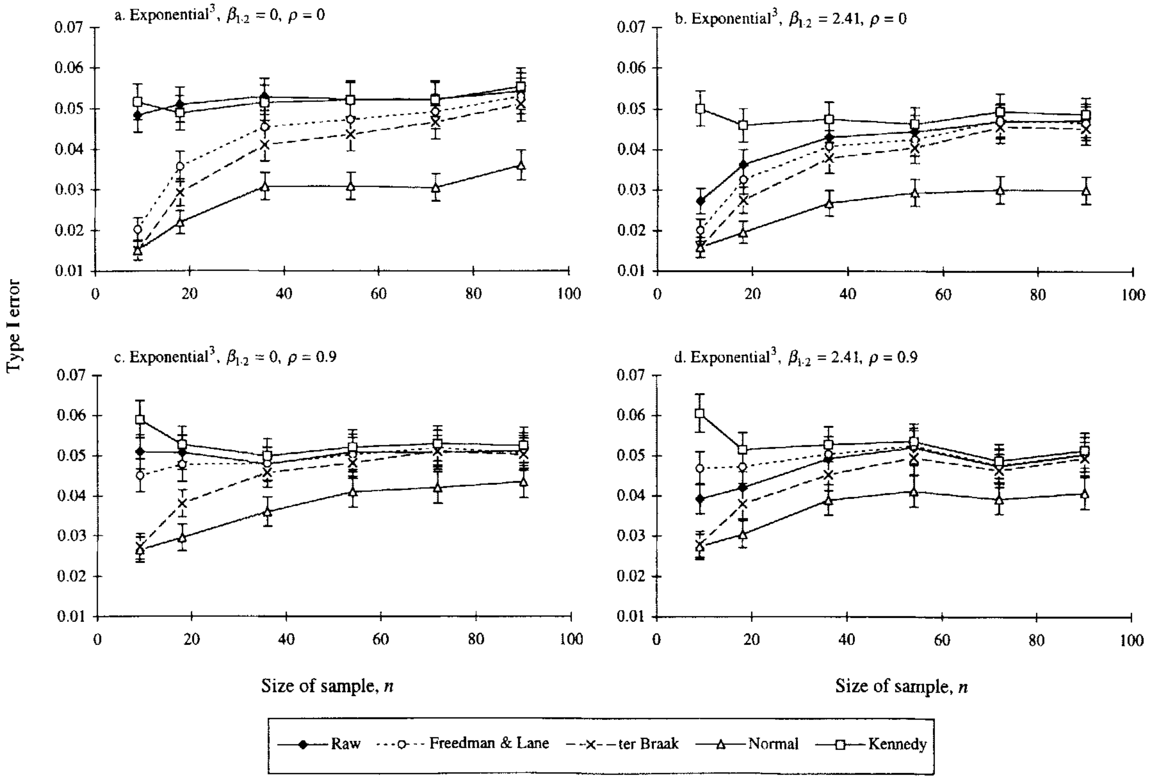

FIGURE 3 As for Figure 2, but where errors were exponential cubed.

