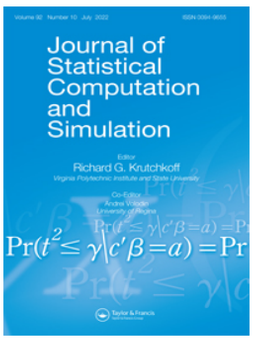

Taylor & Francis
Taylor & Francis Group

Journal of Statistical Computation and Simulation

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/gscs20

Predictor versus response permutation for significance testing in weighted regression and redundancy analysis

Cajo J. F. ter Braak

To cite this article: Cajo J. F. ter Braak (2022) Predictor versus response permutation for significance testing in weighted regression and redundancy analysis, Journal of Statistical Computation and Simulation, 92:10, 2041-2059, DOI: 10.1080/00949655.2021.2019256

To link to this article: https://doi.org/10.1080/00949655.2021.2019256

JOURNAL OF STATISTICAL COMPUTATION AND SIMULATION
2022, VOL. 92, NO. 10, 2041–2059
https://doi.org/10.1080/00949655.2021.2019256

Taylor & Francis
Taylor & Francis Group

Predictor versus response permutation for significance testing in weighted regression and redundancy analysis

Cajo J. F. ter Braak

Biometris, Wageningen University & Research, Wageningen, the Netherlands

ABSTRACT

In testing an overall null hypothesis, it does not matter whether to permute the response variables (Y) while keeping the predictors fixed or to permute the predictor variables (X) while keeping the response variables fixed. However, in weighted (univariate and multivariate) regression and in partial tests these options yield different results. This paper defines and evaluates by simulation ordinary and standardized versions of X- and Y-permutation for tests which encompass the existing Double Semi-Partialling, here called Collins-Dekker, and Freedman-Lane permutation methods. When the error variance is inversely proportional to the weights, the standardized permutation methods (which use standardized residuals) are most powerful, as expected, but otherwise they can be extremely liberal. In contrast, ordinary X-permutation (which permutes the residuals of X given any covariates 'as is') is by far the most robust against variability in the weights and against the error variance-weight relationship and is thus recommended for general use.

ARTICLE HISTORY
Received 31 August 2021
Accepted 13 December 2021

KEYWORDS
Permutation test; weighted regression; redundancy analysis; weighted log-ratio analysis; canonical correspondence analysis

1. Introduction

Permutation testing for statistical significance avoids the assumption of normality that is needed for the validity of the usual *t*- and *F*-tests in linear regression. Multivariate normality is needed in multivariate regression and is particularly difficult to ascertain in wide data sets, with a high number of response variables, in which case redundancy analysis [1] is often used [2]. Redundancy analysis (RDA) is also known as least-squares reduced rank regression [3,4] or principal components analysis with respect to instrumental variables [5,6]. In testing an overall null hypothesis (no effect of the predictors on the response variables), it does not matter whether the rows of the response variables (Y) are permuted, while keeping the values of the predictors fixed or the rows of the predictor variables (X) are permuted, while keeping the values of the response variables fixed. However, in weighted regression these options yield surprisingly different results. Their results may also differ in partial tests, i.e. testing one (set of) predictors X in the presence of another set of predictors Z (here called covariates to distinguish them from the tested predictor(s)). In partial

CONTACT Cajo J. F. ter Braak cajo.terbraak@wur.nl Droevendaalsesteeg 1, 6708 PB Wageningen, the Netherlands

Supplemental data for this article can be accessed here. https://doi.org/10.1080/00949655.2021.2019256

Table 1. Definition of four different permutation methods in weighted regression and redundancy analysis based on their analysis using weighted least-squares (WLS). The four methods are the cross-classification of Y-permutation (using Freedman-Lane) and X-permutation (using Collins-Dekker) with permutation of ordinary or of standardized residuals of the WLS regression and redundancy analysis.

WLS-residuals	Y-permutation Freedman-Lane[a]	X-permutation Collins-Dekker[b]
Standardized	$\mathrm{W}^{-1/2}\pi(\mathrm{W}^{1/2}\mathrm{R}_\mathrm{Z}^w\mathrm{Y}) + \mathrm{H}_\mathrm{Z}^w\mathrm{Y}\tilde{\sim}\mathrm{X} + \mathrm{Z}$	$\mathrm{Y}\tilde{\sim}\mathrm{W}^{-1/2}\pi(\mathrm{W}^{1/2}\mathrm{R}_\mathrm{Z}^w\mathrm{X}) + \mathrm{Z}$
Ordinary	$\pi(\mathrm{R}_\mathrm{Z}^w\mathrm{Y}) + \mathrm{H}_\mathrm{Z}^w\mathrm{Y}\tilde{\sim}\mathrm{X} + \mathrm{Z}$	$\mathrm{Y}\tilde{\sim}\pi(\mathrm{R}_\mathrm{Z}^w\mathrm{X}) + \mathrm{Z}$

Notes: [a]Freedman and Lane (9); [b]Collins (10) and Dekker et al. (2003, 2007). $\pi(\cdot)$ is a permutation of rows of the argument, $\mathrm{H}_\mathrm{Z}^w = \mathrm{Z}(\mathrm{Z}'\mathrm{WZ})^{-1}\mathrm{ZW}$ is the hat matrix in WLS, $\mathrm{R}_\mathrm{Z}^w = \mathrm{I} - \mathrm{H}_\mathrm{Z}^w$ is the residual-forming matrix of the WLS-regression on **Z**, and $\cdot\tilde{\sim}\cdot$ indicates that the left-hand argument is regressed using WLS with weights **w** using the model formula on the right-hand side. The test-statistic is the *F*-ratio in both methods, but could be replaced in Collins-Dekker by the regression sum of squares due to **X** after fitting **Y** to **Z**. In computational shortcuts, $\mathrm{H}_\mathrm{Z}^w\mathrm{Y}$ in Freedman-Lane can be omitted from the response matrix as this is a linear combination of **Z** that has no effect on the *F*-ratio, and **Y** in Collins-Dekker can be replaced by $\mathrm{R}_\mathrm{Z}^w\mathrm{Y}$ for the same reason.

tests, there is the option [7,8] either to permute residuals of the response variables of the model $\mathrm{Y} \sim \mathrm{Z}$ [9] or to permute the residuals of the predictor(s) of the model $\mathrm{X} \sim \mathrm{Z}$ [10–12], which we call here Freedman-Lane and Collins-Dekker, respectively. The Freedman-Lane method was shown to be the best among rival Y-permutation methods by Anderson and Robinson [13]. Based on extensive simulations of eight methods of permutation, Winkler *et al.* [7] recommended Freedman-Lane and the Smith method which we call here Collins-Dekker in view of priority. In a weighted regression that is fitted by weighted least-squares (WLS), there is then the option to either permute the ordinary residuals or the standardized residuals. In the former, the residuals of the weighted regression are permuted 'as is', whereas in the latter, the WLS-residuals are standardized before permutation by multiplication by the square-root of the weights and de-standardized (by division by the square-root of the weights) after permutation (Table 1). This standarization is motivated by the fact that, if the weights are inversely proportional to the error variance, the errors standardized in this way would have equal variance. The description, evaluation and comparison of these options is the topic of this paper.

Weighted regression is called for in the situation that the error variance is non-constant. Ideally, weights should then be chosen inversely proportional to the error variance. In a regression with responses that are means of different samples sizes, the sample size is the optimal weight. However, the error variance may not be known precisely, due to other sources of error. If the error variance (up to proportionality) is uncertain, is it then prudent to use this imperfect knowledge, or is it safer, e.g. in terms of type I error rate, to ignore it? Rather than following from the error variance, weights may alternatively be key to the method. This can be illustrated with two multivariate examples.

The first example is canonical correspondence analysis [14], a method that is widely used in community ecology to relate a cases-by-species abundance table to environmental (predictor) variables [15]. But, it can be carried out by weighted RDA of transformed data, in which the row weights are the row totals of the non-negative response table [2,16]. These weights can be understood, as in correspondence analysis [17], in terms of the variance of the contingency ratio obtained from independent Poisson counts, but not so when the counts are overdispersed. One might then consider to change the weights so as to improve

Table 2. Short-cut formulas for the four different permutation methods of Table 1 after transformation of the WLS-problem to ordinary least-squares (OLS).

