Petra Buzkova*

Interaction Testing: Residuals-Based Permutations and Parametric Bootstrap in Continuous, Count, and Binary Data

DOI 10.1515/em-2015-0010

Abstract: To obtain statistical inference about interaction hypotheses without making strong distributional assumptions, permutation tests based on permuting the outcomes are often being used. It was shown that in continuous and binary data these tests might not be even approximately valid and parametric bootstrap was suggested as a viable alternative, outperforming such permutation tests. We describe an alternative permutation test, permuting the null hypothesis residuals rather than the outcome. Using simulations, we compare accuracy across the permutation tests and parametric bootstrap, studying continuous, binary, and additionally count data. Finally, we address power.

Keywords: interaction testing, permutation methods, parametric bootstrap

1 Introduction

Testing in a regression model framework requires computing the distribution of the test statistic under sampling from the null-hypothesis model. When standard distributional assumptions are doubtful or analytic approaches are intractable, resampling methods might be a viable option for moderate and large sample size data.

Permutations tests (Ernst 2004, Higgins 2004) are non-parametric randomization tests that date back to Fisher (1935). They received more interest with accessible computer power (Brown and Maritz 1982, Freedman and Lane 1983, Manly 1997). A permutation test compares the observed test statistic to a distribution generated by a group of permutations that would not affect the distribution of the test statistic if the null hypothesis were true (Cox and Hinkley 1979). Thus, permutation tests are only applicable when the null hypothesis being tested specifies a suitable group of permutations under which the distribution of the statistic would be unaffected.

Exact permutation tests for interactions are generally not available (Anderson 2001). Permutation of the outcome Y within levels of the two covariates generates data with the interaction effect of the covariates unchanged rather than removed. It was shown with simulations in continuous and binary data that it is typically not possible to use exact permutation tests for combinations of main effects and the interaction effect to test for interaction (Bůžková et al. 2011). Such permutation tests, permuting the outcome Y within the entire data, or permuting the outcome Y within levels of one of the covariates, demonstrated substantial errors. These permutation methods did perform acceptably only in some scenarios when approximately pivotal statistics were used.

The most often used permutations are approaches permuting the outcome. More appropriate for testing an interaction might instead be approaches permuting the null hypothesis residuals. While permuting the outcome is always feasible, permuting the residuals is only possible in models that are collapsible, such as linear regression models for continuous data or log-linear regression models for count data.

A good approximation to the distribution of the test statistic under sampling from the true null-hypothesis model is the distribution of the test statistic under sampling from the fitted null-hypothesis model. This is the idea behind parametric bootstrap (Davison and Hinkley 1997). In moderate and large

***Corresponding author: Petra Buzkova,** Department of Biostatistics, University of Washington, Seattle, WA, USA, E-mail: buzkova@u.washington.edu

sample size parametric bootstrap is a practical alternative for interaction testing in linear and logistic regression, valid even with the non-pivotal statistics (Bůžková et al. 2011).

In this paper we study a test for interactions based on permutations of null model residuals. We describe this approach in linear regression with continuous data and log-linear regression with count data. We demonstrate its superior accuracy to approaches that permute the outcome. Further, we extend the application of parametric bootstrap to count data. We compare the accuracy of the test for interactions based on permuting the null model residuals and parametric bootstrap. Finally, we address power in all approximately accurate approaches, including power in continuous, count, and binary data.

2 Methods

For continuous outcome Y the typical null hypothesis with two covariates E and G is

$$\mathbb{E}[Y] = \alpha + \beta_G G + \beta_E E. \tag{1}$$

For count and binary data, respectively, the null hypothesis may be

$$\log \mathbb{E}[Y] = \alpha + \beta_G G + \beta_E E \tag{2}$$

$$\operatorname{logit} P[Y = 1] = \alpha + \beta_G G + \beta_E E. \tag{3}$$

An alternative hypothesis for continuous, count, and binary data, respectively, may be

$$\mathbb{E}[Y] = \alpha + \beta_G G + \beta_E E + \gamma E \times G \tag{4}$$

$$\log \mathbb{E}[Y] = \alpha + \beta_G G + \beta_E E + \gamma E \times G \tag{5}$$

$$\operatorname{logit} P[Y = 1] = \alpha + \beta_G G + \beta_E E + \gamma E \times G, \tag{6}$$

including the interaction term $E \times G$, Cox (1984).

