Aust. N. Z. J. Stat. **43**(1), 2001, 75–88

PERMUTATION TESTS FOR LINEAR MODELS

MARTI J. ANDERSON[1] AND JOHN ROBINSON[2]*

University of Sydney

Summary

Several approximate permutation tests have been proposed for tests of partial regression coefficients in a linear model based on sample partial correlations. This paper begins with an explanation and notation for an exact test. It then compares the distributions of the test statistics under the various permutation methods proposed, and shows that the partial correlations under permutation are asymptotically jointly normal with means 0 and variances 1. The method of Freedman & Lane (1983) is found to have asymptotic correlation 1 with the exact test, and the other methods are found to have smaller correlations with this test. Under local alternatives the critical values of all the approximate permutation tests converge to the same constant, so they all have the same asymptotic power. Simulations demonstrate these theoretical results.

Key words: asymptotics; partial correlations; power; resampling.

1. Introduction

The first descriptions of permutation (or randomization) tests for linear statistical models can be traced back to the first half of the 20th century in the work of Fisher (1935) and Pitman (1937a, b, c). However, these tests are computationally intensive and the use of them, as opposed to the traditional normal-theory tests, did not receive much attention in the natural and behavioural sciences until the emergence of widely accessible computer power (Edgington, 1995; Manly, 1997).

There is general agreement concerning an appropriate method of permutation for exact tests of hypotheses in simple linear regression (or, more simply, tests for the relationship between two variables, e.g. Edgington, 1995; Manly, 1997). This is not the case, however, for partial tests in multiple linear regression, i.e. tests in the presence of concomitant variables. Several different methods of permutation have been proposed to test the significance of one or more regression coefficients in a multiple linear regression model (Brown & Maritz, 1982; Freedman & Lane, 1983; Collins, 1987; Oja, 1987; Gail, Tan & Piantadosi, 1988; Welch, 1990; ter Braak, 1992; Kennedy, 1995; Manly, 1997). Complex designs such as these are commonly used in biological and ecological studies, where several factors are of interest, concomitant environmental variables are measured, or nested hierarchies of sampling at various temporal and spatial scales are necessary.

Received April 1999; revised December 1999; accepted January 2000.

* Author to whom correspondence should be addressed.

[1] Centre for Research on Ecological Impacts of Coastal Cities, University of Sydney, NSW 2006, Australia.
Current address: Dept of Statistics, University of Auckland, Private Bag 92019, Auckland, New Zealand.

[2] School of Mathematics and Statistics, University of Sydney, NSW 2006, Australia.
e-mail: johnr@maths.usyd.edu.au

Acknowledgments. M.J. Anderson was supported by a University of Sydney U2000 Fellowship and funds from the Centre for Research on Ecological Impacts of Coastal Cities. The authors thank Cajo J.F. ter Braak for his contributions to this work.

Although proponents of the various permutational strategies have each provided a rationale supporting their approach, the methods have not been formally compared. In a recent empirical comparison of the type 1 error and power of the various permutation methods proposed, Anderson & Legendre (1999), using Monte Carlo methods, showed that none of the proposed tests was an exact test and that no two of them were identical. Our purpose is to theoretically examine the relationships among the proposed methods of permutation for tests of partial regression coefficients in linear models. The results explain the results of empirical simulations given in Anderson & Legendre (1999).

In Section 2, we give the model and notation for the problem and introduce an exact permutation test for a partial regression coefficient. Then we introduce the approximate permutation tests of Freedman & Lane (1983), Kennedy (1995), Manly (1997) and ter Braak (1992). Section 3 derives the asymptotic joint distributions of these statistics under permutation under the null hypothesis. Section 4 shows that the asymptotic powers based on these test statistics are the same under local alternative hypotheses. The last two sections give some results of empirical simulations and a comparative discussion of the methods.

2. Notation and description of methods

First, consider the paired observations, Z_i and Y_i, $i = 1, \dots, n$, where Y is the dependent variable of interest and Z contains fixed values. Without loss of generality, let Z and Y be standardized to have mean zero. The simple linear model is $Y_i = \beta Z_i + \epsilon'_i$. We wish to test the null hypothesis $H_0\colon \beta = 0$. We assume that the ϵ' are independent and identically distributed (iid) random variables. An appropriate test statistic for the two-tailed test is the square of the correlation coefficient, r_S^2. If the null hypothesis is true, then any of the $n!$ random pairings of the observations Y with Z are equi-probable. An exact test is therefore given by calculating the correlation for each of the $n!$ permutations of the observations Y, while keeping Z fixed.

We use the notation of Freedman & Lane (1983) to denote the permutation π of $I_n = \{1, 2, \dots, n\}$. So π moves $i \in I_n$ to $\pi i \in I_n$ as a 1 to 1 mapping of I_n onto itself. Equal weight $1/n!$ is assigned to each of the $n!$ possible permutations π. When used as a superscript, π denotes that the superscripted variable itself is permuted, whereas when used as a subscript, π denotes that the subscripted variable is derived from permuted and unpermuted variables. Thus, the value of the statistic for any particular permutation is $r_\pi^2 = (\sum_{i=1}^n Y_i^\pi Z_i)^2 / (\sum_{i=1}^n Y_i^2 \sum_{i=1}^n Z_i^2)$.

The probability for this test is $p = \Pr(r_\pi^2 \geq r_S^2)$, the fraction of the permutations for which $r_\pi^2 \geq r_S^2$. The test is conditional only on the order statistics. Also, the assumption of iid errors can be relaxed if units have been randomly allocated to treatments *a priori*. As the number of possible permutations $n!$ increases rapidly with n, a practical strategy is to perform the test using a random subset $M < n!$ of all possible permutations. Such a test is still exact.