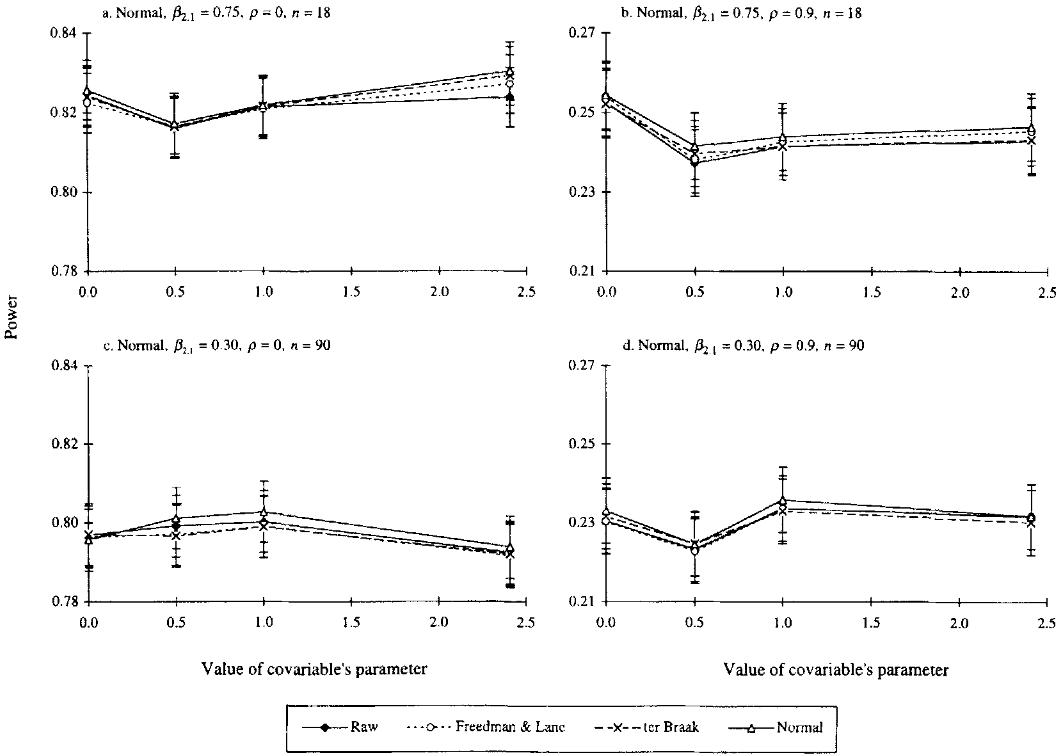

FIGURE 4 Power (± 95% C.I.) with increasing values of $\beta_{1.2}$ (covariable's parameter) for three methods of permutation and the normal-theory t-test obtained from 10,000 simulations where errors were standard normal with (a) X and Z uncorrelated, $n = 18$; (b) X and Z collinear ($\rho = 0.9$), $n = 18$; (c) X and Z uncorrelated, $n = 90$; and (d) X and Z collinear ($\rho = 0.9$), $n = 90$.

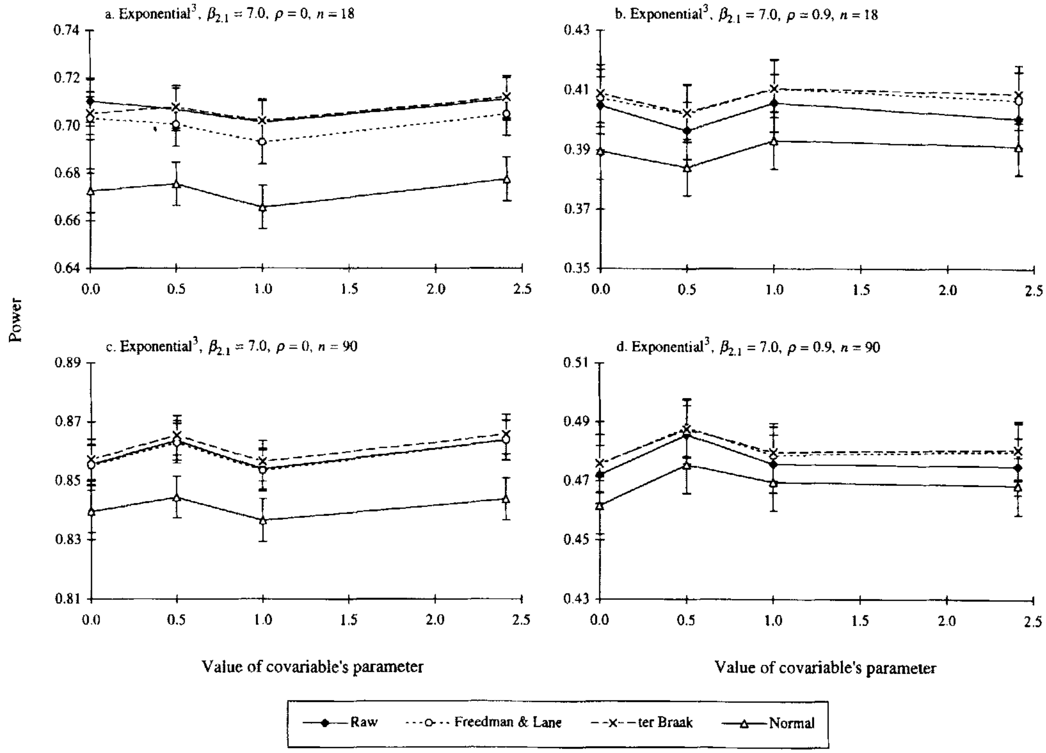

FIGURE 5 As for Figure 4, but where errors were exponential cubed.