Rather than comparing the observed value of a test statistic to its distribution under repeated sampling, permutation test compares the observed value to a distribution generated by a group of permutations that would not affect the distribution if the null hypothesis were true. For a valid permutation test of the hypothesis of no interaction, we would require a group of permutations that preserve the main effects β_G and β_E in equation (1), (2), and (3), but also ensure that the interaction coefficient γ is zero. That is impossible to construct (Anderson 2001), even if both main effects were categorical. If we restricted permutations to occur within combinations of the categories of both predictors to control for main effects, there would be generated no other interactions than those from the original data.

2.1 Resampling approaches

We use the notation of Anderson and Robinson (2001) to denote the permutation π of $I_n = \{1, 2, \ldots, n\}$ as a 1 to 1 mapping of I_n onto itself with equal weight $1/n!$ for each permutation. When we use π as a superscript, the superscripted variable is permuted.

The two most widely used types of permutations (Bůžková et al., 2011) are based on permuting the outcome Y. Permutation A, proposed for the test of partial correlation (Manly 1997), and permutation B are:

- A: Keep covariate pairs (G, E) and permute Y to obtain $Y^{\pi(A)}$;
- B: Keep covariate pairs (G, E) and permute Y within levels of E to obtain $Y^{\pi(B)}$.

We now consider permutation C based on permuting the residuals from null-model regression. Without loss of generality, we center covariates at their means. Centering covariates does not modify the interaction coefficient γ in the alternative models,

$$\begin{aligned}
&\alpha + \beta_G(G-\bar{G}) + \beta_E(E-\bar{E}) + \gamma(E-\bar{E})\times(G-\bar{G}) \\
&= \alpha + \beta_G G - \beta_G\bar{G} + \beta_E E - \beta_E\bar{E} + \gamma E\times G - \gamma E\times\bar{G} - \gamma\bar{E}\times G + \gamma\bar{E}\times\bar{G} \\
&= \left(\alpha - \beta_G\bar{G} - \beta_E\bar{E} + \gamma\bar{E}\times\bar{G}\right) + (\beta_G - \gamma\bar{E})G + (\beta_E - \gamma\bar{G})E + \gamma E\times G \\
&= \tilde{\alpha} + \tilde{\beta}_G G + \tilde{\beta}_E E + \gamma E\times G.
\end{aligned}$$

Permutation C is a multi stage method. For continuous data it can be summarized in the following steps:

1. Obtain residuals $R = Y - \hat{Y}$ from the original data by fitting a null-hypothesis model (1).
2. Permute residuals R to obtain $R^{\pi(C)}$.
3. Compute the test statistic based on fitting the model $\mathbb{E}[R^{\pi(C)}] = \alpha^\star + \beta_G^\star G + \beta_E^\star E + \gamma E\times G$.
4. Repeat Steps 2 and 3 many times, to obtain an empirical distribution of the test statistic.
5. Compute the test statistic from the original data, based on fitting the alternative-hypothesis model (4).
6. Compute the p-value, by comparing the test statistic in Step 5 to the distribution in Step 4.

The log-link regression model (5) used for count data is also collapsible. That is $\mathbb{E}[Y] = \exp\{\alpha + \beta_G G + \beta_E E\} \times \exp\{\gamma E\times G\}$. We obtain residuals $R = Y/\hat{Y}$, where $\hat{Y}$ is the fitted mean $\exp\{\hat{\alpha} + \hat{\beta}_G G + \hat{\beta}_E E\}$ from the null model (2). Step 3 proceeds with the model $\log\mathbb{E}[R^{\pi(C)}] = \alpha^\star + \beta_G^\star G + \beta_E^\star E + \gamma E\times G$.

Permutation C is exactly valid if the covariates G and E are independent of each other. If not, the interaction and the main effects are dependent and this permutation is not exact, similarly to permutations A and B.