Next, consider the familiar multiple regression model

$$Y = \mu'' + \beta_1 X + \beta_2 Z + \epsilon'', \tag{1}$$

where Y is the dependent variable of interest, X and Z are each a set of fixed values, and each of these is a value from samples of size n. In (1), Y, X and Z are considered to be typical values for these variables; the subscript $i = 1, \dots, n$ has been omitted here and in what follows, as a simplification. Exactly analogous results follow if Y, X or Z is multivariate, but for simplicity of notation we restrict attention throughout this paper to univariate Y, X and Z. For further simplicity and without loss of generality, we standardize Y, X and Z to have mean zero. Interest lies in testing the null hypothesis H_0: $\beta_2 = 0$ (i.e. that the partial regression coefficient for Z is not significantly different from zero). We wish to isolate a test of the relationship between Y and Z, while controlling for any concomitant effect of X. Now we describe an exact permutation test for H_0. The idea for such a test is implicit in Freedman & Lane (1983), but here we give it an explicit notation by reference to an appropriate test statistic.

First consider the linear model of the relationship between the dependent and the concomitant variable: $Y = \alpha X + \epsilon$. An appropriate test statistic for the null hypothesis of no relationship between Y and Z, over and above any relationship of Y with X, is the partial correlation coefficient r, or, restricting the discussion here to the two-tailed test, its square

$$r^2 = \frac{\sum (R_{Y|X} R_{Z|X})^2}{\sum R_{Y|X}^2 \sum R_{Z|X}^2}, \tag{2}$$

where $R_{Y\,|\,X} = Y - aX$ denotes the residuals of the regression of Y on X alone and $R_{Z\,|\,X} = Z - \gamma X$ denotes the residuals of the regression of Z on X alone. Note that $a = \sum YX / \sum X^2$ is an estimate of the unknown regression parameter α, while γ is known for fixed X and Z.

Now, if we knew what the relationship was between Y and X, i.e. if we knew α, then we could obtain an exact test by constructing random permutations of Y given X that were exchangeable under the null hypothesis. Specifically, the Y themselves are not exchangeable under H_0, for these include some portion of variability explained by X. It is the errors ϵ that are exchangeable random variables under H_0. Although they cannot be observed, they are conceptually available to us as that part of Y not explained by X. If α were known, we could consider the exact conditional distribution of Y under permutation given X. We would first obtain the true errors ϵ, given the observed values of Y and X as $\epsilon = Y - \alpha X$. These would be permuted to obtain ϵ^{π}. We could then create new observations that are genuine realizations of alternative possible values of Y conditional on X under a true null hypothesis as $Y_{\pi(E)} = \alpha X + \epsilon^{\pi}$. The test statistic under permutation for this exact test would therefore be

$$r_E^2 = \frac{\left(\sum (Y_{\pi(E)} - a_{\pi(E)} X) R_{Z|X}\right)^2}{\sum (Y_{\pi(E)} - a_{\pi(E)} X)^2 \sum R_{Z|X}^2}, \tag{3}$$

where $a_{\pi(E)} = \sum Y_{\pi(E)} X / \sum X^2$ and $Y_{\pi(E)} - a_{\pi(E)} X$ is the residual of $Y_{\pi(E)}$ removing the effect of X.

It may seem strange that to calculate r_E^2 we need to estimate the regression coefficient $a_{\pi(E)}$, when we have just created new possible values by pretending to know the value of α. It is important to remember, however, that the calculation of the original value of the test statistic (2) relies on just such an estimate. This estimation must therefore also occur under permutation for the values of the statistic under permutation (3) to be commensurate with the original observed value of r^2. We use our knowledge of α only to create new observations conditional on X that could occur under a true null hypothesis.

Given r^2 and all possible $n!$ values of r_E^2 under the null hypothesis, $p = \Pr(r_E^2 \geq r^2)$ gives an exact test of H_0. This exact test is conditional on the order statistics of the original observations Y and has two assumptions: (i) the relationship between Y and X conforms to a linear model, and (ii) the errors ϵ'' are iid. Again, the last assumption can be relaxed if Z consists of codes for an experimental design allocated randomly to units *a priori*.

It is not possible to carry out an exact permutation test like this in practice because we cannot know α. It is at this point that the literature diverges into many opinions concerning an appropriate approximate permutation method for such a test. Only in the special case of X containing several replicates of each of several fixed values can an exact test be done. In that case, by restricting permutations within sets of observations that take similar values for X, α remains invariant under permutation (Brown & Maritz, 1982). Otherwise, an approximate method is necessary.

The first approximate test we consider is that provided by Freedman & Lane (1983). This method of permutation has also been called permutation under the reduced model. It is similar to the exact test, except that ϵ and α are replaced by their least-squares estimates $R_{Y\,|\,X}$ and a, respectively. So $Y_{\pi(F)} = aX + R^{\pi}_{Y\,|\,X}$, and the test statistic under permutation is

$$r_F^2 = \frac{\left(\sum(Y_{\pi(F)} - a_{\pi(F)}X)R_{Z|X}\right)^2}{\sum\left(Y_{\pi(F)} - a_{\pi(F)}X\right)^2 \sum R^2_{Z|X}}, \tag{4}$$

where $a_{\pi(F)} = \sum Y_{\pi(F)}X / \sum X^2$.

Kennedy (1995) proposed another method of permutation which he claimed was identical to the method of Freedman and Lane. The test is based on the general idea that the partial regression coefficient is equivalent to the simple regression coefficient of residuals. Here, the test statistic under permutation is the simple correlation coefficient between $R^{\pi}_{Y\,|\,X}$ and $R_{Z\,|\,X}$, thus

$$r_K^2 = \frac{\left(\sum R^{\pi}_{Y|X}R_{Z|X}\right)^2}{\sum R^2_{Y|X}\sum R^2_{Z|X}}. \tag{5}$$

Although both Kennedy's and Freedman and Lane's methods permute residuals $R_{Y\,|\,X}$, the values r_K^2 and r_F^2 differ under permutation.