shown). On the other hand, for data generated with radically non-normal errors, the normal-theory t-test was significantly less powerful then the permutation methods when $\rho = 0.0$ or $\rho = 0.5$ (*e.g.*, Fig. 5a, c). When $\rho = 0.9$, the normal-theory t-test was less powerful, on average, than the permutation methods, but not significantly so (*e.g.*, Fig. 5b, d). None of the permutation methods differed significantly in terms of power for any of the simulations done here (Figs. 4 and 5 and results for exponential errors, not shown).

Further tests on power of the reduced and full-model methods showed more details concerning any possible difference in these two methods with increasing values of $\beta_{2\cdot 1}$ (Fig. 6). Note how the power curves for the two methods are virtually identical (Fig. 6a, b, c). Differences between the two methods were detectable, however, at the smallest sample sizes, but disappeared as sample size increased. In Figure 6 (d, e, f), the difference between the two methods is plotted as [power(Freedman and Lane) − power(ter Braak)], so negative values indicate comparatively greater power for the full-model method of permutation. At the lowest values of $\beta_{2\cdot 1}$, the Freedman and Lane method had slightly greater power. As $\beta_{2\cdot 1}$ increased, the ter Braak method became more powerful than the Freedman and Lane method. As power approached 100%, the two methods converged. The size of the difference in power between the methods was, at most, only about 1.5% (Fig. 6d, $n = 9$). The size of the difference also decreased as sample size increased, becoming barely detectable by $n = 36$. No confidence intervals are plotted in Figure 6, for clarity, but out of 132 sets of simulations (11 values of $\beta_{1\cdot 2}$ for each of $n = \{9, 18, 36, 54, 72, 90\}$ and for each of exponential or exponential^3 errors), only three tests showed a significant difference between the two methods. This is certainly no more than could be expected by chance alone with this number of tests.

4.3. Effect of an Outlier in *X*

The presence of an outlier in X (the covariable) caused the type 1 error for the method of raw data permutation to be destabilized when $\beta_{1\cdot 2} \neq 0$ (Figs. 7, 8). This destabilization was not systematic for small to intermediate sample sizes (*e.g.*, Fig. 7a, b). That is, sometimes the