We note that in Step 3 the method of Freedman and Lane (1983) sums the permuted residuals $R^{\pi(C)}$ and the fitted mean under the null hypothesis $\hat{Y}$ to obtain a new vector of responses. For interaction testing this is not needed because the interaction tests statistics are identical under both approaches.

In moderate to large samples, a good approximation to the distribution of the test statistic under sampling from the null model is the distribution of the test statistic under sampling from the fitted null model. The algorithm for the parametric bootstrap (Bůžková et al. 2011) is as follows:

1. Obtain parameter estimates and fitted mean from the original data by fitting a null-hypothesis model, such as equation (1).
2. Generate new responses from the null-hypothesis model using the fitted mean obtained in Step 1.
3. Compute the test statistic, based on fitting the alternative-hypothesis model such as equation (4) to the responses obtained in Step 2.

Steps 4, 5, and 6 are identical to the permutation C algorithm. If the distribution of the test statistic depends smoothly on the regression parameter values, which is true in all standard examples, the parametric bootstrap approach gives a valid test (Davison and Hinkley 1997, 4.2.3), regardless of dependence of covariates E and G.

2.2 Test statistics and significance

Adopted from Bůžková et al. (2011), we compute test statistic $\hat{\gamma}$ and its corresponding approximately pivotal z-statistic.

In permutation testing the empirical p-value is calculated as

$$\left(1 + \sum_{i=1}^{N} I(|s_i| \geq s_o)\right)/(1+N), \tag{7}$$

and in parametric bootstrap as

$$\sum_{i=1}^{N} I(|s_i| \geq s_o)/N, \tag{8}$$

where s_o is the test statistic from the original data, s_i is the statistic from permutation or bootstrap i, and N is the number of permutations or bootstrapped data sets. We note that the test statistic s_o from the original data is considered a permutation statistic in p-value calculation in (7). It is not used as such in the parametric bootstrap in (8). This is also reflected in the differential denominators in (7) and in (8).

For valid tests under the null hypothesis the empirical p-value is uniformly distributed on the set $i/(N+1)$ or i/N, which for large N is close to the uniform distribution on $(0,1)$. We use quantile-quantile plots to compare the distribution of the computed p-values to the continuous uniform $(0,1)$ ideal. These are plotted on the $-\log_{10}$ scale to emphasize small p-values. We provide tables specifically for p-values of 0.05 and 0.01.

We compare power across methods and statistics by computing the proportion of empirical p-values below 0.05.

Our results are based on 10000 simulations. Within each simulation we took $N = 1000$ resamples. We study sample size $n = 20$, 100, 500, and 2000 for test size and for power.

Simulations were performed in R (R Core Team 2014).

3 Results

3.1 Simulations

Using simulations we explore tests of no interaction in regression models for univariate outcomes in samples of independent individuals.

We evaluate permutation C in linear regression for continuous data and in log-linear regression for count data. For count data we further explore finite sample properties of permutation A, permutation B, and parametric bootstrap. We compare power of viable resampling approaches for continuous, binary, and count data.

We generate G as a binary exposure, such as a genetic polymorphism with dominant or recessive inheritance. We denote $P[G=1]=p_G$. We generate binary E with $P[E=1]=p_E$ and $\text{logit}_{p_E} = a + bG$. Hence, b denotes the log odds ratio of association between G and E. A non-zero b specifies a correlation between G and E. We set $p_G = 0.4, a = \text{logit}(0.2), b = \log(2)$.

We generate the continuous outcomes as $Y \sim N(\mu_Y, 1)$, where the model for the mean μ_Y is (4) and $\alpha = 2, \beta_E = 2$, and $\beta_G = 3$. We generate the count outcomes using a Poisson distribution with mean (5) with $\alpha = 0.6, \beta_E = 0.3$ and $\beta_G = 2$. For binary outcomes, we generate data using model (6) with $\alpha = 0.6, \beta_E = 0.3$, and $\beta_G = 3$. Additionally, we generate skewed continuous outcome with identical mean model but χ^2_k errors with $k = 3$ degrees of freedom. Detailed results for the skewed continuous data are reported in the Appendix. To study accuracy, we simulate data under the null hypothesis with the interaction coefficient $\gamma = 0$.