Transformation from WLS to OLS	$\tilde{Y} = W^{1/2}Y$ $\tilde{X} = W^{1/2}X$ and $\tilde{Z} = W^{1/2}Z$	
WLS-residuals	Y-permutation Freedman-Lane[a]	X-permutation Collins-Dekker[b]
Standardized	$\pi(R_{\tilde{Z}}\tilde{Y}) \sim \tilde{X} + \tilde{Z}$	$\tilde{Y} \sim \pi(R_{\tilde{Z}}\tilde{X}) + \tilde{Z}$
Ordinary	$W^{1/2}\pi(W^{-1/2}R_{\tilde{Z}}\tilde{Y}) \sim \tilde{X} + \tilde{Z}$	$\tilde{Y} \sim W^{1/2}\pi(W^{-1/2}R_{\tilde{Z}}\tilde{X}) + \tilde{Z}$

Notes: [a]Freedman and Lane (9); [b]Collins (10) and Dekker et al. (2003, 2007). $\pi(\cdot)$ is a permutation of rows of the argument, $R_{\tilde{Z}} = I - \tilde{Z}(\tilde{Z}'\tilde{Z})^{-1}\tilde{Z}$ is the residual-forming matrix in OLS-regression on $\tilde{Z}$, and $\cdot \sim \cdot$ indicates here that the left-hand argument is regressed using OLS using the model formula on the right-hand side. In Collins-Dekker, $\tilde{Y}$ can be replaced by $R_{\tilde{Z}}\tilde{Y}$, but this does not reduce the computational costs as the sum of squares due to $\tilde{Z}$ is fixed during permutation.

the inverse relation with variance, but this then gives a method that no longer has the properties that are key to canonical correspondence analysis, namely that it is an approximation to maximum likelihood fitting of a Gaussian response model [14], that it fits an unfolding model [15,18] and maximizes 'niche separation' [19].

The second example is weighted log-ratio RDA which is the weighted least-squares regression form of log-ratio principal component analysis [20], a dimension reduction method for compositional data. The role of the weights is to ensure that the method is not only 'subcompositional coherent', a cornerstone of log-ratio analysis [21], but also shows 'distributional equivalence', which is a cornerstone of correspondence analysis [22]. Weights are useful as well in log-ratio RDA. But how should the effect of a set of predictors on the composition be tested in weighted log-ratio RDA, with Y-permutation using Freedman-Lane or X-permutation using Collins-Dekker and should it be with ordinary or standardized residuals?

Note that permutation of standardized residuals is implicit, if the permutations are carried out after transformation of the WLS-problem issued by the weighted regression to an ordinary least-squares problem (OLS) and the weighs are further disregarded after the transformation (Table 2). The transformation is obtained by multiplying both Y and X (and Z) by the square-root of the case weights, which is a standard way of solving a WLS-problem [23].

The null hypothesis in this paper is that the predictors have no effect on the distribution of the response variables after taking account for the effects of covariates. Detected effects may thus either be an effect on the mean response or on the variance (or weight) or on both. Thus, this paper does not consider Behrens-Fisher type problems in which the more restricted null hypothesis of an effect on the mean response only is tested. For permutation tests of this more restricted null hypothesis, see DiCiccio and Romano [24] and Helwig [25]. The null hypothesis in this paper is: the predictors have neither an effect on the distribution of the responses nor on the distribution of the weight. The weights were therefore drawn independently of the predictors in our simulations, except for a simulation to study sensitivity of the tests to the correlation of the weight with the predictors. Note also that, with covariates, the permutation tests are only approximately valid [13].

Section 2 describes the various permutation methods and Section 3 illustrates the issues in the simplest possible way with real ecological data. Section 4 evaluates and compares the methods by simulation, featuring three scenarios (simple regression, multiple regression

with model y ~ x + z and reduced rank regression) with error variance scenarios that range from variance inversely-proportional-to-the-weight to constant variance, all analyzed by Monte Carlo permutation with and without weights, i.e. by WLS and OLS. In all scenarios the type I error rate and power to detect the effect of x are investigated; in the reduced rank scenario the rank is investigated by testing each axis conditional on the earlier axes [26]. Section 5 provides discussion and conclusions.

2. Permutation methods for significance tests

This section describes the fitting of the multivariate linear model by WLS to data and testing the effect of one set of predictors in the presence of a set of covariates using Monte Carlo significance tests. While summarized in Tables 1 and 2, the two versions of Freedman-Lane and of Collins-Dekker are presented in full in Sections 2.3 and 2.4, respectively, with their relation with raw data permutation for testing the overall null hypothesis (Section 2.6). Computational aspects are discussed in Section 2.5. Examples using R-code are in Appendix S1.

2.1. Weighted least-squares fitting and testing in the linear model

The starting point is the multivariate linear model

$$\mathbf{Y} = \mathbf{ZA} + \mathbf{XB} + \mathbf{E} \tag{1}$$

with $\mathbf{Y}$, $\mathbf{Z}$ and $\mathbf{X}$ real matrices with n rows (cases) containing, as columns, m response variables, q covariates and p predictor variables, respectively. $\mathbf{E}$ is an $n \times m$ matrix of errors with zero expectation, $E(\mathbf{E}) = 0_{n\times m}$, and $\mathbf{A}$ and $\mathbf{B}$ are $q \times m$ and $p \times m$ matrices of unknown regression parameters, respectively, which need to be estimated from the data $\mathbf{Y}$, $\mathbf{Z}$ and $\mathbf{X}$. The number of columns of each is at least 1; $\mathbf{Z}$ is assumed to contain a column of ones for the intercept. This model encompasses multiple regression when $m = 1$ and simple regression when $m = q = p = 1$ and $\mathbf{Z} = \mathbf{1}_n$, so that $\mathbf{A}$ is the intercept and $\mathbf{B}$ is the slope parameter with respect to the only predictor in $\mathbf{X}$. The task is to test the effect of the predictors ($\mathbf{X}$) on the responses ($\mathbf{Y}$) after accounting for (in the presences of) the covariates ($\mathbf{Z}$), i.e. the null hypothesis is $\mathbf{B} = 0_{p\times m}$, in the context of WLS, i.e. when estimation and testing are carried out with weights $\mathbf{w} = w_1, \ldots, w_n$ for the n cases. Dekker et al. [12] take a similar starting point 'for data organized in square matrices of relatedness among n objects', but without weights. Equation (1) is also denoted by the model formula $\mathbf{Y} \dot{\sim} \mathbf{Z} + \mathbf{X}$, where the dot above the usual tilde is added as a reminder that the model is fitted using WLS with weights $\mathbf{w}$. To highlight the effect of $\mathbf{X}$ given $\mathbf{Z}$, it is also denoted by $\mathbf{Y} \dot{\sim} \mathbf{X}|\mathbf{Z}$.

To obtain a significance test, both the null and the alternative model must be fitted to the data. Under the null hypothesis, $\mathbf{Y} = \mathbf{ZA} + \mathbf{E}$, the WLS estimates of the regression coefficients $\mathbf{A}$ are

$$\hat{\mathbf{A}}_0 = (\mathbf{Z}'\mathbf{W}\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{W}\mathbf{Y} \text{ with } \mathbf{W} = diag(w_1, \ldots, w_n), \tag{2}$$

giving fitted values $\hat{\mathbf{Y}}_0 = \mathbf{Z}\hat{\mathbf{A}}_0$, residuals $\hat{\mathbf{E}}_0^Y = \mathbf{Y} - \hat{\mathbf{Y}}_0$ and weighted residual sum of squares $RSS_0 = trace(\hat{\mathbf{E}}_0^{Y\prime}\mathbf{W}\hat{\mathbf{E}}_0^Y)$. Under the alternative hypothesis (Equation (1)), $\mathbf{Z}$ in Equation (2) must be replaced by $\mathbf{Z}_a = (\mathbf{Z} : \mathbf{X})_{n\times(q+p)}$, the horizontal concatenation of

the columns of $\mathbf{Z}$ and $\mathbf{X}$, to give fitted values $\hat{\mathbf{Y}}_a = (\mathbf{Z} : \mathbf{X})(\hat{\mathbf{A}} : \widehat{\mathbf{B}})' = \mathbf{Z}\hat{\mathbf{A}} + \mathbf{X}\hat{\mathbf{B}}$ and residuals $\hat{\mathbf{E}}_a^Y = \mathbf{Y} - \hat{\mathbf{Y}}_a$ and weighted residual sum of squares $RSS_a = trace(\hat{\mathbf{E}}_a^{Y'}\mathbf{W}\hat{\mathbf{E}}_a^Y)$. The test-statistic, the F-ratio

$$F = [(RSS_0 - RSS_a)/p]/[RSS_a/(n - p - q)], \tag{3}$$

is likely to be closer to asymptotic pivotality than $RSS_0 - RSS_a$, which is desirable in Freedman-Lane [13]. In multiple regression and normal error with constant variance and equal weights, the F-ratio follows an F-distribution with p and $n - p - q$ degrees of freedom. If $m > 1$, other test statistics may be considered (e.g. a stacked t^2-ratio), but Equation (3) is often used in RDA implementations. Another test statistic that is often used, is based on the first eigen value of a principal components analysis of the fitted values of the model $\mathbf{Y} \sim \mathbf{X}$, when there are no nontrivial covariates and, in general, of the model $\hat{\mathbf{E}}_0^Y \dot{\sim} \hat{\mathbf{E}}^X$ (i.e. the response and predictor matrices after the covariates are partialled out) [26]. This test statistic, F_{eig}, is like Equation (3), but with RSS_a replaced by the residual sums of squares of the rank 1 model. It is often used in establishing the rank of the model. For this, each axis is tested after including the earlier axes as covariates [26]. The so-calculated P-values for subsequent axes are adjusted for multiple testing by replacing them by their cumulative maximum [27]. This is a sequential and closed testing procedure that guarantees that an axis is judged significant only when the earlier axes are judged significant [27].

The focus of this paper is on permutation tests using these F-ratios to produce a statistical test that has the nominal type I error rate (size) and, preferably, high power.