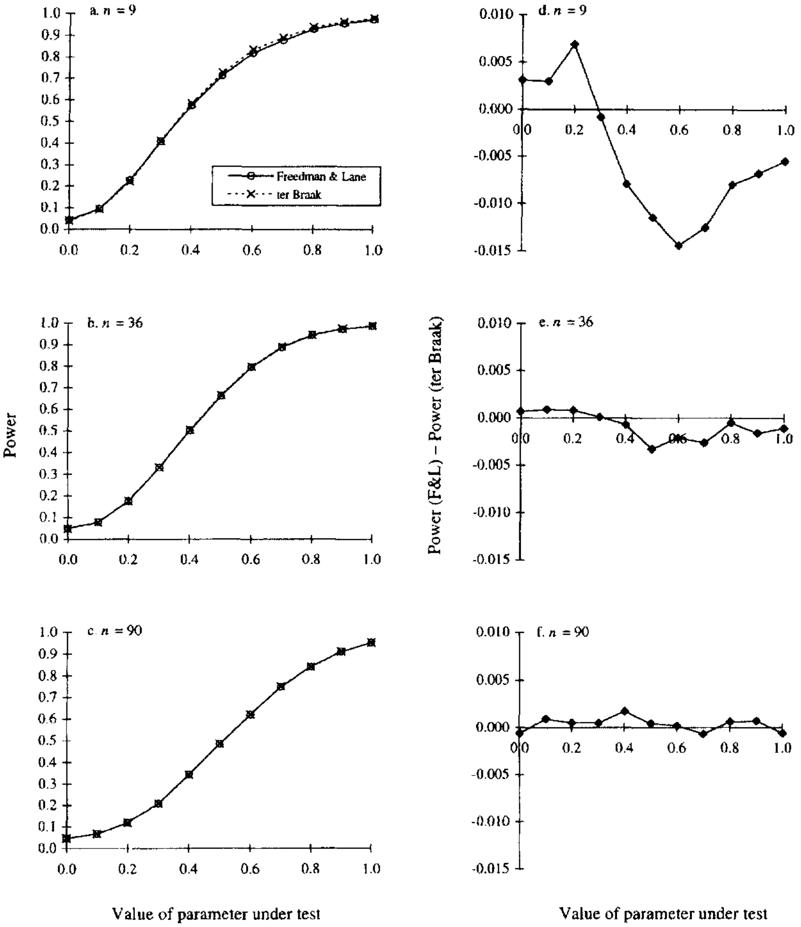

FIGURE 6 Power with increasing values of $\beta_{2\cdot 1}$ (parameter under test) for two methods of permutation obtained from 10,000 simulations where errors were exponential with (a) $n = 9$, (b) $n = 36$, (c) $n = 90$. Also, the difference in power between the Freedman and Lane method and the ter Braak method of permutation for increasing values of $\beta_{2\cdot 1}$ with (d) $n = 9$, (e) $n = 36$, (*f*) $n = 90$.

method erred on the side of giving too many rejections (inflated type 1 error), while at other times the test was too conservative. With larger sample sizes, however, the problem was a consistent inflation of type 1 error (*e.g.*, $n = 100$, Fig. 7c).

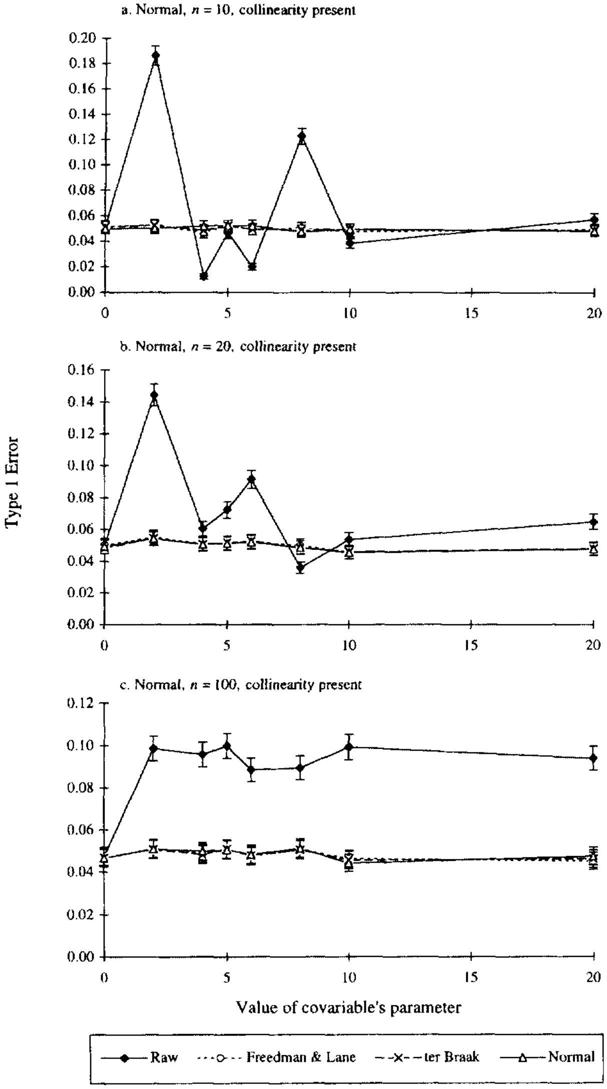

FIGURE 7 Type I error (±95% C.I.) with increasing values of $\beta_{1.2}$ (covariable's parameter) for three methods of permutation and the normal-theory t-test obtained from 10,000 simulations where X contains an outlier and errors were standard normal for (a) $n = 10$, (b) $n = 20$ and (c) $n = 100$. Results are shown for the situation where there was collinearity between X and Z. Similar results were obtained where no explicit collinearity was created (not shown).