To study power we allow γ to vary across sample sizes. We set $\gamma = c/\sqrt{n}$, where c is a positive constant. For continuous data we set $c = 10\sqrt{5}$, resulting in $\gamma = 5, \sqrt{5}, 1$, and 0.5 for sample size 20, 100, 500, and 2000, respectively. For count data we set $c = 30\sqrt{5}$, resulting in $\gamma = 15, 3\sqrt{5}, 3$, and 1.5. For binary data we set $c = 60\sqrt{5}$, resulting in $\gamma = 30, 6\sqrt{5}, 6$, and 3. We study efficiency only in approximately accurate approaches.

3.2 Simulations results – Type I error rate

Figure 1 shows the typical results when using permutation C for continuous outcome in linear regression, specifically for sample size of 500. Table 1 summarizes results for nominal levels of 0.05 and 0.01 across all

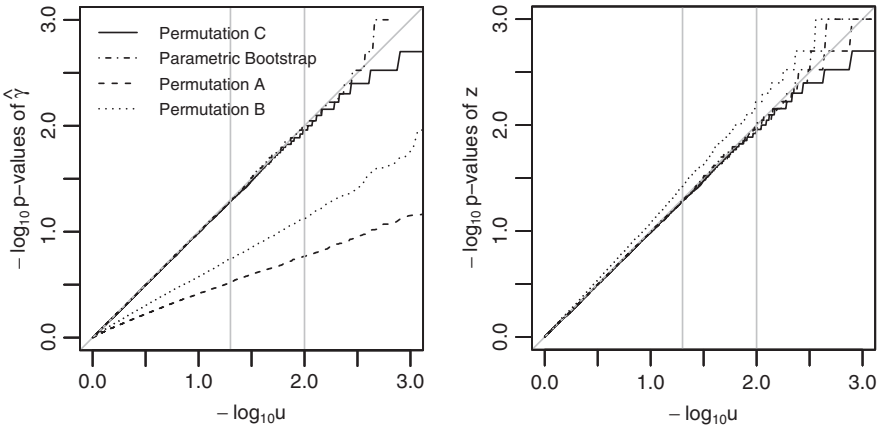

Figure 1: QQ plots for $-\log_{10}$ of p-values of $\hat{\gamma}$ and z statistics for continuous data under four resampling approaches for $n = 500$.

Table 1: P-values of $\hat{\gamma}$ and z statistics for continuous data for nominal size of 0.05 and 0.01 under four resampling approaches.

n	Permutation C		Permutation A		Permutation B		Parametric Bootstrap	
	$\hat{\gamma}$	z	$\hat{\gamma}$	z	$\hat{\gamma}$	z	$\hat{\gamma}$	z
$\alpha = 0.05$								
20	0.035	0.045	0.287	0.052	0.188	0.048	0.053	0.048
100	0.048	0.051	0.301	0.052	0.160	0.035	0.051	0.050
500	0.052	0.052	0.301	0.052	0.180	0.038	0.051	0.051
2000	0.052	0.052	0.295	0.050	0.179	0.036	0.048	0.048
$\alpha = 0.01$								
20	0.006	0.008	0.147	0.009	0.050	0.005	0.014	0.009
100	0.009	0.010	0.171	0.010	0.057	0.003	0.010	0.010
500	0.011	0.011	0.170	0.010	0.074	0.006	0.010	0.009
2000	0.010	0.010	0.180	0.011	0.084	0.007	0.010	0.010

sample sizes. When using approximately pivotal statistic z, it proves to be a valid approximate approach across all sample sizes, including $n = 20$. When using the non-pivotal statistic $\hat{\gamma}$, the test is slightly anti-conservative for $n = 20$, i. e. too small p-values and too large Type I error rates. The accuracy of the p-values increases with sample size. It is approximately valid for $n = 100$ and larger samples.