From Equation (2) and the formulas leading to Equation (3), it can be seen that the same value of the F-ratio of Equation (3) can be obtained by OLS, i.e. by unweighted regression and RDA, by a transformation of the data to $\tilde{\mathbf{Y}} = \mathbf{W}^{1/2}\mathbf{Y}$, $\tilde{\mathbf{X}} = \mathbf{W}^{1/2}\mathbf{X}$ and $\tilde{\mathbf{Z}} = \mathbf{W}^{1/2}\mathbf{Z}$. The OLS-analysis of this data gives fitted values $\widehat{\tilde{\mathbf{Y}}}_0 = \mathbf{W}^{1/2}\hat{\mathbf{Y}}_0$, $\widehat{\tilde{\mathbf{Y}}}_a = \mathbf{W}^{1/2}\hat{\mathbf{Y}}_a$ and residuals

$$\widehat{\tilde{\mathbf{E}}}_0^Y = \tilde{\mathbf{Y}} - \widehat{\tilde{\mathbf{Y}}}_0 = \mathbf{W}^{1/2}\hat{\mathbf{E}}_0^Y \text{ and } \widehat{\tilde{\mathbf{E}}}_a^Y = \tilde{\mathbf{Y}} - \widehat{\tilde{\mathbf{Y}}}_a = \mathbf{W}^{1/2}\hat{\mathbf{E}}_a^Y. \tag{4}$$

under the null and alternative hypothesis, respectively. Freedman-Lane uses the residuals $\hat{\mathbf{E}}_0^Y$ and $\widehat{\tilde{\mathbf{E}}}_0^Y$ from the models $\mathbf{Y}\dot{\sim}\mathbf{Z}$ and $\tilde{\mathbf{Y}} \sim \tilde{\mathbf{Z}}$, respectively.

For Collins-Dekker we need the residuals of the linear model

$$\mathbf{X} = \mathbf{Z}\mathbf{C} + \mathbf{E}^X \tag{5}$$

with $\mathbf{E}^X$ an $n \times p$ matrix of errors with zero expectation, $E(\mathbf{E}^X) = 0_{n\times p}$, and $\mathbf{C}$ is a $p \times q$ matrix of unknown regression parameters, which are estimated from the data $\mathbf{X}$ and $\mathbf{Z}$ using weighted regression with weights $w_1, \ldots, w_n$, giving WLS estimates $\hat{\mathbf{C}} = (\mathbf{Z}'\mathbf{W}\mathbf{Z})^{-1}\mathbf{Z}'\,\mathbf{W}\mathbf{X}$ and residuals $\hat{\mathbf{E}}^X = \mathbf{X}–\mathbf{Z}\hat{\mathbf{C}}$. Transformation to an unweighted regression of $\mathbf{W}^{1/2}\mathbf{X}$ on $\mathbf{W}^{1/2}\mathbf{Z}$ gives the standardized residuals $\widehat{\tilde{\mathbf{E}}}^X = \mathbf{W}^{1/2}\hat{\mathbf{E}}^X$.

2.2. Permutation algorithm for testing statistical significance

The procedure of a Monte Carlo permutation test is as follows.

(1) Calculate the test statistic for the data, leading to the value denoted by F_0.

(2) Generate K new data sets that are equally likely under the null hypothesis. In this paper, new data sets are generated using either Freedman-Lane or Collins-Dekker.
(3) Calculate the test statistic for each new data set, leading to values $F_1, F_2, \dots, F_K$.
(4) Calculate the Monte Carlo significance level by placing F_0 among $F_1, F_2, \dots, F_K$ and calculating the proportion of values greater than or equal to F_0, i.e. $\#(F_k \geq F_0, k = 0,1,2, \dots, K)/(K+1)$.

P-values should never be zero [28]. Step 4 guarantees this and leads to an exact test of the overall null hypothesis in the unweighted situation with independent identically distributed errors for the cases, even for small K, e.g. 19, 99 [29]. The simplest rationale for step 4 is that, under the overall null hypothesis with equal weights, the cases are exchangeable (nothing distinguishes the cases) so that the data as observed and its test statistic F_0 are equally likely as each new data set and its test statistic F_k. However, with differential weights, the cases are no longer exchangeable so that the Monte Carlo permutation test can only be approximate. It is explored in the next sections, which ways of performing the test have been proposed and which way might then be best.

2.3. Y-permutation (Freedman-Lane)

This section describes the ordinary and standardized versions of the Freedman and Lane [9] method of Y-permutation, which was singled out as the best one out of several other Y-permutation versions by Anderson and Robinson [13]. An early rationale for Freedman-Lane stems from Kempthorne (30: section 8.2) [30] who proposed a randomization model for the analysis of randomized experiments in which a fixed plot error ($\mathbf{E}$) is randomly assigned to treatments ($\mathbf{X}$).

In ordinary Freedman-Lane, each new data set in the permutation procedure contains the original $\mathbf{Z}$ and $\mathbf{X}$, but the response matrix is changed from $\mathbf{Y}$ to

$$\mathbf{Y}^{\pi} = \hat{\mathbf{Y}}_0 + \pi(\hat{\mathbf{E}}_0^Y) \tag{6}$$

with $\pi(\cdot)$ a random permutation of the rows of a matrix. The new data set is thus the sum of the fitted values plus a random row permutation of the residuals, both under the null model, and the F-ratio is obtained therefrom by WLS.

In standardized Freedman-Lane, each new data set in the permutation procedure contains the transformed covariate and predictor data $\tilde{\mathbf{Z}} = \mathbf{W}^{1/2}\mathbf{Z}$ and $\tilde{\mathbf{X}} = \mathbf{W}^{1/2}\mathbf{X}$, but the transformed response matrix $\tilde{\mathbf{Y}} = \mathbf{W}^{1/2}\mathbf{Y}$ is changed to

$$\tilde{\mathbf{Y}}^{\pi} = \widehat{\tilde{\mathbf{Y}}}_0 + \pi(\widehat{\tilde{\mathbf{E}}}_0^Y), \tag{7}$$

the sum of the fitted values plus a random row permutation of the residuals, both under the null model after data transformation to an equal-weights-situation and the F-ratio is obtained therefrom by OLS. The corresponding WLS version can be obtained by back-transformation of Equation (7), giving

$$\mathbf{W}^{-1/2}\tilde{\mathbf{Y}}^{\pi} = \mathbf{W}^{-1/2}\widehat{\tilde{\mathbf{Y}}}_0 + \mathbf{W}^{-1/2}\pi(\widehat{\tilde{\mathbf{E}}}_0^Y) = \hat{\mathbf{Y}}_0 + \mathbf{W}^{-1/2}\pi(\mathbf{W}^{1/2}\hat{\mathbf{E}}_0^Y), \tag{8}$$

the sum of the original fit $\hat{\mathbf{Y}}_0$ and a reweighted version of the permuted transformed residuals $\hat{\mathbf{E}}_{w0}^Y$. The F-ratio is then obtained by WLS applied to the new data set $(\mathbf{W}^{-1/2}\mathbf{Y}_w^{\pi}, \mathbf{Z}, \mathbf{X})$.

Equation (8) shows that standardized Freedman-Lane weighs the original residuals $\hat{\mathbf{E}}_0^Y$ by the square-root of ratios of the weights of cases. If weight is inversely proportional to the variance, the three elements in the last term in Equation (8), $\mathbf{W}^{1/2}\hat{\mathbf{E}}_0^Y$, $\pi(\cdot)$ and $\mathbf{W}^{-1/2}$, reflect the three logical steps 'standardize the residuals, permute and de-standardize them'.

There is little to motivate ordinary Freedman-Lane when the weights appear for a different reason in the analysis. The term 'residualized response permutation' will therefore refer to standardized Freedman-Lane only. The standardized version, which equals, of course, the ordinary one for equal weights, appears closest in spirit to the Freedman and Lane procedure.

2.4. X-permutation (Collins-Dekker)

This section describes the ordinary and standardized versions of the DSP [12] method of X-permutation. DSP is identical to the algorithm derived in regression context by Collins [10] from an early proposal by Oja [31] and differs from raw X-data permutation [7,32] in permuting the residuals from the model $\mathrm{X} \sim \mathrm{Z}$ instead of permuting X itself. The advantage of this is that the F-ratio of each newly created data set examines the same contrast (specified by the projection of $\mathbf{X}$ on the orthocomplement of $\mathbf{Z}$), whereas the F-ratio in raw X-data permutation would examine different contrasts [33]. The method is named Smith in [7] and O'Gorman-Smith in [25], but, based on priority, Collins-Dekker in this paper.

In ordinary Collins-Dekker, each new data set in the permutation procedure contains the original $\mathbf{Y}$ and $\mathbf{Z}$, but the predictor matrix $\mathbf{X}$ is replaced by $\pi(\hat{\mathbf{E}}^X)$, a random row permutation of the WLS residuals of the regression of $\mathbf{X}$ on to $\mathbf{Z}$, and the F-ratio is obtained from the new data set by WLS. Note that, the identity permutation (i.e. no permutation) yields the same F-ratio (F_0) as the original data set ($\mathbf{Y}, \mathbf{Z}, \mathbf{X}$).