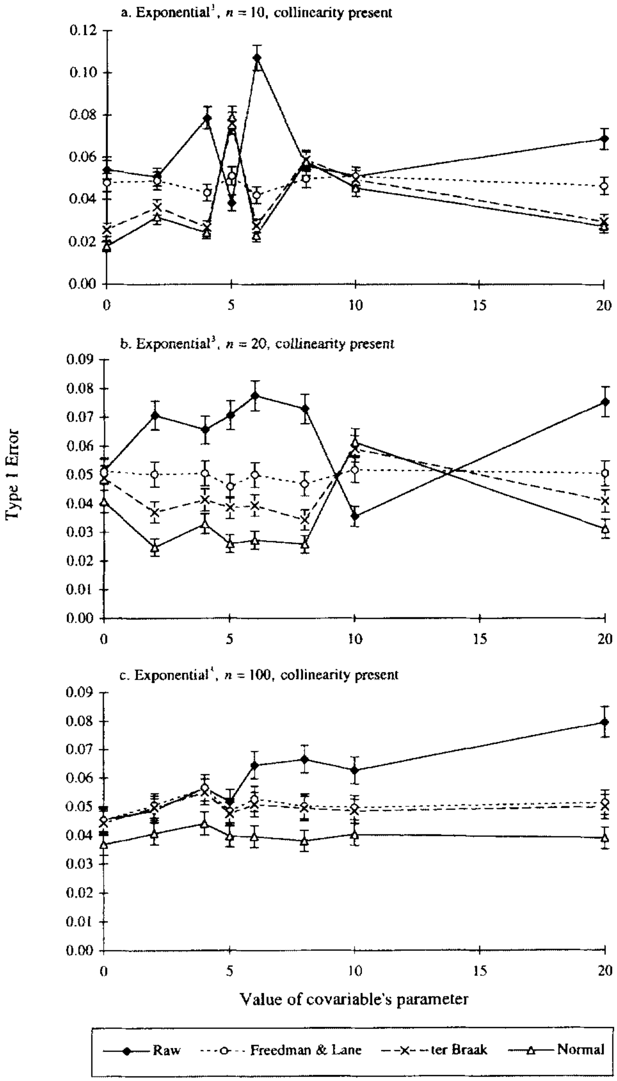

FIGURE 8 As for Figure 7, but where errors were exponential cubed.

In contrast, Freedman and Lane and ter Braak's method, along with the normal-theory t-test, showed no such problems and maintained type 1 error at the chosen $\alpha = 0.05$, provided errors were normally distributed (Fig. 7). When errors were radically non-normal, however, the type 1 error of the normal-theory t-test and ter Braak's method were destabilized in a similar manner to (but not to the same extent as) the destabilization of the test by raw data permutation. For a given set of simulations, these methods resulted in the opposite problem to that for raw data permutation: *i.e.*, when Manly's method gave too many rejections, ter Braak's method and the normal-theory t-test gave too few and *vice versa* (Fig. 8). Theoretical results supporting this observation will be published elsewhere by M. J. Anderson and J. Robinson.

Nevertheless, it is important to note that the problem disappears for ter Braak's method with large sample sizes ($n = 100$, Fig. 8c) or with more reasonable error structures (*i.e.*, normal or exponential, Fig. 7). This is not the case with raw data permutation, which for large sample sizes has inflated type 1 error when X contains such an outlier, regardless of the nature of the errors (Figs. 7c, 8c).

In all of the simulations we did that included an outlier in X, even in the most extreme situations, the type 1 error of the Freedman and Lane method never differed significantly from 0.05.