When comparing permutation C permuting residuals with the permutation A and B permuting the response, it always outperforms them for the non-pivotal statistic $\hat{\gamma}$. Note, permutations type A and B provide largely conservative results for $\hat{\gamma}$ across all sample sizes (Bůžková et al. 2011). When using the approximately pivotal statistic z, permutations C and A perform comparably well and provide approximately valid answers over the whole range of sample sizes. Permutation B is slightly anti-conservative. Parametric bootstrap and permutation C perform comparably across all sample sizes for both statistics $\hat{\gamma}$ and z. For moderate and large samples and for both statistics both approaches provide test size that is approximately nominal.

Permutation C with skewed data takes a larger sample size to provide approximately nominal levels. When using z-statistic, the test is slightly anti-conservative for $n = 20$, but for $n = 100$ it is at its nominal level. When using $\hat{\gamma}$ statistic, the test is anti-conservative for $n = 20$ and still slightly anti-conservative for $n = 100$. For permutation A, using z-statistic provides approximately valid answers starting at sample size of $n = 500$. For smaller samples permutation A is anti-conservative. Permutation B stays anti-conservative across all sample sizes. When using $\hat{\gamma}$ statistic, permutations A and B show poor performance across the

whole range of sample sizes. The test is conservative across the whole range of sample sizes for permutation A and anti-conservative for small sample size and conservative for larger sample size for permutation B. Parametric bootstrap performs comparably to permutation C, providing valid p-values for sample sizes of $n = 500$ or larger for both approximately pivotal and non-pivotal statistics. It is anti-conservative for $n = 20$ and 100. Appendix Figures 4 and 5 show QQ-plots across all approaches, and Table 4 summarizes results for nominal levels of 0.05 and 0.01.

Figure 2 shows the typical results for log-linear regression in count data under permutation C, specifically for sample size of 500. For nominal level of 0.05 and 0.01, see left part of Table 2. Using z-statistic permutation C seems to be a valid approximate approach across all sample sizes. When using $\hat{\gamma}$ statistic, the test is conservative across the whole range of sample sizes, i. e. too large p-values and too small Type I error rates. In our simulations the method never reached an approximately nominal size, it remained conservative even for a sample size of $n = 10000$. We believe that with partially adjusting for the main effects we decrease the variation only partially. The residual variance is still inflated, making the p-values too large.

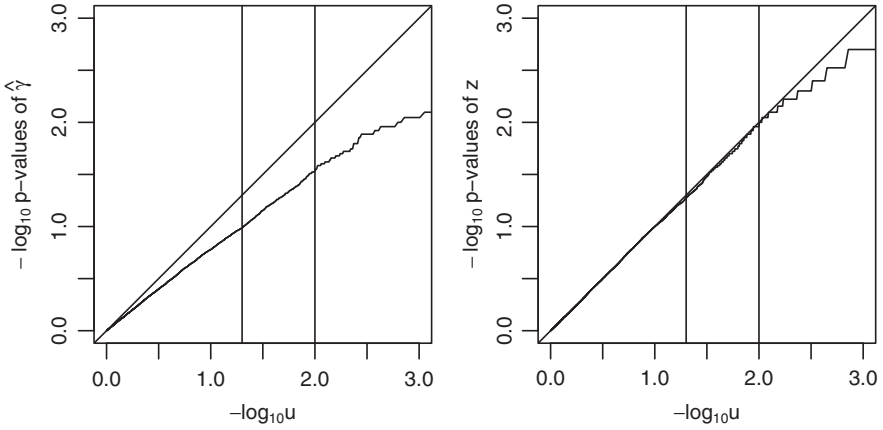

Figure 2: QQ plots for $-\log_{10}$ of p-values of $\hat{\gamma}$ and z statistics for count data under permutation C for $n = 500$.