Manly (1997) proposed simply permuting observed values Y for the test of partial correlation. This gives, under permutation,

$$r_M^2 = \frac{\left(\sum(Y^{\pi} - a_{\pi(M)}X)R_{Z|X}\right)^2}{\sum(Y^{\pi} - a_{\pi(M)}X)^2\sum R^2_{Z|X}}, \tag{6}$$

where $a_{\pi(M)} = \sum Y^{\pi}X / \sum X^2$. Although permuting observations gives an exact test of multiple correlation, there has been some controversy concerning the validity of this approach for partial tests (Kennedy & Cade, 1996; Manly, 1997).

Ter Braak (1992) proposed permutation of residuals of the full model, rather than residuals of the reduced model as in Freedman & Lane (1983). The test statistic under permutation for this method is

$$r_T^2 = \frac{\left(\sum(R^{\pi}_{Y|XZ} - k_{\pi}X)R_{Z|X}\right)^2}{\sum(R^{\pi}_{Y|XZ} - k_{\pi}X)^2\sum R^2_{Z|X}}, \tag{7}$$

where $k_{\pi} = \sum R^{\pi}_{Y\,|\,XZ}X / \sum X^2$ and $R^{\pi}_{Y\,|\,XZ}$ are the permuted least-squares residuals of the full model (1).

Note that all these methods use the partial least-squares correlation coefficient r^2 calculated on the original data as the value against which to compare distributions of permuted values r_F^2, r_K^2, r_M^2 or r_T^2.

3. Distribution associated with the null

3.1. Permutation under the reduced model

Let $R_{Y\,|\,X}$ and $R_{Z\,|\,X}$ satisfy Condition C of the Appendix. Then $\sqrt{n}\, r_K \xrightarrow{d} \mathrm{N}(0,1)$ by Theorem 2. Now, we can write the relationship between the partial correlation coefficients used in the Freedman & Lane (1983) and Kennedy (1995) methods under permutation as

$$r_F^2 = \frac{r_K^2}{(1 - A_\pi^2)}\,, \tag{8}$$

where $A_\pi^2 = (\sum R_{Y\,|\,X}^{\pi} X)^2 / \sum R_{Y\,|\,X}^2 \sum X^2$, the squared correlation coefficient between $R_{Y\,|\,X}^{\pi}$ and X. If X also satisfies Condition C, then $\sqrt{n}\, A_\pi \xrightarrow{d} \mathrm{N}(0,1)$ and $A_\pi \xrightarrow{p} 0$, and so $\sqrt{n}\, r_F \xrightarrow{d} \mathrm{N}(0,1)$.

Although both methods use test statistics that converge to the same distribution under permutation for large n, there is an important distinction between them. For the Kennedy method, the permuted residuals $R_{Y\,|\,X}^{\pi}$ are regressed directly on $R_{Z\,|\,X}$. This means that the value of a remains fixed throughout the permutations. For the Freedman and Lane method, in contrast, the permuted residuals $R_{Y\,|\,X}^{\pi}$ are added back onto the fitted values to create new values $Y_{\pi(F)}$ under permutation. The important point here is that the value $a_{\pi(F)}$ does not stay constant, but changes with each permutation π. There is no linear relationship between $R_{Y\,|\,X}$ and X. It is only by permuting the $R_{Y\,|\,X}$ to form $R_{Y\,|\,X}^{\pi}$ that a small relationship is re-introduced, so A_π is non-zero for any particular permutation π.

The numerator of r_F^2 is the same as that for r_K^2 (cf. (4) and (5)). The value of the cross-product, and therefore of the estimated partial regression coefficient for Z, is the same for the two methods under permutation. This is what led Kennedy (1995) to suggest that the methods were equivalent. Although Kennedy & Cade (1996) discussed the importance of using an asymptotically pivotal statistic, such as r^2, t or F, for partial tests in multiple regression, they did not appear to note that the equivalence of the Kennedy method with that of Freedman and Lane extends only to the value of the partial regression coefficient. Under the null hypothesis, we wish to perform a permutation test which is completely conditional on X. The Freedman and Lane method ensures that this conditioning on X is maintained throughout the permutations, whereas the Kennedy method does not.

Although the difference between the two methods disappears asymptotically, (8) shows that r_K^2 will be consistently smaller than or equal to values of r_F^2. The observed value r^2 will thus appear more extreme more often for the Kennedy method under permutation, resulting in tail probabilities that are too small, and inflating the type 1 error. Empirical simulations support these results (Anderson & Legendre, 1999).

3.2. Permutation of raw data

The method of raw data permutation does not suffer from the same problem as the Kennedy method. After each permutation the full model is applied, so the test of Y^π versus

Z is properly conditioned on the covariable X throughout the permutations. By conditioning we intend approximate conditioning, insofar as it is only the linear relationship of the response variable with the covariable X that is removed.

We can also consider the permuted residuals used in the Freedman and Lane method as $R^{\pi}_{Y\,|\,X} = Y^{\pi} - aX^{\pi}$. The numerator of (6) reduces to $(\sum Y^{\pi} R_{Z\,|\,X})^2$, because $\sum X R_{Z\,|\,X} = 0$. Now, replacing Y^{π} with $R^{\pi}_{Y\,|\,X} + aX^{\pi}$, and recalling that $\sum Y^{\pi 2} = \sum Y^2 = \sum R^2_{Y\,|\,X}(1 + a^2 \sum X^2 / \sum R^2_{Y\,|\,X})$, some algebra gives

$$r_M^2 = \frac{(r_K + gB_{\pi})^2}{(1+g^2)(1-C_{\pi}^2)}, \tag{9}$$

where

$$B_{\pi}^2 = \frac{\left(\sum X^{\pi} R_{Z\,|\,X}\right)^2}{\sum X^2 \sum R^2_{Z\,|\,X}}, \qquad C_{\pi}^2 = \frac{\left(\sum XY^{\pi}\right)^2}{\sum X^2 \sum Y^2} \quad \text{and} \quad g^2 = \frac{\left(\sum XY\right)^2}{\sum X^2 \sum R^2_{Y\,|\,X}}.$$