In standardized Collins-Dekker, each new data set in the permutation procedure contains the transformed response and covariate data $\tilde{\mathbf{Y}} = \mathbf{W}^{\frac{1}{2}}\mathbf{Y}$ and $\tilde{\mathbf{Z}} = \mathbf{W}^{\frac{1}{2}}\mathbf{Z}$, but the predictor matrix $\tilde{\mathbf{X}} = \mathbf{W}^{\frac{1}{2}}\mathbf{X}$ is replaced by $\widetilde{\hat{\mathbf{E}}}^X$, a random row permutation of the residuals of the regression of $\tilde{\mathbf{X}}$ on to $\tilde{\mathbf{Z}}$. The F-ratio is obtained from the new data set by OLS.

Back-transformation leads to a new data set ($\mathbf{Y}, \mathbf{Z}, \mathbf{X}^{\pi *}$) with

$$\mathbf{X}^{\pi *} = \mathbf{W}^{-1/2}\pi(\widetilde{\hat{\mathbf{E}}}^X) = \mathbf{W}^{-1/2}\pi(\mathbf{W}^{1/2}\hat{\mathbf{E}}^X). \tag{9}$$

from which the F-ratio is calculated by WLS. Equation (9) shows that standardized Collins-Dekker weighs the original residuals $\hat{\mathbf{E}}^X$ by the square-root of ratios of weights of cases. This ratio may have a good reason if the error variance in the $\mathrm{X} \sim \mathrm{Z}$ model is proportional to the weight, but arguments for such variance-weight relation are lacking. It was the error variance in the $\mathrm{Y} \sim \mathrm{Z} + \mathrm{X}$ model that motivated standardized Freedman-Lane.

Randomization in experiments forms a strong basis for Collins-Dekker as it randomly assigns treatments ($\mathbf{X}$) to units (cases). In observational studies, Collins-Dekker can be motivated by assuming exchangeability of the rows of $\mathbf{X}$. In both randomized experiments and observational studies standardized Collins-Dekker appears less attractive than ordinary Collins-Dekker, as the latter exchanges the raw X-residuals of the weighted analysis instead of using an additional, rather unnatural assumption such as 'variance of the X-residual is inversely proportional to the weight', leading to standardized X-residuals. The

term 'residualized predictor permutation' will therefore refer to ordinary Collins-Dekker only.

2.5. Computation

The F-ratio in ordinary Freedman-Lane must be computed from the model $\mathbf{Y}^{\pi} \dot{\sim} \mathbf{Z} + \mathbf{X}$, using WLS but, because $\hat{\mathbf{Y}}_0 = \mathbf{Z}\hat{\mathbf{A}}_0$ is in the span of $\mathbf{Z}$, it can also be computed from the model $\pi(\hat{\mathbf{E}}_0^Y) \dot{\sim} \mathbf{Z} + \mathbf{X}$. The F-ratio of the latter can be computed in two steps, namely by fitting $\pi(\hat{\mathbf{E}}_0^Y) \dot{\sim} \mathbf{Z}$ and $\pi(\hat{\mathbf{E}}_0^Y) \dot{\sim} \hat{\mathbf{E}}^X$ as $\hat{\mathbf{E}}^X$ is $\mathbf{W}$-orthogonal to $\mathbf{Z}$. Both steps are needed, as $\pi(\hat{\mathbf{E}}_0^Y)$ and $\mathbf{Z}$ are not necessarily $\mathbf{W}$-orthogonal. For this reason, regression sum of squares due to $\mathbf{X}$ and the F-ratio are not monotonically related (and thus yield different P-values) when $q > 1$ or the weights are unequal.

The F-ratio in ordinary Collins-Dekker must be computed from the model $\mathbf{Y} \dot{\sim} \mathbf{Z} + \pi(\hat{\mathbf{E}}^X)$, but it can also be computed from the model $\hat{\mathbf{E}}_0^Y \dot{\sim} \mathbf{Z} + \pi(\hat{\mathbf{E}}^X)$, because $\mathbf{Y}$ and $\hat{\mathbf{E}}_0^Y$ differ by $\mathbf{Z}\hat{\mathbf{A}}_0$ which is in the span of $\mathbf{Z}$. The F-ratio of this model can also be computed in two steps, namely by first calculating the residuals of the model $\pi(\hat{\mathbf{E}}^X) \dot{\sim} \mathbf{Z}$, denoted by $\hat{\mathbf{E}}_Z^{\pi X}$, and then fitting the model $\hat{\mathbf{E}}_0^Y \dot{\sim} \hat{\mathbf{E}}_Z^{\pi X}$. Note that these steps differ from the ones hereabove; in particular, the model $\hat{\mathbf{E}}_0^Y \dot{\sim} \mathbf{Z}$ does not need to be computed as $\hat{\mathbf{E}}_0^Y$ and $\mathbf{Z}$ are $\mathbf{W}$-orthogonal. As the contribution of $\mathbf{Z}$ to the fit of $\mathbf{Y}$ is constant during permutation, the regression sum of squares due to $\hat{\mathbf{E}}_Z^{\pi X}$ and the F-ratio are monotonically related in Collins-Dekker.

For the standardized permutation methods analogous procedures apply with all matrices replaced by the ones with superscript '$\sim$' and analysis of each model by OLS with $\mathbf{W}$-orthogonality replaced by $\mathbf{I}_n$-orthogonality.

It may be advantageous for computational speed to carry out the ordinary permutation methods by first transforming the WLS problem to an OLS problem by multiplication by the square-root of the weights and then carrying out the permutation using OLS only. This requires that any OLS residual is reweighted (i.e. back-transformed to a WLS residual), permuted and then reweighted (i.e. transformed to an OLS residual) before it is used further in the OLS analysis. The transformations are a division and a multiplication by the square-root of the weights, respectively, precisely the opposite of those in Eqs. (8) and (9). Illustrations of these computations are given in Appendix S2.

The essential computational difference between Freedman-Lane and Collins-Dekker is the regression of the Y-residuals with respect to $\mathbf{Z}$ in Freedman-Lane and the regression of the X-residuals with respect to $\mathbf{Z}$ in Collins-Dekker (Table 2). If the number of responses (m) is larger than the number of predictor variables (p), Collins-Dekker is thus slightly quicker, whereas Freedman-Lane is quicker in the reverse case ($m < p$).

2.6. Raw data permutation versus permutation of residuals

Raw data permutation, permuting the rows of $\mathbf{Y}$ or of $\mathbf{X}$ instead of permuting residuals as in this paper, has been proposed for multiple regression by Manly [34] and Draper and Stoneman [32], but these methods are not recommended for general usage [7,13]. When testing the overall null hypothesis only (i.e. $\mathbf{Z} = \mathbf{1}_n$), Freedman-Lane and Collins-Dekker

can be equivalent with the Manly and Draper-Stoneman methods. This subsection studies when these methods are equivalent.

With raw Y-data permutation and $\mathbf{Z} = \mathbf{1}_n$, $\hat{\mathbf{Y}}_0$ is invariant under permutation, as all rows of $\hat{\mathbf{Y}}_0$ are equal (all equal to the weighted means of the response variables). Therefore, $\mathbf{Y}^{\pi}$ in Equation (6) is equal to $\pi(\mathbf{Y})$, implying that raw Y-data permutation is equivalent with permuting Y-residuals, but only if $\mathbf{Z} = \mathbf{1}_n$. Ordinary Freedman-Lane is thus equivalent with the Manly method of raw Y-data permutation in testing the overall null hypothesis. However, this equivalence does not extend to standardized Freedman-Lane as the rows of $\tilde{\hat{\mathbf{Y}}}_0 = \mathbf{W}^{1/2}\hat{\mathbf{Y}}_0$ are equal for equal weights only.

If $\mathbf{Z} = \mathbf{1}_n$, raw X-data permutation uses $\mathbf{Y} \dot{\sim} (1_n : \pi(\mathbf{X}))$ which gives the same F-ratio as $\mathbf{Y} \dot{\sim} (1_n : \pi(\hat{\mathbf{E}}^X))$ as $\pi(\mathbf{X}) \dot{\sim} 1_n$ yields residuals $\pi(\hat{\mathbf{E}}^X))$. Ordinary Collins-Dekker is therefore equivalent with the Draper-Stoneman method of raw X-data permutation in testing the overall null hypothesis.

With $\mathbf{Z} = \mathbf{w} = 1_n$, $\pi(\mathbf{Y}) \sim \mathbf{X}$ gives the same F-ratio as $\mathbf{Y} \sim \pi^{-1}(\mathbf{X})$, so that, with the previous equivalences, Manly, Draper-Stoneman, Freedman-Lane and Collins-Dekker are all equivalent when testing the overall null hypothesis with equal weights. In this simple equi-weight case, these methods thus result in the same P-values in Monte Carlo testing when each permutation π in Y-permutation using either Manly or Freedman-Lane is replaced by its inverse in X-permutation using either Draper-Stoneman or Collins-Dekker.