Table 2: P-values of $\hat{\gamma}$ and z statistics for count data for nominal size of 0.05 and 0.01 under four resampling approaches.

n	Permutation C		Permutation A		Permutation B		Parametric Bootstrap	
	$\hat{\gamma}$	z	$\hat{\gamma}$	z	$\hat{\gamma}$	z	$\hat{\gamma}$	z
$\alpha = 0.05$								
20	0.069	0.044	0.345	0.031	0.275	0.039	0.072	0.050
100	0.094	0.51	0.328	0.049	0.250	0.055	0.051	0.052
500	0.103	0.052	0.336	0.052	0.267	0.056	0.052	0.052
2000	0.102	0.050	0.331	0.051	0.263	0.054	0.051	0.050
$\alpha = 0.01$								
20	0.014	0.008	0.024	0.003	0.136	0.003	0.020	0.010
100	0.025	0.012	0.193	0.008	0.115	0.009	0.012	0.010
500	0.029	0.010	0.197	0.011	0.137	0.012	0.010	0.011
2000	0.030	0.011	0.201	0.011	0.139	0.012	0.010	0.010

Figure 3 demonstrates the typical results for log-linear regression in count data with permutation A, permutation B, and parametric bootstrap, specifically for sample size of 500. The right part of Table 2 summarizes results for nominal levels of 0.05 and 0.01 across all sample sizes. Similar to

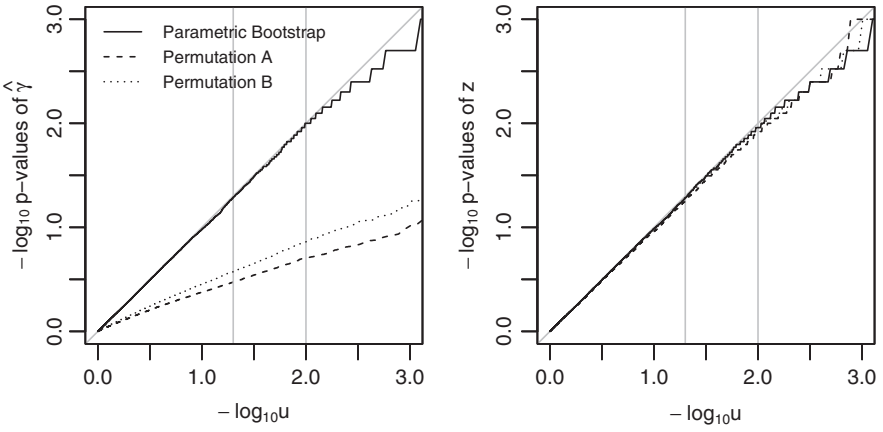

Figure 3: QQ plots for $-\log_{10}$ of p-values of $\hat{\gamma}$ and z statistics for count data for $n = 500$.

continuous outcome, methods A and B for $\hat{\gamma}$ provide conservative tests. The accuracy does not improve with sample size increase. We believe that the permutation tests A and B are conservative because setting the main effects to zero puts the variation explained by the main effects back into the residual variance. We note that permutations A and B are much more conservative than permutation C, which does attempt to decrease the variance due to main effects. For the z-statistic this is not the case. Starting at sample size of 100 both permutation types provide approximately nominal size. Parametric bootstrap proves to be a valid approximate approach throughout, except when using the $\hat{\gamma}$ statistic in a small sample with n = 20.

3.3 Simulations results – power

In Table 3 we compare power across all combinations of resampling methods and statistics that provided approximately valid nominal size. We note that the interaction coefficient varies across sample size.

Table 3: Comparison of power at $\alpha = 0.05$ for continuous, count, and binary data under four resampling approaches and 2 statistics for a range of sample sizes. The interaction coefficients vary with sample size. Available only in approaches that exist and are approximately accurate.

n	Permutation C		Permutation A		Permutation B		Parametric Bootstrap	
	$\hat{\gamma}$	z	$\hat{\gamma}$	z	$\hat{\gamma}$	z	$\hat{\gamma}$	z
Continuous								
20	–	0.960	–	0.938	–	–	0.965	0.942
100	0.968	0.944	–	0.941	–	–	0.980	0.940
500	0.981	0.954	–	0.953	–	–	0.972	0.954
2000	0.970	0.966	–	0.965	–	–	0.972	0.961
Count								
20	–	0.421	–	–	–	–	–	0.398
100	–	0.598	–	0.543	–	0.531	0.760	0.605
500	–	0.625	–	0.531	–	0.592	0.755	0.618
2000	–	0.643	–	0.557	–	0.589	0.741	0.611
Binary								
20			–	–	–	–	–	–
100			–	–	–	–	0.988	0.528
500			–	0.411	–	0.374	0.984	0.497
2000			–	0.529	–	0.530	0.999	0.532

In linear regression with continuous data, all considered approaches demonstrate similar power. Specifically for using $\hat{\gamma}$, the parametric bootstrap and permutation C are comparable. With skewed continuous data the findings are similar, see Table 5. We note that simulations in continuous data with two-way designs showed little difference in size and power comparing normal theory tests to residual based permutation (O'Gorman 2012, Ch. 6.3.1).