The role of $(1 - C_{\pi}^2)$ in (9) is directly analogous to the role of $(1 - A_{\pi}^2)$ in the Freedman and Lane method (i.e. providing complete conditioning on X throughout the permutations), where C_{π} is the correlation coefficient between Y^{π} and X. Consider also that the random variables r_K and B_{π} are each correlation coefficients between the variable pairs $(R_{Z\,|\,X}, R^{\pi}_{Y\,|\,X})$ and $(R_{Z\,|\,X}, X^{\pi})$, respectively. Subject to Condition C holding for $R_{Z\,|\,X}$, $R_{Y\,|\,X}$ and X, we can apply Theorem 2 to show that $\sqrt{n}\, r_K$ and $\sqrt{n}\, B_{\pi}$ behave asymptotically as independent standard normal variables. If Y also satisfies Condition C, we note that $C_{\pi} \xrightarrow{p} 0$; so from (9), $\sqrt{n}\, r_M \xrightarrow{d} \mathrm{N}(0, 1)$. The asymptotic distribution of r_M^2 under permutation is thus the same as the asymptotic distribution of r_F^2. This result means that, by virtue of the good qualities of the pivotal statistic, including conditioning on nuisance variables, the permutation of raw data gives an approximate test for non-zero partial regression coefficients.

3.3. Permutation under the full model

The method of ter Braak, like that of Freedman and Lane or Manly, maintains the conditioning on X throughout the permutations. The permuted residuals of the full model can be written $R^{\pi}_{Y\,|\,XZ} = R^{\pi}_{Y\,|\,X} - bR^{\pi}_{Z\,|\,X}$ where $b = \sum R_{Y\,|\,X} R_{Z\,|\,X} / \sum R^2_{Z\,|\,X}$. So replacing $R^{\pi}_{Y\,|\,XZ}$ in (7) and multiplying this out gives

$$r_T^2 = \frac{(r_K - rF_{\pi})^2}{(1-r^2)(1-G_{\pi}^2)}, \tag{10}$$

where $F_{\pi}^2 = \left(\sum R^{\pi}_{Z\,|\,X} R_{Z\,|\,X}\right)^2 / \left(\sum R^2_{Z\,|\,X}\right)^2$ and $G_{\pi}^2 = \left(\sum R^{\pi}_{Y\,|\,XZ} X\right)^2 / \left(\sum R^2_{Y\,|\,XZ} \sum X^2\right)$. Note that r in (10) is the value of the test statistic for the original data before permutation (2), used for all methods. Compare (10) with the result obtained for the method of raw data permutation in (9). The role of $(1 - G_{\pi}^2)$, where G_{π} is the correlation coefficient between $R^{\pi}_{Y\,|\,XZ}$ and X, is analogous to the role of $(1 - A_{\pi}^2)$ in the Freedman and Lane method and the role of $(1 - C_{\pi}^2)$ in the method of Manly: they ensure conditioning on the covariable X throughout the permutations. As in these previous analogous situations, if $R_{Y\,|\,XZ}$ satisfies Condition C, $G_{\pi} \xrightarrow{p} 0$. Furthermore, applying Theorem 2 as before, $\sqrt{n}\, r_K$ and $\sqrt{n}\, F_{\pi}$

TABLE 1

Correlations among the test statistics under permutation for a single set of simulated data. The expected correlations according to equation (11) are given first, and the observed correlations, calculated from 999 permutations are in parentheses.

	$\sqrt{n}\, r_E$	$\sqrt{n}\, r_F$	$\sqrt{n}\, r_M$
$\sqrt{n}\, r_F$	1.000 (0.999)		
$\sqrt{n}\, r_M$	0.820 (0.854)	0.820 (0.830)	
$\sqrt{n}\, r_T$	0.675 (0.666)	0.675 (0.668)	0.553 (0.547)

behave as standard normal variables with correlation r. Thus, from (10), $\sqrt{n}\, r_T \xrightarrow{d} \mathrm{N}(0, 1)$. Therefore, permutation of residuals under the full model gives a distribution under the null hypothesis that is asymptotically the same as that obtained by the Freedman and Lane method and the permutation of raw data. Like the latter, ter Braak's method relies on the qualities of the pivotal statistic as a ratio, including the conditioning on nuisance variables.

3.4. Correlations under permutation

Consider equations (8), (9) and (10). Using Theorem 2 we can show $\sqrt{n}(r_F, r_M, r_T)$ converges in distribution to a trivariate normal with means zero and covariance matrix

$$\begin{bmatrix} 1 & u & v \\ u & 1 & uv \\ v & uv & 1 \end{bmatrix}, \qquad \text{where } u = (1 + g^2)^{-1/2},\ v = (1 - r^2)^{1/2}. \tag{11}$$

So, even though the value of the test statistic for each method is not the same for any single permutation π, the distributions of the permuted statistics, r_F, r_M and r_T, converge to the same distribution, a standard normal. For the test by permutation, it is the distribution of the permuted values that matters, for it is against this that we scale the observed value of the test statistic, r^2, to obtain a P-value. So, all the tests (except the Kennedy method) work as approximate permutation tests for a partial regression coefficient. All the methods use the same observed value and they all produce permutation distributions under H_0 that converge to the same distribution, so the expected values of the probabilities obtained using the tests are asymptotically equivalent. The method of Freedman and Lane comes the closest to attaining the status of an exact test in the current framework. It is the only test to have an expected correlation of 1 with the exact test.

The expected and observed correlation matrices among the four statistics (Table 1) show that the Freedman and Lane method is indeed the closest to the exact test for each permutation π. In addition, all observed correlations correspond well to their expected values obtained using (11). To derive the tabulated values a single set of data was simulated with X and Z each chosen randomly from a uniform distribution on the interval (0, 3), $\beta_1 = \beta_2 = 1$ and errors ϵ'' were drawn randomly from a standard normal distribution. The linear model in equation (1) was then used to create observations Y and the sample size was $n = 40$.