Manly and Draper-Stoneman are not equivalent in WLS when weights vary. However, in some methods, for example canonical correspondence analysis, the weights are linked to $\mathbf{Y}$ (the weights are the row sums of $\mathbf{Y}$), so that they 'travel with' the rows of $\mathbf{Y}$. In such situation, raw Y-data permutation is equivalent with raw X-data permutation, as the data set $(\pi(\mathbf{Y}),\ \mathbf{X})$ with weights $\pi(\mathbf{w})$ is a reordering of the data set $(\mathbf{Y}, \pi^{-1}(\mathbf{X}))$ with weights $\mathbf{w}$, so that their analyses are equivalent.

3. A univariate data example

The example, taken from ter Braak [35], is on trait-environment association in ecology and serves to illustrate that the four permutation methods can yield wildly different P-values. It considers 10,695 individuals of 75 different species and the values of an environmental variable at which individuals occurs. As trait values are available for species only, these values are summarized by their mean per species. The mean environment is then related to the trait of interest. In this description, the data collection process has been simplified and idealized for the sake of argument. As the response values are means, there is reason to use the sample size, on which each mean is based, as weight (Figure S1). Also, this weighted regression is related to what is known as the fourth-corner correlation [36]. This illustrative example is extreme, as there are outliers in the weights resulting in a coefficient of variation of 3.4, so that the weighted regression is dominated by a few data points (Figure S1). With these weights, the standardized versions of the permutation methods resulted both in a P-value of 0.001 and the ordinary Freedman-Lane and ordinary Collins-Dekker resulted in the P-values 0.122and 0.060, respectively (9999 permutations). For wildly different weights, the four permutation methods can thus yield wildly different P-values. For comparison, the parametric P-values of unweighted and weighted regression are 0.035 and < 0.0001, respectively. The permutational P-value using OLS-regression is 0.035 also; the methods of

permutation are all equivalent in the simple equi-weight case (Section 2.6). Disregarding the weights altogether looks most safe, but removes the link of the regression to the fourth-corner correlation.

Note that all four permutation methods are based on weighted regression; they differ in how the residuals are treated (permuted unchanged in the ordinary methods, but standardized and reweighted during permutation in the standardized methods).

4. Simulation

In this section, the performance of the four permutation methods in weighted regression and RDA is investigated by simulation. Weights $\{w_i\}(i = 1, \ldots, n)$ were either all set equal or drawn independently from a negative binomial distribution with mean 10, increased by 1 to avoid zero weights, and used in each analysis unless explicitly noted otherwise.

4.1. Testing effects by weighted simple and multiple regression

In this section, data ($n = 30$) was simulated from the univariate linear regression model

$$y_i = a_0 + a_1 z_i + b x_i + e_i \text{ with weights } w_i, \tag{10}$$

with $a_0 = 1$, (z_i, x_i) bivariate Student-t distributed with 5 degrees of freedom and a correlation 0.7 and e_i an independent normal error with variance that is a convex combination of variance inversely related to the simulated weights ($var(e_i) \propto 1/w_i$) and constant variance, representing measurement error and error due to a missing covariate, respectively. The Student-t distribution was chosen so as to avoid the danger that the simulation set-up was advantageous for Collins-Dekker. Note that in a multivariate Student-t distribution, the condition variance of a variable depends on the values of the other variables [25,37], whereas this variance is constant in the multivariate normal distribution. The null hypothesis $b = 0$ was tested with either $a_1 = 0$ (simple regression) or $a_1 = 1$ (a single non-trivial covariate) using both the usual parametric F-test (in both weighted and unweighted regression) and the four variants of Monte Carlo permutation from Section 2 with $K = 199$. Each permutation in Collins-Dekker was the inverse of each permutation in Freedman-Lane, so that the resulting P-values are identical in simple regression with equal weights ($w_i = 1$), despite the small number of K.

The type I error rate in 1000 simulated data sets ($b = 0$) was determined for eleven scenarios for the variance, running from variance purely inverse of weight to constant variance and six levels of weight variability, quantified by the overdispersion of the negative binomial, including the equal-weights scenario.

Figures 1 ($a_1 = 1$) and S2 ($a_1 = 0$) show that all methods control the type I error when the weights are equal or when the variance is inversely proportional to the weight, but also that the standardized parametric and both standardized permutation methods become more and more liberal when the weight variability and the constant variance part in the error variance increase. Of the ordinary methods, Freedman-Lane can be very conservative, which then results in lower power than ordinary Collins-Dekker (Figure 2). The coefficient of variation of the weights of the data example lies in between those of weight variability 10 and 20 in Figures 1 and 2.

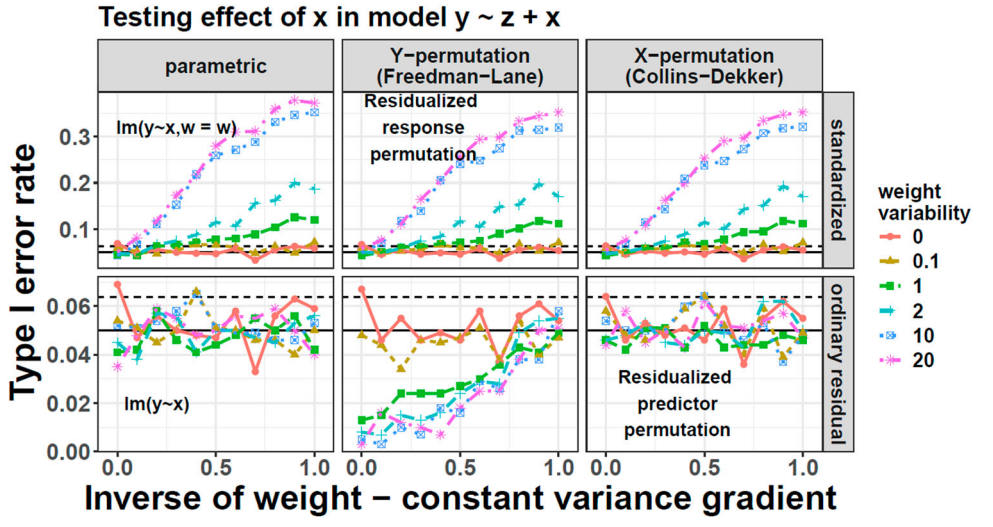

Figure 1. Type I error rates of testing the effect of x in the linear model $y \sim 1 + z + x$ with error variance that runs from variance inversely to weight (scale value 0) to constant variance (scale value 1). All panels use WLS, except for the lower-left panel which uses OLS. Weights were drawn from a negative binomial with mean 10, increased by 1 to avoid zero weights; weight variability is its overdispersion, except that the value 0 indicates 'no variability', i.e. equal weights. The six levels of increasing weight variability yield coefficients of variation of the weights of 0, 0.41, 0.95, 1.3, 2.9 and 4.1. The horizontal solid line is at the nominal significance threshold; rates (from 1000 simulations) above the dashed line (at 0.064) are significantly greater than 0.05.

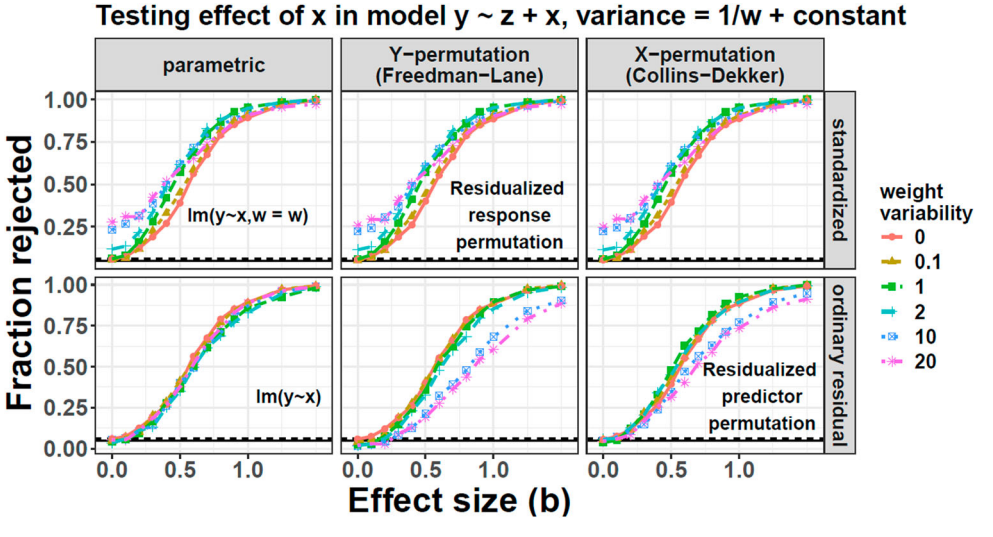

Figure 2. Rejection rates of testing the effect of x in the linear model $y \sim 1 + z + x$ against the effect size, when the error variance is halfway through the scale of Figure 1, i.e. a constant plus variance inversely proportional to the weights used in WLS, except for the lower-left panel which uses OLS. Weights were drawn as in Figure 1. The horizontal solid line is at the nominal significance threshold; rates (from 1000 simulations) above the dashed line (at 0.064) are significantly greater than 0.05.

As predicted by theory, Freedman-Lane and Collins-Dekker gave exactly the same rejection rates in simple regression with equal weights (Figure S2), but also with the covariate (Figures 1 and 2), the rates differed little (at most 0.02), even when compared with those of the parametric F-test, so that the red lines in Figures 1, 2 and S2 are (almost) identical in the six panels of each of these figures.