Simulations for count data in log-linear models suggest that parametric bootstrap with $\hat{\gamma}$ is highly outperforming all other approaches, see middle part of Table 3. The second best performance demonstrates permutation C with z. In logistic regression with binary data, see lower part of Table 3, parametric bootstrap with $\hat{\gamma}$ shows again the most power across all sample sizes, largely dominating all other approaches.

The parametric bootstrap with $\hat{\gamma}$ statistic achieved the highest power across all sample sizes for count and binary data, and the power was similar to other approaches for continuous data.

4 Discussion

Exact permutation tests for interactions are not available in most situations. Bůžková et al. (2011) showed that when using outcome based permutation tests the Type I error rates can be substantial, specially when non-pivotal quantities are used.

We have studied permutations based on null hypothesis residuals. We concluded that, while not exact, such permutation tests always outperformed the outcome based permutation tests. In moderate and large samples Type I error rates were approximately nominal with the pivotal z-statistic in both continuous and count data. When using the non-pivotal $\hat{\gamma}$ accuracy was acceptable in continuous data, but we found the permutation tests to be conservative for count data.

Bůžková et al. (2011) suggested that a practical alternative to permutation tests for interaction testing is the parametric bootstrap and studied it in continuous and binary data. Parametric bootstrap is extendable to count data and we showed it to be an accurate approach with both pivotal and non-pivotal statistics in moderate and large samples. When using the pivotal z-statistic it was valid even for a small sample size of $n = 20$.

While efficiency in continuous data with both normal and skewed errors was similar across all approaches where the size was approximately nominal, parametric bootstrap with the non-pivotal $\hat{\gamma}$ statistic dominated other approximately accurate approaches in count and binary data.

Appendix

Table 4: P-values of $\hat{\gamma}$ and z statistics for skewed continuous data for nominal size of 0.05 and 0.01 under Permutation C, as compared to permutation A and B and the parametric bootstrap.

n	Permutation C		Permutation A		Permutation B		Parametric Bootstrap	
	$\hat{\gamma}$	z	$\hat{\gamma}$	z	$\hat{\gamma}$	z	$\hat{\gamma}$	z
$\alpha = 0.05$								
20	0.031	0.044	0.087	0.045	0.037	0.027	0.045	0.041
100	0.044	0.045	0.089	0.047	0.057	0.037	0.045	0.044
500	0.052	0.052	0.103	0.052	0.074	0.050	0.051	0.051
2000	0.051	0.052	0.101	0.052	0.075	0.047	0.051	0.052
$\alpha = 0.01$								
20	0.003	0.004	0.016	0.005	0.002	0.002	0.007	0.005
100	0.009	0.009	0.020	0.007	0.007	0.004	0.008	0.009
500	0.011	0.011	0.032	0.012	0.017	0.009	0.010	0.010
2000	0.010	0.010	0.032	0.011	0.017	0.009	0.010	0.010