4. Power of the tests

A few comments concerning the power of these approximate permutation tests are appropriate. Previous work by Hoeffding (1952), Robinson (1973), Vadeviloo (1983) and Hall

& Titterington (1989) is relevant. In the present case, for simplicity, consider the one-tailed test for H_0: $\beta_2 = 0$, with alternative H_1: $\beta_2 > 0$. Let r denote the observed value of the test statistic as in (2). Consider its value under permutation as

$$r_\pi = \frac{\sum R^{\pi}_{Y|X} R_{Z|X}}{\sqrt{\sum R^2_{Y|X} \sum R^2_{Z|X}}} \,,$$

the square root of r_K^2. Taking M random permutations, we have a subset $\{\pi_1, \ldots, \pi_M\}$ randomly sampled with (or without) replacement from $\{\pi_1, \ldots, \pi_P\}$, where P is the total number of possible permutations. Then $\{r_{\pi_1}, \ldots, r_{\pi_M}\}$ are the values of the statistic for this subset of permutations and $\{r^{(1)} \le r^{(2)} \le \cdots \le r^{(M)}\}$ is the ordered set of these values.

The test is: reject H_0 for $r \ge r^{(k)}$, where $k = M - [M\alpha]$, α is the *a priori* level of significance chosen for the test and $[M\alpha]$ denotes the largest integer less than or equal to $M\alpha$. We assume that α is fixed for the test and that $M \to \infty$ as $n \to \infty$, so $k/M \to 1 - \alpha$ as $n \to \infty$. The probability of rejecting H_0 when it is true (i.e. the size of the test) is

$$\mathrm{Pr}_0(r \ge r^{(k)}) = \mathrm{E}_{H_0}\big(\mathrm{Pr}(r \ge r^{(k)} \mid Y)\big) \,.$$

Similarly, the probability of rejecting H_0 when it is false (i.e. the power of the test) is

$$\mathrm{Pr}_1(r \ge r^{(k)}) = \mathrm{E}_{H_1}\big(\mathrm{Pr}(r \ge r^{(k)} \mid Y)\big) \,.$$

Theorem 1. *Assume that* $\{X\}$ *and* $\{R_{Z\,|\,X}\}$ *satisfy Condition C of the Appendix and that* $\mathrm{var}(Y) > 0$ *and* $\mathrm{E}|Y|^3 < \infty$. *If we consider a sequence of alternatives such that* $\sqrt{n}\,\mathrm{E}_{H_1}(r)$ *converges to a constant* ζ, *then the asymptotic power of the test is* $1 - \Phi(u_\alpha - \zeta)$, *where* $1 - \Phi(u_\alpha) = \alpha$ *and* Φ *is the standard normal cdf.*

Proof. We first show Condition A of Hoeffding (1952): $\sqrt{n} r^{(k)} \xrightarrow{p} u_\alpha$, where $|k/M - (1 - \alpha)| \le 1/M$. Our treatment differs from Hoeffding's in that we have a random subset M of all possible permutations P. As in the Monte Carlo approach outlined by Hall & Titterington (1989), let r_π denote a generic value r under permutation. Define $p_u = \mathrm{Pr}(\sqrt{n}\, r_\pi \le u \mid Y)$. Let $F_M(u \mid Y)$ be the empirical distribution function of $\{r_{\pi_1}, \ldots, r_{\pi_M}\}$; that is, the number of r_{π_i} less than or equal to u divided by M. Then $r^{(k)} \le u$ if and only if $F_M(u \mid Y) \ge k/M$, so

$$\mathrm{Pr}(r^{(k)} \le u \mid Y) = \mathrm{Pr}\big(MF_M(r \mid Y) \ge k\big) = \sum_{i=k}^{M} \binom{M}{k} p_u^i (1 - p_u)^{M-i} ,$$

which is an upper-tail probability of a binomial (M, p_u), conditional on Y.

Let A_n denote the set of $\{Y\}$ such that $\{R_{Y\,|\,X}\}$ and $\{R_{Z\,|\,X}\}$ satisfy Condition C of the Appendix. Then for $Y \in A_n$, from Theorem 2, given $\delta > 0$,

$$\big|\mathrm{Pr}(\sqrt{n}\, r_\pi \le u \mid Y) - \Phi(u)\big| < \delta$$

for large enough n. The set A_n is such that $\frac{1}{n} \sum R^2_{Y\,|\,X} > C$ and $\frac{1}{n} \sum |R_{Y\,|\,X}|^3 < C'$ for some constants C and C', so from the assumptions, using the weak law of large numbers and

Hölder's inequality we can show that, given $\varepsilon > 0$, $\Pr(Y \in A_n) > 1 - \varepsilon$ for large enough n. Then

$$\Pr\big(\big|\Pr(\sqrt{n}\,r_\pi \le u \mid Y) - \Phi(u)\big| < \delta\big) > 1 - \epsilon. \tag{12}$$

Now, if

$$F_M(u \mid Y) \xrightarrow{p} \Phi(u), \tag{13}$$

then for $u' < u_\alpha < u''$, $\Pr(\sqrt{n}\,r^{(k)} \le u' \,|\, Y) \to 0$ and $\Pr(\sqrt{n}\,r^{(k)} \le u'' \,|\, Y) \to 1$ so $\sqrt{n}\,r^{(k)} \xrightarrow{p} u_\alpha$. Next we show that (13) holds. From (12), $p(u) = \Pr(\sqrt{n}\,r_\pi \le u \,|\, Y) \xrightarrow{p} \Phi(u)$. Also because, given Y, $M F_M(u \,|\, Y)$ is binomial (M, p_u), $\Pr\big(\big|F_M(u \,|\, Y) - p_u\big| < \epsilon \,|\, Y\big) \xrightarrow{p} 1$. Also the conditional probabilities on the left are bounded by 1 so

$$\Pr\big(\big|F_M(u \mid Y) - p_u\big| < \epsilon\big) = \mathrm{E}\big[\Pr\big(\big|F_M(r \mid Y) - p_u\big| < \epsilon \mid Y\big)\big] \to 1,$$

and therefore (13) holds, which is what we require.