5. DISCUSSION

The primary conclusions obtained by this simulation study were the following:

1. The Kennedy method of permutation will not give equivalent results to the Freedman and Lane method of permutation under the reduced model for tests using the t-statistic. The Kennedy method has inflated type I error, especially with small sample sizes. Using $(n-2)$ rather than $(n-p)$ degrees of freedom (under permutation only), where p is the number of independent variables plus 1 (for the intercept), reduces the bias in this method when the number of covariables is small; the type I error is still seriously inflated with increases in the number of covariables at small sample sizes ($n \leq 18$). Permutation under the reduced model should therefore be done using the Freedman and Lane method.
2. Permutation of raw data, permutation under the reduced model (in the manner of Freedman and Lane) and permutation under the full model all gave asymptotically equivalent results in most situations

and provide good approximate tests for a partial regression coefficient by permutation. They have significantly greater power and type 1 error closer to nominal α than the normal-theory t-test for data with non-normal error structures.

3. Permutation of raw data resulted in destabilized type I error when the covariable contained an extreme outlier, whether or not there was collinearity between predictor variables or the data were normal or non-normal. This problem was not amended with increasing sample sizes, but rather resulted in consistently inflated type 1 error in these situations.
4. Permutation under the reduced model (Freedman and Lane, 1983) and under the full model (ter Braak, 1992) generally gave very similar results and would probably be equally appropriate for most situations. In the extreme situation of a remote outlier in the covariable (X) coupled with extremely non-normal errors and small sample sizes, type 1 error for the ter Braak method may be destabilized. This problem is generally in the direction of a more conservative test, however, and disappears for large n.
5. There was no significant difference in power among any of the permutation methods (excluding the Kennedy method, which was not included in power analyses).

Some comparisons of these methods of permutation have been done in the context of least absolute deviation (LAD) regression by Cade and Richards (1996). Their results suggested that permutation of raw data and permutation of residuals under either the reduced or full models had similar type I error when predictor variables were not correlated and when the covariable's parameter ($\beta_{1\cdot 2}$) was zero. When there was collinearity among the independent variables or when $\beta_{1\cdot 2} \neq 0$, however, permutations under the reduced model maintained type I error closer to nominal α than the other two methods for LAD regression. Our results with least squares are consistent with the few simulations for type I error provided by Cade and Richards (1996) and by Kennedy and Cade (1996). Formal theoretical comparisons of these methods for least squares will be published elsewhere by M. J. Anderson and J. Robinson.

We do not recommend the use of the Kennedy method as a substitute for the Freedman and Lane method. These methods are not equivalent. By not explicitly conditioning the test on the covariable(s) throughout the permutations, the type I error is inflated substantially with Kennedy's method. This becomes especially important as the number of covariables increases, especially with small sample sizes. We found type I error averaging as large as 0.602 with this method (with $n = 12$ and 9 covariables in the model). Attempts to adjust the inflated type I error by estimating the t-statistic with $(n - 2)$ df, as an intuitive remedy, did not completely cure this problem.

Although Kennedy (1995) showed how the proposed method was equivalent to the Freedman and Lane method of permutation in terms of producing the same estimate of the slope coefficient under permutation, the value of the pivotal t-statistic under permutation differs for the two methods. Insofar as it is necessary to use the pivotal t-statistic for tests of partial regression coefficients in a linear model, the Kennedy method of permutation is not equivalent to Freedman and Lane's method, causes inflated type 1 error and cannot be used. These results have implications for tests of partial correlation among distance matrices by permutation, proposed by Smouse *et al.* (1986) as an extension of the Mantel test. One of their proposed methods is the same as that suggested by Kennedy for the univariate test, but applied to distance matrices. We expect that it will suffer from the same problems as Kennedy's method. Simulation results supporting this will be published elsewhere by P. Legendre.

We also cannot unreservedly recommend the unrestricted use of permutation of raw data, which had some problems and differed significantly from results obtained with the model-based methods when the value of the covariable's parameter was not zero and X contained an outlier. It would appear that this method cannot handle these particular situations due to the fact that the relationship between X and Y is the held constant throughout the permutations, as suggested by Kennedy and Cade (1996). There are several possible reasons for the discrepancy between our results and those of Manly (1997) concerning the effect of an outlier in X. Manly (1997) did not obtain new values of X and Z for each simulation, but simply re-randomized the error terms to obtain new simulations of data. Also, Manly used 5000 simulations, whereas we used 10,000 simulations. In addition, Manly used 99 permutations for each simulation, whereas we used 999 permutations per simulation.