In Figure 2, the power of standardized Freedman-Lane is at most 0.26 higher than that of ordinary Collins-Dekker, but this higher power comes at a cost of an inflated type I error of 0.25.

Permutation tests using OLS instead of WLS yielded results that are almost indistinguable from those of the parametric unweighted analyses in the lower-left boxes of Figures 1 and 2. When the weight-variance relation is uncertain, it is therefore safe to ignore the weights, if possible.

The same data were also analysed after binarization of the covariate and predictor data (1 if positive, 0 elsewere) and after exponentiation ($(\mathbf{X},\mathbf{Z}) \leftarrow \exp(\mathbf{X},\mathbf{Z})$). Both analyses lead to qualitatively the same results (e.g. Figure S3), except that the power was decreased with skewed predictor data.

4.2. Testing effects by weighted RDA

Multivariate data with $n = 30$ cases and $m = 50$ response variables was simulated from Equation (1) using a reduced rank regression model of rank three, with $\mathbf{B}$ of rank 2, two predictor dimensions each defined by four predictors, and also with $\mathbf{A}$ of rank 2, one for the intercept and one for a non-trivial covariate dimension defined by two covariates. In addition, four variables without effect on the responses were simulated, so that $p = 4+4+4 = 12$ and $q = 1+2 = 3$. The correlation of the four predictors of the first dimension with the covariate dimension was 0.7, as were the correlation of the four predictors of second dimension with the first dimension, the correlation of the four predictors without effect with the second dimension and the correlation between successive variables within each of the four sets constituting the covariate and predictor variables. The error was independent normal with zero mean and variance that was either inversely related to the simulated weights ($var(e_{ik}) \propto 1/w_i$, $k = 1, \ldots, m$) or constant (1). Two more scenarios were created by addition of rank-1 error to the former two scenarios. Rank-1 error was of the form $d_k^* e_i^*$ (Appendix S1) mimicking the effects (d_k^*) of a latent (unobserved) predictor (e_i^*) on the m responses. Effect size b was introduced in the model as a multiple of a matrix $\tilde{\mathbf{B}}$ of non-zero regression coefficients, so that $\mathbf{B} = b\tilde{\mathbf{B}}$ and the importance of the predictor dimensions increases with effect size. We used two scenarios for the distribution of $\mathbf{X}$ and $\mathbf{Z}$. In the first, $\mathbf{X}$ and $\mathbf{Z}$ were multivariate normal, whereas in the second scenario their distribution was jointly Student-t with 5 degrees of freedom and a scale matrix that is equal to $(5–2)/5 = 0.6$ times the covariance as in the normal case, so as to give approximately the same expected covariance in both scenarios. Details and alternative formulations of the model are given in Appendix S1. In each simulated data set, the effect of $\mathbf{X}$ was tested with the Monte Carlo significance tests of Section 2, i.e. applying the permutation methods in weighted and unweighted partial RDA (partial as there are covariates in the analysis). Weights and further analysis of the simulation results were as for simple and multiple regression.

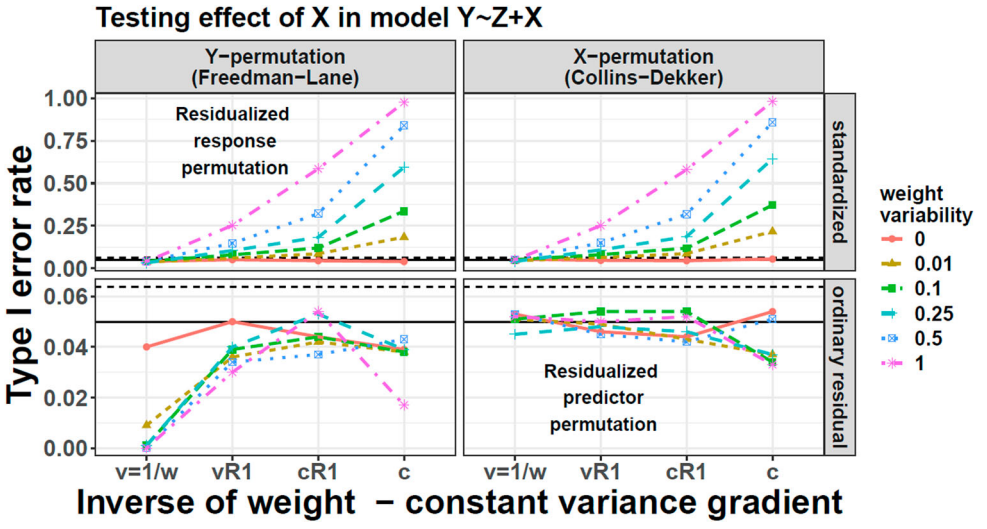

Figure 3. Type I error rates of testing the effect of **X** in a reduced rank model $\mathbf{Y} \sim \mathbf{Z} + \mathbf{X}$ with error variance that runs from variance purely inversely to weight (v = 1/w) to constant variance (c) with, in between, the two scenarios vR1 and cR1 that include rank-1 error in addition to the error as in the utmost left (v = /1w) and utmost right (c) scenarios, respectively. All panels use weighted partial redundancy analysis (RDA) with two non-trivial covariates (**Z**) and twelve predictors (**X**) with weights drawn as in Figure 1 ($n = 30, m = 50, q = 1 + 2, p = 12$). The six levels of increasing weight variability yield coefficients of variation of the weights of 0, 0.30, 0.41, 0.54, 0.70, and 0.95. The horizontal solid line is at the nominal significance threshold; rates (from 1000 simulations) above the dashed line (at 0.064) are significantly greater than 0.05.

As the results of two scenarios for the distribution of **X** and **Z** are qualitatively similar, we report only one of them, namely the Student-*t* scenario. Even with weight variability that is modest in terms of Figure 1, the standardized permutation methods gave grossly inflated type-I error rates in this multivariate context, except in the scenario with variance inversely proportional to the weights (Figure 3). Their power was rather similar (Figure 4). Ordinary Collins-Dekker outperformed ordinary Freedman-Lane, both in terms of type I error rate (Figure 3) and power (Figure 4), except that they had similar power when the variance was constant (Figure 4). Applying OLS instead of WLS yielded type I error rates close to 0.05; its power was somewhat higher than ordinary Collins-Dekker, except in the no rank 1 noise case when the variance was inversely related to the weights (Figure 4).

The same data were also analysed after binarization of the covariate and predictor data (1 if positive, 0 elsewere) and after exponentiation ($(\mathbf{X},\mathbf{Z}) \leftarrow \exp(\mathbf{X}, \mathbf{Z})$). Both analyses lead to qualitatively the same results (e.g. Figure S4), except that the power was decreased.

Neither ordinary Freedman-Lane nor standardized Collins-Dekker is further considered: the former because of its inferior performance, and the latter, because its rationale is less strong and its performance not better than standardized Freedman-Lane.

4.3. Testing of the axes of weighted RDA

This section examines the two remaining permutation tests, standardized Freedman-Lane and ordinary Collins-Dekker for testing the significance of the dimensions of the RDA. It

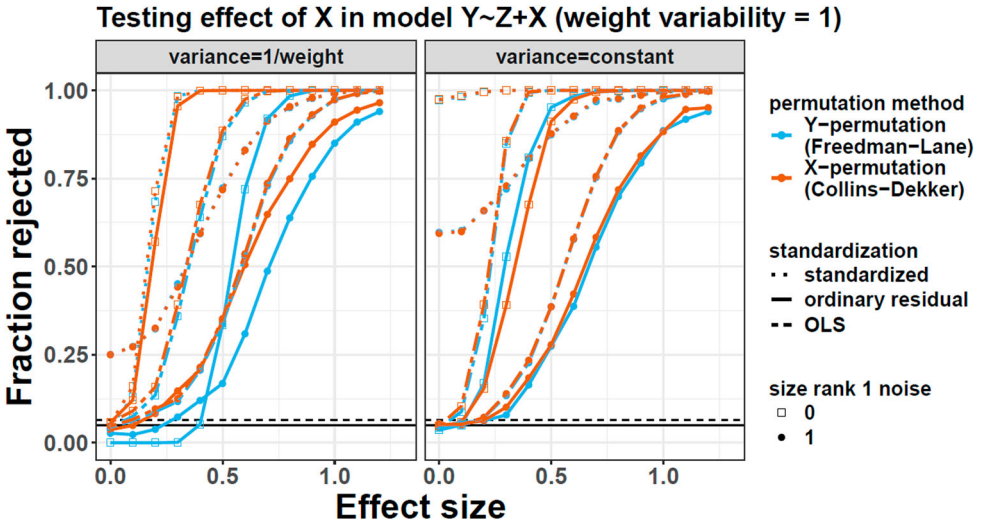

Figure 4. Rejection rates of testing the effect of **X** in a reduced rank model $\mathbf{Y} \sim \mathbf{Z} + \mathbf{X}$ by weighted partial redundancy analysis (RDA) with a weight variability of 1 against the effect size, with, as in Figure 3, two non-trivial covariates (**Z**), twelve predictors (**X**) and weights drawn as in Figure 1 ($n = 30, m = 50$, $q = 1 + 2, p = 12$). OLS (dashed lines) disregarded the weights and was obtained by unweighted RDA. The horizontal solid line is at the nominal significance threshold; rates (from 1000 simulations) above the dashed line (at 0.064) are significantly greater than 0.05.

uses the same data generation model and scenarios as the previous section, with an effect size of 2 for the first dimension, so that it is nearly always considered significant and a varying effect size of the second dimension. Conditional on the covariates **Z**, the expected responses lie in a two-dimensional space; the third eigen value of $E(\mathbf{Y}) \sim \mathbf{X}|\mathbf{Z}$ is therefore zero. Testing of the second and third axis proceeds by either the test statistic F of Equation (3) or that based on the first eigenvalue (F_e), given all covariates and previously tested dimensions.