Finally, under H_1 and the assumptions of the theorem, $\sqrt{n}\,\mathrm{E}_{H_1} r \xrightarrow{p} \zeta$ and $n\,\mathrm{var}(r) \xrightarrow{p} 1$, so we can use the Lyapounov version of the central limit theorem to show that

$$\Pr(\sqrt{n}\,r \le u) \to \Phi(u - \zeta).$$

All the permutation tests considered here use r^2 for the observed test statistic (the two-tailed equivalent to the above one-tailed test). The values of the test statistics under permutation (i.e. r_K^2, r_F^2, r_M^2 and r_T^2) have all been shown to converge asymptotically to the same distribution and so all have comparable asymptotic power. These results are also supported by extensive empirical simulations (Anderson & Legendre, 1999).

5. Empirical simulations

Some empirical simulations additional to those provided by Anderson & Legendre (1999) are given here. In particular, we show the exaggerated effect of inflated type 1 error for Kennedy's method with increases in the number of covariables in the model. Data were simulated as follows: predictor variables were each chosen randomly and independently from a uniform distribution on the interval $(0, 3)$. The slope parameter for one of the variables (Z, the one under test) was set at zero, while the slope parameter for each of the other variables was set at 1. Errors ϵ'' were chosen independently and randomly from a standard normal distribution and the linear model was used to obtain Y. Ten thousand such datasets were generated and for each we obtained P-values using 999 permutations in the manner of Kennedy and of Freedman and Lane. The P-value for the t-test was also calculated for each dataset. The type 1 error for each method was recorded as the proportion of rejections of H_0 at significance level 0.05. This was done for $n = 10, 20, 30, 40, 50, 60$ and with the number of covariables $= 5, 10$. When the number of covariables was 10, the smallest sample size was $n = 12$ (rather than 10). The entire procedure was repeated with errors ϵ'' chosen from an $\exp(1)$ distribution raised to the third power.

Figure 1 shows the results, and inflated type 1 error is apparent for the Kennedy method, especially with small sample sizes. In the most extreme cases, type 1 error is as high as 60% for the Kennedy method. Note also that it is only for quite radically non-normal errors (i.e.

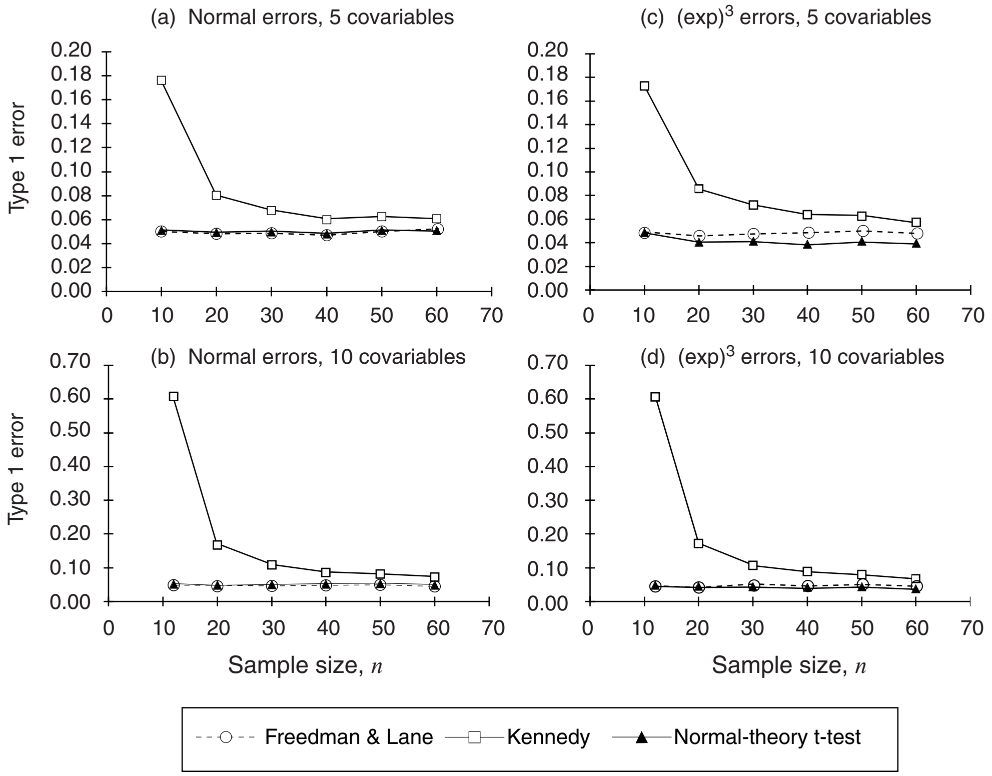

Figure 1. Results of empirical simulations comparing type 1 error for the methods of Freedman and Lane, Kennedy and the normal-theory t-test. The method of simulation of data is described in the text. The *a priori* significance level was 0.05. Each point represents the proportion of the number of rejections of the null hypothesis out of 10 000 simulations, each with 999 permutations.

$\exp(1)^3$) that the normal-theory t-test noticeably deviates from the *a priori* significance level of 0.05 (Figure 1c).

The simulation study by Anderson & Legendre (1999) suggests that the method of raw data permutation is prone to error in the manner suggested initially by Kennedy & Cade (1996). In the case of an extreme outlier in X with either normal or extremely non-normal errors, the type 1 error was de-stabilized for permutation of raw data, often being inflated (see Anderson & Legendre, 1999 Figure 7). However, the presence of such an extreme outlier is certainly a situation that violates the conditions of boundedness used in the proofs provided here. Mathematical proofs demonstrating the problem with Manly's method caused by an outlier in X would require an Edgeworth expansion of r_M^2 at least out to the fourth moment (examining kurtosis), which is not pursued here.