Depending on the circumstances, permutation of raw data may err on the side of too many rejections or on the side of too few. However, it was necessary to simulate a very extreme outlier in X in order to see a destabilization of type 1 error for raw data permutation. When there are such extreme outliers in predictor variables, these should be identifiable as high leverage points in diagnostic analyses prior to the regression analysis. Outliers in covariables may be removed from the data set, so that the potential problem may be caught ahead of time. Nevertheless, the presence of outliers in a multiple regression context with more than 2 predictor variables may not always be readily apparent or easy to find. Furthermore, permutation methods are prized for their lack of assumptions concerning distributions of variables, meaning we should hope that much diagnostic checking of distributions of variables would become unnecessary with the permutational approach. At the very least, these results highlight that the method of raw data permutation is not an exact test for a partial regression coefficient in a linear model unless all other parameters in the model are truly equal to zero. The method provides an approximate test, as stated by Manly (1997), relying on the pivotal statistic, like the model-based permutational strategies that permute residuals.

The methods of Freedman and Lane (permutation under the reduced model) and ter Braak (permutation under the full model) gave the best overall results in terms of level accuracy and power whether the errors were normal or exponential, in the presence or absence of collinearity or an outlier in the predictors, and when the value of the covariable's parameter was not zero. Permutation under the full model suffered from some destabilization of type 1 error in the situation with an outlier in X and radically non-normal errors coupled with small sample sizes. Generally in these cases type 1 error was too small, making the test slightly too conservative. This problem disappeared in any event with larger sample sizes. The introduction of an outlier in X had no effect on the level accuracy of the Freedman and Lane method of permutation, in any situation.

In terms of relative power, we detected no significant differences among any of the permutation methods. For the model-based methods, greater power should theoretically attend the ter Braak method of permutation under the full model. In situations of increasing values of the parameter under test, it was possible to demonstrate a slight difference in power between ter Braak's method and Freedman and Lane's method for small sample sizes. This difference was not significant, however, and never measured more than about 1.5% and thus, for practical purposes, is not an important distinction between the methods.

In general, in cases with radically non-normal errors, we found that the full-model method of permutation did not maintain level accuracy at small sample sizes as well as the reduced-model method (which was expected: ter Braak, 1992; Cade and Richards, 1996), although the nature of this deviation was on the conservative side, generally resulting in even smaller type I error, at least for the data we simulated. For each of the model-based permutation methods, level accuracy deviated most from α with small sample sizes. This is because these methods have only asymptotically unbiased type I error. The estimates of the regression coefficients used in these procedures are less accurate with smaller samples, causing the inexactness.

Although the Freedman and Lane method might be preferable to use with smaller sample sizes, for some situations there is a computational advantage in using the ter Braak method. One can use the permutation of a single set of residuals from the full model to test a number of different hypotheses concerning individual partial regression coefficients in a multiple regression model. With the Freedman and Lane method, on the contrary, testing several hypotheses about different coefficients in a multiple regression model will each require a different set of residuals from several different reduced models (Manly, pers. comm.). This is computationally more demanding and requires more thought and care on the part of the experimenter regarding which particular term or set of terms is being tested with any particular set of permutations.

The principles investigated here are not restricted to the univariate model with two predictor variables. We expect our results to hold generally for greater numbers of predictor variables in multiple regression or analysis of variance. We have not compared here the permutation methods for their use with multivariate distance matrices (*e.g.*, Smouse *et al.*, 1986; Legendre, in prep.), nor with multiple response variables as in canonical analysis (*e.g.*, ter Braak, 1987). It also remains to be investigated how well restricted randomization methods (*e.g.*, Brown and Maritz, 1982; Edgington, 1995) will perform when compared to the model-based permutation methods examined here, in situations where such restricted randomizations can actually be done. The type I error for the restricted randomization methods will be assured (unlike the method of unrestricted raw data permutation), but it remains to be seen how powerful these tests will be compared to the model-based permutation tests.