With both test statistics, residualized response permutation (standardized Freedman-Lane) had (up to hugely) inflated type I error rates in testing the second and third dimensions of the model (Figure 5) when weights were unequal, except in the scenario without rank 1 noise but with variance inversely proportional to weight (Figure 5).

With both test statistics, residualized predictor permutation (ordinary Collins-Dekker) controlled the type I error rate and both test statistics resulted in similar power (Figure 5).

4.4. Sensitivity to weight-predictor correlation

This section examines the sensitivity of the methods to the correlation between the weight and the predictors when there is no effect of the predictors on the mean response. The correlation was simulated by linearly combining the logarithm of the weight with the first predictor so as to set a particular correlation. After backtransformation, the weight was used in the weigthed analyses and for the variance of the responses in the scenarios that require this.

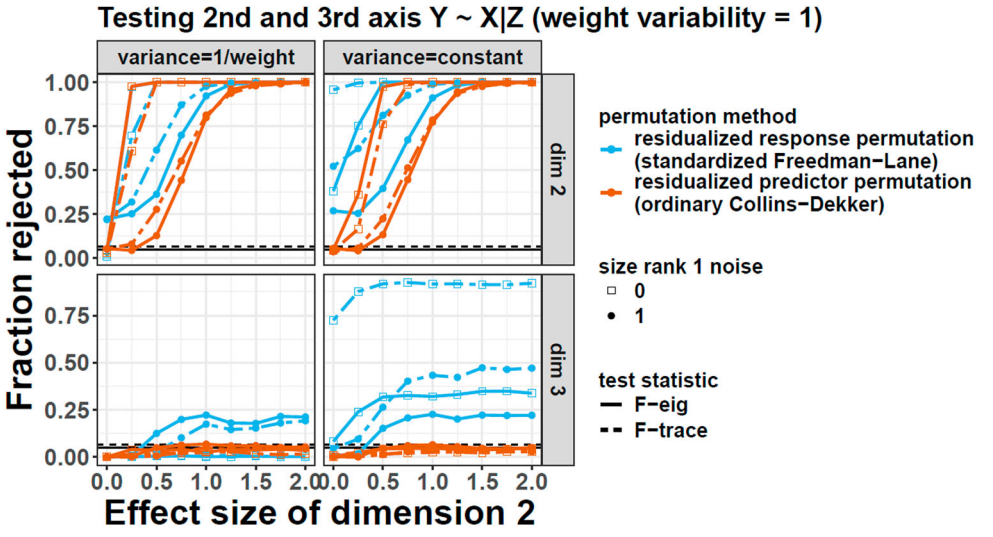

Figure 5. Rejection rates of testing the second and third axes against the effect size of the second axis in the reduced rank model $\mathbf{Y} \sim \mathbf{X}|\mathbf{Z}$ by weighted partial redundancy analysis (RDA) with a weight variability of 1 using two alternative test statistics (F_{eig} and $F_{trace} = F$ of Equation (3)) with, as in 3 and 4, two non-trivial covariates ($\mathbf{Z}$) and twelve predictors ($\mathbf{X}$) with weights drawn as in Figure 1 ($n = 30$, $m = 50$, $q = 1 + 2$, $p = 12$). The rejection rate for the first axis was close to 1 everywhere. The horizontal solid line is at the nominal significance threshold; rates (from 1000 simulations) above the dashed line (at 0.064) are significantly greater than 0.05.

Using WLS, both ordinary permutation methods showed hardly any type I error inflation for $n = 30$ (Figure 6). But after increasing the sample size to $n = 100$, ordinary Collins-Dekker showed huge type I error inflation (up to 0.50), whereas ordinary Freedman-Lane did not. In the constant variance scenario OLS performed, of course, perfect, but in the other scenario, with the variance inversely proportional to the weight, OLS showed type I error inflation for both sample sizes (Figure 6). The standardized permutation methods are not shown in Figure 6. With weight variability and rank 1 noise both 1, as in Figure 6, their type I error rates were between 0.23 and 1.0.

5. Discussion and conclusion

This paper defines and evaluates by simulation four permutation methods for statistical testing of the effect of a set of predictors on the distribution of one or more response variables in weighted linear regression and redundancy analysis. The simulation results (Figures 1–5, Figures S2–S4) show that the standardized permutation methods (methods that weigh or standardize the residuals and then reweigh after permutation) can be extremely liberal if the weights are not exactly inversely proportional to the error variance. They behave like parametric tests in weighted multiple regression (Figure 1). The Freedman-Lane and Collins-Dekker versions of the standardized permutation methods are rather similar. In contrast, the ordinary permutation methods (methods that permute the WLS-residuals 'as is') differ significantly in that Freedman-Lane can be very conservative and is less powerful than Collins-Dekker. In terms of type I error rate, ordinary

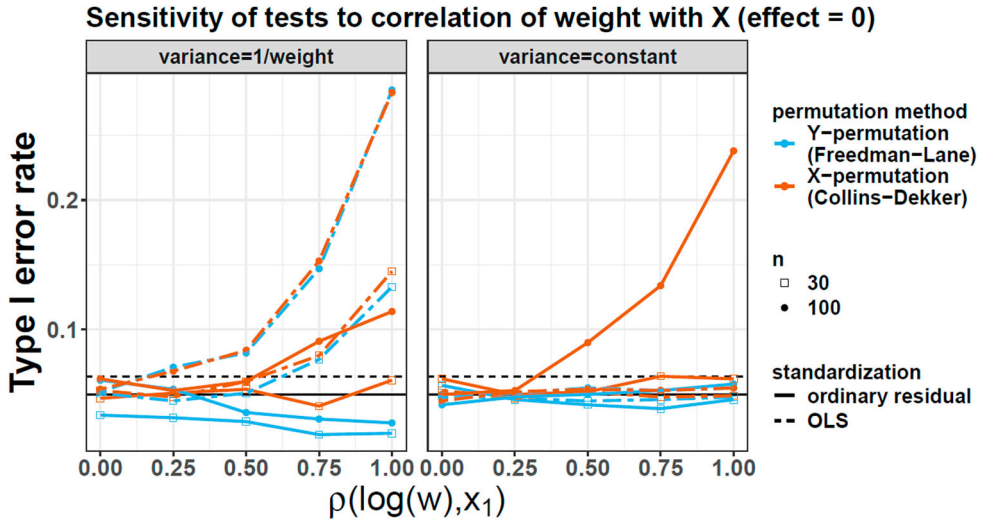

Figure 6. Type I error rates of testing the effect of **X** in a reduced rank model $\mathbf{Y} \sim \mathbf{Z} + \mathbf{X}$ by weighted partial redundancy analysis (RDA), with weight variability and rank 1 noise both 1, against the correlation of the weights with the first predictor, with two non-trivial covariates (**Z**), twelve predictors (**X**) and weights drawn as in Figure 1 ($n = 30$ and 100, $m = 50$, $q = 1 + 2$, $p = 12$). OLS (dashed lines) disregarded the weights and was obtained by unweighted RDA. The horizontal solid line is at the nominal significance threshold; rates (from 1000 simulations) above the dashed line (at 0.064) are significantly greater than 0.05.

Collins-Dekker is unbiased and behaves in multiple regression like a parametric test that ignores the weights (Figure 1).

Whereas the standardized permutation methods are slightly more powerful when the error variance is inversely proportional to the weights, ordinary Collins-Dekker is by far the most robust against variability in the weights and relationships of the error variance to the weights. Similar remarks apply when testing the dimensionality of a reduced rank regression. Alternatively, if this is an option, one may altogehter ignore the weights and use OLS instead of WLS; this avoids Type I error inflation when weights are independent of the predictors and was in the majority of scenarios more powerful than ordinary Collins-Dekker. In the light of these results, the permutational *P*-value of 0.035 of OLS in the data example of Section 3 would be the most trustworthy, if one could ignore the weights. However, the weights are an essential ingredient of the fourth-corner correlation. The *P*-value of 0.06 of ordinary Collins-Dekker is thus the most trustworthy for testing the fourth-corner correlation.

We considered including the Huh-Jhun method [38] in this paper, but did not because we failed to get a robust ordinary version of the method. The key of the method is to transform the residuals so that these are uncorrelated with equal first and second moments. Many such transformations exist, which introduces some non-trivial arbitrariness [39]. We add that this arbitrainess could be removed by making it a rotation test [40] instead of a permutation only test. The standardized version of Huh-Jhun yielded in our simulation type I error inflation where the other standardized methods yielded inflation, and did not give higher power elsewhere, in line with results reported in [41].