If the partial regression coefficient alone were used for the test (a non-pivotal statistic), then neither the method of raw data permutation nor permutation under the reduced model would work. This has been noted by others (Kennedy & Cade, 1996). We show explicitly that the slope coefficient under permutation of raw data would be affected by the added value of the non-zero fixed constant g defined in (9), thus inflating the values of the permuted statistics and decreasing the rejection rate. Table 2 shows results of simulations demonstrating this effect, where the *a priori* significance level of 0.05 is not achieved by either method when

TABLE 2

Type 1 error obtained using the pivotal t-statistic (equivalent to using r^2 or the F-statistic under permutation) or the non-pivotal slope coefficient (b_2) for either permutation of raw data (Raw) or permutation under a reduced model (Reduced)

No. of regressors	Sample size n	Raw (t)	Reduced (t)	Raw (b_2)	Reduced (b_2)
2	12	0.054	0.053	0.012	0.079
5	12	0.047	0.047	0.000	0.131
10	12	0.052	0.052	0.000	0.610
2	20	0.050	0.051	0.009	0.057
5	20	0.051	0.052	0.000	0.086
10	20	0.052	0.052	0.000	0.167
2	40	0.051	0.049	0.011	0.052
5	40	0.048	0.047	0.000	0.060
10	40	0.051	0.050	0.000	0.090

using the slope coefficient alone for the permutation test. Note that permutation under the reduced model using the slope coefficient alone as the test statistic is equivalent to using the Kennedy method. As the number of rejections under the true null hypothesis has a binomial distribution, the 95% confidence interval for 10 000 trials (simulations) is (0.046–0.054). Any values falling outside this range can be considered to differ significantly from the expected value of 0.05. The data were simulated as for Table 1 with the obvious generalization to multidimensional X when the number of regressors is 5 or 10.

Further simulations show that the method of ter Braak, in virtually all situations considered, gives results highly comparable with those obtained for the Freedman and Lane method (Anderson & Legendre, 1999). In particular, the issue of increased power, suggested by ter Braak (1992) as being a potential advantage of the method, does not seem to occur. Power curves consistently show results for the Freedman and Lane method and the ter Braak method of permutation as virtually identical: they are almost impossible to distinguish on the graphs (see Anderson & Legendre, 1999 Figure 6). The method of ter Braak also suffers somewhat from de-stabilization of type 1 error when the covariable X contains an outlier, but only in the presence of extremely non-normal errors and to a much lesser extent than for Manly's method. This effect also disappears with increases in the sample size (see Anderson & Legendre, 1999 Figure 8).

6. Discussion

The distributions of the statistics for these methods ($\sqrt{n}\, r_F$, $\sqrt{n}\, r_M$ and $\sqrt{n}\, r_T$) all converge asymptotically under permutation to a standard normal distribution. We have also demonstrated that the critical values for all these permutation tests converge to the same constant under sequences of contiguous alternative hypotheses, so they all have the same asymptotic power.

In empirical studies, the Freedman & Lane (1983) method generally gives the best results (in terms of type 1 error or power; Anderson & Legendre, 1999) and the theoretical results here demonstrate why this is so. It estimates what would be an exact test if the relationship between the response variable Y and the concomitant variable X were known by creating alternative possible observed values under the null hypothesis that are completely conditional on X through permutation of errors ϵ. None of the other methods approximates the exact test.

We have shown that the reason for the inflated type 1 error with the Kennedy method (1995) is a lack of appropriate conditioning on the nuisance variable X throughout the permutations. The lack of conditioning causes the estimated variance of the partial regression coefficient being tested to be too large under permutation, causing inflated type 1 error. This is a consistent error that becomes worse with increases in the number of nuisance variables in the model.

The methods of Manly (1997) and ter Braak (1992) do not suffer from this error of a lack of complete conditioning on the nuisance variable(s). For each of these methods, as for that of Freedman and Lane, X and Z are kept fixed and are included in the partial regression done for each permutation. Neither of these methods, however, follows the rationale of an exact test. Instead, we have demonstrated how they rely on the asymptotically pivotal test statistic under permutation. By asymptotically pivotal in the present context, we mean that the test statistic used has a distribution that, asymptotically, does not rely on any unknown parameters and thereby adjusts for any nuisance parameters that are not of interest for the test. The importance of using a pivotal statistic has been discussed in the context of bootstrapping for constructing confidence limits and tests (e.g. Hall & Titterington, 1989; Fisher & Hall, 1990), but has not been considered fully in the present context of permutation tests for partial regression.

Unlike Freedman and Lane's method, these two methods can have problems when the covariable X contains an extreme outlier. In that case, the asymptotic statements concerning the equivalence of the methods begin to lose their meaning as the conditions of Theorem 2 no longer apply. However, such extreme situations do not adversely affect the Freedman and Lane permutation method.

Although we have considered only the two-tailed tests, the same arguments apply to one-tailed tests of partial slope coefficients using these methods. The results obtained also extend to the general situation with (i) greater numbers of covariables or nuisance variables in the model, or (ii) greater numbers of predictor variables being tested simultaneously under the null hypothesis.

In addition, we note that these results have special relevance for the case of multivariate data (i.e. multiple response variables). For univariate analysis, the normal-theory tests are fairly robust and, in situations where this is in doubt, appropriate transformations of the raw data can usually be found. Traditional multivariate tests, however, are not so robust to departures from non-normality, and permutation tests now abound for non-parametric analysis of multivariate data, particularly in the biological and ecological sciences, especially for tests based on distance matrices (e.g. Mantel, 1967; Mielke, Berry & Johnson, 1976; Smouse, Long & Sokal, 1986; Clarke, 1993). It was for multivariate canonical analysis of ecological and agricultural data that ter Braak (1992) first suggested his permutational approach. Smouse *et al.* (1986) considered an extension of the simple Mantel test (1967) to a partial Mantel test, suggesting a permutational strategy equivalent to the method proposed later by Kennedy (1995) for such tests. We consider that the use of Kennedy's method with multivariate distance matrices will also suffer with inflated type 1 error, due to lack of conditioning on nuisance terms, as for the univariate case. All our results can be readily extended to the case of multivariate response variables. Indeed, in the above, it can be considered that Y is a matrix rather than a vector, and that correspondingly entire rows of this matrix (or matrices of appropriate residuals) are permuted for the tests, and that the products indicate vector products or matrix multiplication (depending on the context).