Substantial computer power is now available, enabling researchers to investigate the behavior of computationally intensive methods. Obtaining empirical measures of type I error or power allows direct practical comparisons of permutation methods. Current theoretical comparisons of the methods cannot provide us with complete information on how the methods will compare in different situations in practice.

Acknowledgements

We thank C. J. F. ter Braak, B. F. J. Manly, J. Robinson, A. J. Underwood and M. Haddon for their comments on this work. M. J. Anderson gives special thanks to B. Wingett for his continued enthusiasm and support. This work was supported by a U2000 Postdoctoral Fellowship to M. J. Anderson at the University of Sydney and by NSERC grant no. OGP7738 to P. Legendre.

References

Brown, B. M. and Maritz J. S. (1982) Distribution-free methods in regression, *Austral. J. Statist.*, **24**, 318–331.

Cade, B. S. and Richards, J. D. (1996) Permutation tests for least absolute deviation regression, *Biometrics*, **52**, 886–902.

Edgington, E. S. (1995) *Randomization Tests* (Third Edition). New York: Marcel Dekker.

Fisher, N. I. and Hall, P. (1990) On bootstrap hypothesis testing, *Austral. J. Statist.*, **32**, 177–190.

Freedman, D. and Lane, D. (1983) A nonstochastic interpretation of reported significance levels, *J. Bus. Econ. Statist.*, **1**, 292–298.

Furnas, G. W. (1984) The generation of random, binary unordered trees, *J. Classif.*, **1**, 187–233.

Hall, P. and Titterington, D. M. (1989) The effect of simulation order on level accuracy and power of Monte Carlo tests, *J. R. Statist. Soc. B*, **51**, 459–467.

Hope, A. C. A. (1968) A simplified Monte Carlo test procedure, *J. R. Statist. Soc. B*, **30**, 582–598.

Kempthorne, O. (1952) *The Design and Analysis of Experiments*. New York: John Wiley and Sons.

Kennedy, P. E. (1995) Randomization tests in econometrics, *J. Bus. Econ. Statist.*, **13**, 85–94.

Kennedy, P. E. and Cade, B. S. (1996) Randomization tests for multiple regression, *Commun. Statist.–Simulation Comput.*, **25**, 923–936.

l'Ecuyer, P. and Côté, S. (1991) Implementing a random number package with splitting facilities, *ACM Trans. Math. Software*, **17**, 98–111.

Legendre, P. and Fortin, M.-J. (1989) Spatial pattern and ecological analysis, *Vegetatio*, **80**, 107–138.

Legendre, P. and Legendre, L. (1998) *Numerical Ecology* (Second English Edition). Amsterdam: Elsevier Science BV.

Manly, B. F. J. (1991) *Randomization and Monte Carlo Methods in Biology* (First Edition). London: Chapman and Hall.

Manly, B. F. J. (1997) *Randomization, Bootstrap and Monte Carlo Methods in Biology* (Second Edition). London: Chapman and Hall.

Oja, H. (1987) On permutation tests in multiple regression and analysis of covariance problems, *Austral. J. Statist.*, **29**, 91–100.

Smouse, P. E., Long, J. C. and Sokal, R. R. (1986) Multiple regression and correlation extensions of the Mantel test of matrix correspondence, *Syst. Zool.*, **35**, 627–632.

Stapel, M. and ter Braak, C. J. F. Randomization and bootstrap tests in factorial experiments: does analysis follow from design? *International Biometrics Conference*, German-Dutch Region, Münster, 15–18 May 1994.

ter Braak, C. J. F. (1987) Ordination. In: *Data Analysis in Community and Landscape Ecology* (Eds., Jongman, R. H. G., ter Braak, C. J. F. and van Tongeren, O. F. R.), pp. 91–169, Wageningen, The Netherlands: Pudoc. Reissued in 1995 by Cambridge, England: Cambridge University Press.

ter Braak, C. J. F. (1990) *Update Notes: CANOCO version* 3.10. Wageningen, The Netherlands: Agricultural Mathematics Group.

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

Welch, W. J. (1990) Construction of permutation tests, *J. Am. Statist. Ass.*, **85**, 693–698.