If the null hypothesis is restricted to effect on the mean response (as in Behrens-Fisher type problems), then, in view of Figure 6, ordinary Freedman-Lane might be considered instead of ordinary Collins-Dekker, particularly if Type I error control is judged more important than power. However, for testing such restricted null hypothesis it is probably wiser to switch from the *F*-test statistic used in this paper to a robust Wald test-statistic [24,25].

Whereas permutation testing avoids the assumption of normality, it still needs independence. Such independence is violated, among others, in hierarchical study designs and in data with spatial and temporal autocorrelation. In particular cases, approximately valid permutation tests can be constructed by restricted permutation [42]. The simplest example is the restriction to within-block permutation in a randomized block experiment or within-subject in a hierarchical study. Other cases put additional requirement on the study design, e.g. balanced data in an hierarchical design so that all data of a subject can be shuffled with the data of another subject, or equal spacing in space or time [43–45].

This paper uses simulation as the main evaluation tool. The limitation of conclusions drawn from simulation is analogous to the following in mathematics: a single counter example is sufficient to show that a theorem is false, but a theorem cannot be proven by example(s). By simulation we can show that a method does not control the type I error rate by simulation of a particular scenario, where another method does control it in this scenario. However, the other method might fail elsewhere. Thus, it good to keep in mind that our results depend on the scenarios being simulated, however well-chosen.

In conclusion, for general use in weighted linear analysis, ordinary Collins-Dekker is recommended over the other three permutation methods in Tables 1 and 2. Ordinary Collins-Dekker and standardized Freedman-Lane are both are implemented in version 5.15 of the Canoco software [45] under the names of residualized predictor permutation and residualized response permutation, respectively. In these terms, 'residualized' must be understood as being with respect to the covariates only. R-code for all versions is provided in Appendix S3.

Acknowledgments

I thank Dennis te Beest and Petr Šmilauer as well as the associated editor and reviewers for their comments.

Disclosure statement

CJFtB is the senior author of the Canoco software for visualization of multivariate data. He has academic, but no financial, interest in Canoco.

ORCID

Cajo J. F. ter Braak http://orcid.org/0000-0002-0414-8745

References

[1] van den Wollenberg AL. Redundancy analysis an alternative for canonical correlation analysis. Psychometrika. 1977;42:207–219.
[2] Legendre L, Legendre P. Numerical ecology. Amsterdam: Elsevier; 2012.
[3] Davies PT, Tso MK-S. Procedures for reduced-rank regression. Appl Stat. 1982;31:244–255.

[4] ter Braak CJF, Looman CWN. Biplots in reduced-rank regression. Biom J. 1994;36:983–1003.
[5] Rao CR. The use and interpretation of principal component analysis in applied research. Sankhya A. 1964;26:329–358.
[6] Thioulouse J, Dray S, Dufour A-B, et al. Multivariate analysis of ecological data with ade4. New York (NY): Springer New York; 2018.
[7] Winkler AM, Ridgway GR, Webster MA, et al. Permutation inference for the general linear model. NeuroImage. 2014;92:381–397.
[8] Fieberg JR, Vitense K, Johnson DH. Resampling-based methods for biologists. PeerJ. 2020;8:e9089.
[9] Freedman DA, Lane D. A nonstochastic interpretation of reported significance levels. J Bus Econ Stat. 1983;1:292–298.
[10] Collins MF. A permutation test for planar regression. Aust J Stat. 1987;29:303–308.
[11] Dekker D, Krackhardt D, Snijders TAB. Multicollinearity Robust QAP for Multiple-Regressions 2003 Pittsburgh North American Association for Computational Social and Organization Sciences (NAACSOS 2003) http://www.casos.cs.cmu.edu/publications/papers/dekker_2003_multicollinearity.pdf.
[12] Dekker D, Krackhardt D, Snijders TAB. Sensitivity of MRQAP tests to collinearity and autocorrelation conditions. Psychometrika. 2007;72(4):563–581.
[13] Anderson MJ, Robinson J. Permutation tests for linear models. Aust N Z J Stat. 2001;43(1):75–88.
[14] ter Braak CJF. Canonical correspondence analysis: a new eigenvector technique for multivariate direct gradient analysis. Ecology. 1986;67:1167–1179.
[15] ter Braak CJF. History of canonical correspondence analysis. In: Blasius J, Greenacre M, editors. Visualization and verbalization of data. London: Chapman and Hall/CRC; 2014. p. 61–75.
[16] ter Braak CJF, Verdonschot PFM. Canonical correspondence analysis and related multivariate methods in aquatic ecology. Aquat Sci. 1995;57:255–289.
[17] Greenacre MJ. Theory and applications of correspondence analysis. London: Academic Press; 1984.
[18] Heiser WJ. Joint ordination of species and sites: the unfolding technique. In: Legendre P, Legendre L, editors. Developments in numerical ecology. Berlin: Springer-Verlag; 1987. p. 189–224.
[19] ter Braak CJF. The analysis of vegetation-environment relationships by canonical correspondence analysis. Vegetatio. 1987;69:69–77.
[20] Aitchison J. Principal component analysis of compositional data. Biometrika. 1983;70:57–65.
[21] Greenacre M. Compositional data analysis in practice. Boca Raton, Florida: CRC Press; 2018; (Keiding N, Morgan BJT, Wikle CK, et al., editors. Chapman & Hall/CRC Interdisciplinary Statistics).
[22] Greenacre M, Lewi P. Distributional equivalence and subcompositional coherence in the analysis of compositional data, contingency tables and ratio-scale measurements. J Classif. 2009;26(1):29–54.
[23] Seber GAF. Linear regression analysis. New York: Wiley; 1977.
[24] DiCiccio CJ, Romano JP. Robust permutation tests for correlation and regression coefficients. J Am Stat Assoc. 2017;112(519):1211–1220.
[25] Helwig NE. Robust nonparametric tests of general linear model coefficients: a comparison of permutation methods and test statistics. NeuroImage. 2019;201:116030.
[26] Legendre P, Oksanen J, ter Braak CJF. Testing the significance of canonical axes in redundancy analysis. Methods Ecol Evol. 2011;2(3):269–277.
[27] Winkler AM, Renaud O, Smith SM, et al. Permutation inference for canonical correlation analysis. NeuroImage. 2020;220:117065.
[28] Phipson B, Smyth GK. Permutation P-values should never be zero: calculating exact P-values when permutations are randomly drawn. Stat Appl Genet Mol Biol. 2010;9:Article39.
[29] Hemerik J, Goeman J. Exact testing with random permutations. TEST. 2018;27(4):811–825.
[30] Kempthorne O. The design and analysis of experiments. New York: Wiley; 1952.

[31] Oja H. On permutation tests in multiple regression and analysis of covariance analysis problems. Aust J Stat. 1987;29:91–100.
[32] Draper NR, Stoneman DM. Testing for the inclusion of variables in linear regression by a randomisation technique. Technometrics. 1966;8(4):695–699.
[33] ter Braak CJF. Permutation versus bootstrap significance tests in multiple regression and ANOVA. In: Jöckel K-H, Rothe G, Sendler W, editors. Bootstrapping and related techniques. Berlin: Springer Verlag; 1992. p. 79–85.
[34] Manly BFJ. Randomization and regression methods for testing for associations with geographical, environmental and biological distances between populations. Res Popul Ecol (Kyoto). 1986;28(2):201–218.
[35] ter Braak CJF. New robust weighted averaging- and model-based methods for assessing trait–environment relationships. Methods Ecol Evol. 2019;10(11):1962–1971.
[36] Peres-Neto PR, Dray S, ter Braak CJF. Linking trait variation to the environment: critical issues with community-weighted mean correlation resolved by the fourth-corner approach. Ecography. 2017;40:806–816.
[37] Ding P. On the conditional distribution of the multivariate *t* distribution. Am Stat. 2016;70(3):293–295.
[38] Huh M-H, Jhun M. Random permutation testing in multiple linear regression. Commun Stat Theor Methods. 2001;30(10):2023–2032.
[39] Kherad-Pajouh S, Renaud O. An exact permutation method for testing any effect in balanced and unbalanced fixed effect ANOVA. Comput Stat Data Anal. 2010;54(7):1881–1893.
[40] Langsrud Ø. Rotation tests. Stat Comput. 2005;15(1):53–60.
[41] Manly BFJ, Alberto JAN. Randomization, bootstrap and Monte Carlo methods in biology. 4th ed. Boca Raton, Florida: Chapman & Hall/CRC; 2021.
[42] Anderson MJ, ter Braak CJF. Permutation tests for multi-factorial analysis of variance. J Stat Comput Simul. 2003;73(2):85–113.
[43] Simpson GL. permute: functions for generating restricted permutations of data. R package version 0.9–5; 2019. https://CRANR-projectorg/package = permute.
[44] ter Braak CJF. Update notes: CANOCO version 3.1. Wageningen: Agricultural Mathematics Group; 1990. http://edepot.wur.nl/250652.
[45] ter Braak CJF, Šmilauer P. Canoco reference manual and user's guide: software for ordination (version 5.10). Ithaca: Microcomputer Power; 2018.