Appendix

In this appendix we state a multivariate limit theorem that uses conditions which are simpler than those of Wald & Wolfowitz (1944) and which imply the condition of Hájek (1961) in the multivariate case. We use a Cramér–Wold device to get the multivariate version as in, e.g., Fraser (1957, Theorem 6.3 p. 240).

Condition C. *For the triangular array*

$$B = \{b_{nj},\ j = 1, \ldots, n;\ n = 1, 2, \ldots\}$$

there exist positive constants C and C' such that for each $n = 1, 2, \ldots$

$$\frac{1}{n}\sum_{j=1}^{n} b_{nj}^2 > C \qquad \textit{and} \qquad \frac{1}{n}\sum_{j=1}^{n} |b_{nj}|^3 < C'.$$

Theorem 2. *Consider the arrays $\{a_{nj}\}, \{c_{n1j}\}, \ldots, \{c_{nmj}\}$ which are such that $\bar{a}_n = \bar{c}_{n1} = \cdots = \bar{c}_{nm} = 0$. Let $(\pi_1, \ldots, \pi_n)$ be a random equiprobable permutation of $(1, \ldots, n)$ and define, for each $n = 1, 2, \ldots$,*

$$S_{nk} = \frac{\sqrt{n}\sum_{j=1}^{n} a_{nj} c_{nk\pi_j}}{\sqrt{\sum_{j=1}^{n} a_{nj}^2 \sum_{j=1}^{n} c_{nkj}^2}} \qquad (k = 1, \ldots, m).$$

Then if $\{a_{nj}\}, \{c_{n1j}\}, \ldots, \{c_{nmj}\}$ all satisfy Condition C,

$$(S_{n1}, \ldots, S_{nm}) \xrightarrow{d} \mathrm{N}_m(\mathbf{0}, \boldsymbol{R}), \quad \textit{as} \quad n \to \infty,$$

where $\boldsymbol{R} = [R_{k\ell}]$ and

$$R_{k\ell} = \frac{\sum_{j=1}^{n} c_{nkj} c_{n\ell j}}{\sqrt{\sum_{j=1}^{n} c_{nkj}^2 \sum_{j=1}^{n} c_{n\ell j}^2}}.$$

In the main text, we omit the first subscript n throughout.

References

Anderson, M.J. & Legendre, P. (1999). An empirical comparison of permutation methods for tests of partial regression coefficients in a linear model. *J. Statist. Comput. Simulation* **62**, 271–303.

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

Clarke, K.R. (1993). Nonparametric multivariate analysis of changes in community structure. *Aust. J. Ecol.* **18**, 117–143.

Collins, M.F. (1987). A permutation test for planar regression. *Aust. J. Statist.* **29**, 303–308.

Edgington, E.S. (1995). *Randomization Tests*, 3rd edn. New York: Marcel Dekker.

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

Fisher, R.A. (1935). *Design of Experiments.* Edinburgh: Oliver & Boyd.

Fraser, D.A.S. (1957). *Nonparametric Methods in Statistics.* New York: John Wiley & Sons.

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

Gail, M.H., Tan, W.Y. & Piantadosi, S. (1988). Tests for no treatment effect in randomized clinical trials. *Biometrika* **75**, 57–64.

HALL, P. & TITTERINGTON, D.M. (1989). The effect of simulation order on level accuracy and power of Monte Carlo tests. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **51**, 459–467.

HÁJEK, J. (1961). Some extensions of the Wald–Wolfowitz–Noether theorem. *Ann. Math. Statist.* **32**, 506–523.

HOEFFDING, W. (1952). The large-sample power of tests based on permutations of the observations. *Ann. Math. Statist.* **23**, 169–192.

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

KENNEDY, P.E. & CADE, B.S. (1996). Randomization tests for multiple regression. *Comm. Statist. Simulation Comput.* **25**, 923–936.

MANLY, B.F.J. (1997). *Randomization, Bootstrap and Monte Carlo Methods in Biology*, 2nd edn. London: Chapman & Hall.

MANTEL, N. (1967). The detection of disease clustering and a generalized regression approach. *Cancer Res.* **27**, 209–220.

MIELKE, P.W., BERRY, K.J. & JOHNSON, E.S. (1976). Multi-response permutation procedures for *a priori* classifications. *Comm. Statist. Theory Methods* **5**, 1409–1424.

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

PITMAN, E.J.G. (1937a). Significance tests which may be applied to samples from any populations. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **4**, 119–130.

PITMAN, E.J.G. (1937b). Significance tests which may be applied to samples from any populations. II. The correlation coefficient test. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **4**, 225–232.

PITMAN, E.J.G. (1937c). Significance tests which may be applied to samples from any populations. III. The analysis of variance test. *Biometrika* **29**, 322–335.

ROBINSON, J. (1973). The large-sample power of permutation tests for randomization models. *Ann. Statist.* **1**, 291–296.

SMOUSE, P.E., LONG, J.C. & SOKAL, R.R. (1986). Multiple regression and correlation extensions of the Mantel test of matrix correspondence. *System. Zool.* **35**, 627–632.

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

VADEVILOO, J. (1983). On the theory of modified randomization tests for nonparametric hypotheses. *Comm. Statist. Theory Methods* **12**, 1581–1596.

WALD, A. & WOLFOWITZ, J. (1944). Statistical tests based on permutations of the observations. *Ann. Math. Statist.* **15**, 358–372.

WELCH, W. J. (1990). Construction of permutation tests. *J. Amer. Statist. Assoc.* **85**, 693–698.