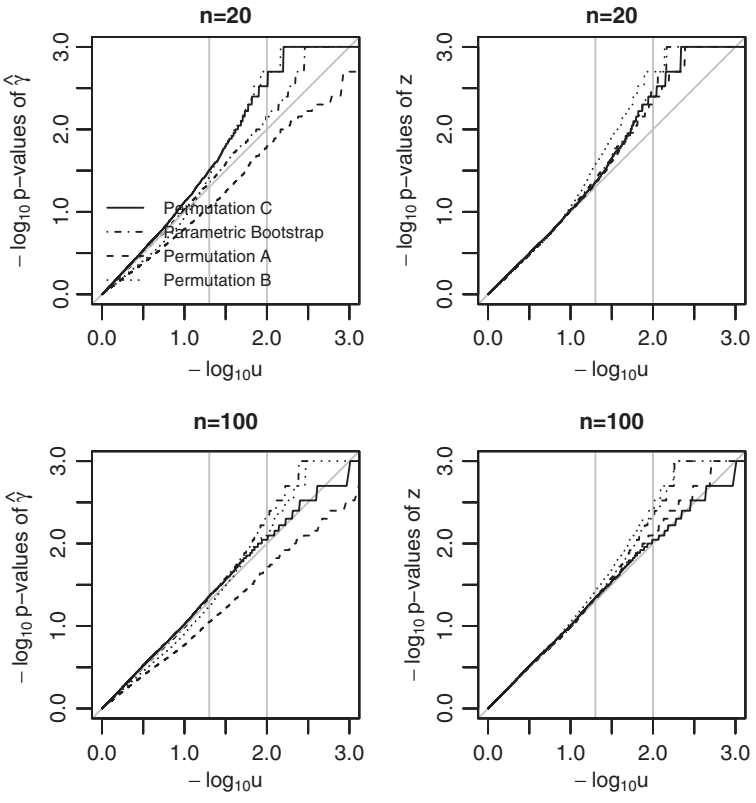

Figure 4: QQ plots for $-\log_{10}$ of p-values of $\hat{\gamma}$ and z statistics for skewed continuous data for $n = 20, 100$.

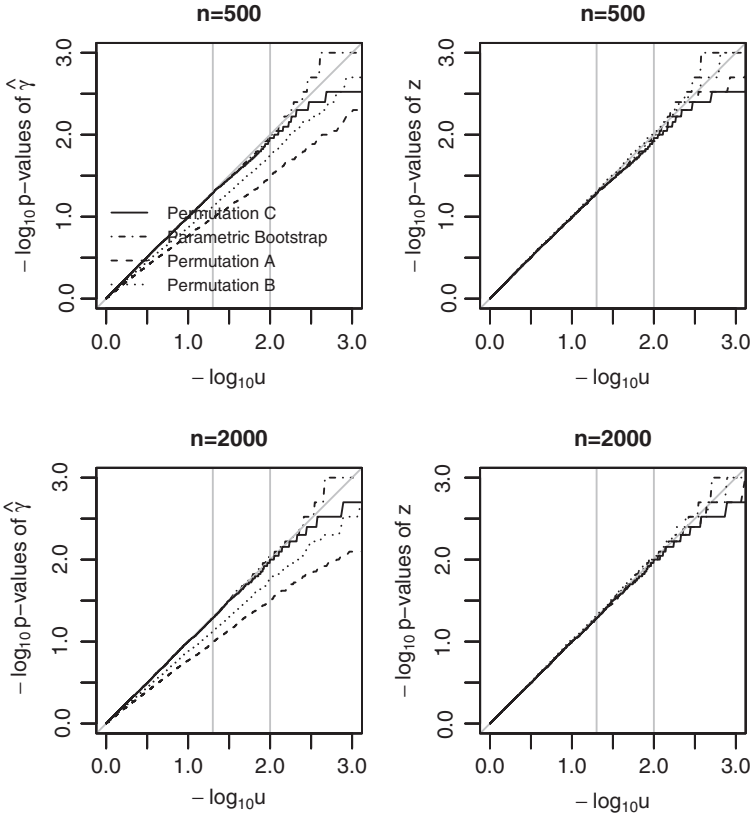

Figure 5: QQ plots for $-\log_{10}$ of p-values of $\hat{\gamma}$ and z statistics for skewed continuous data for $n = 500, 2000$.

Table 5: Comparison of power at $\alpha = 0.05$ for skewed continuous data across 4 methods and 2 statistics for a range of sample sizes. The interaction coefficients vary with sample size.

n	Permutation C		Permutation A		Permutation B		Parametric Bootstrap	
	$\hat{\gamma}$	z	$\hat{\gamma}$	z	$\hat{\gamma}$	z	$\hat{\gamma}$	z
20	–	–	–	–	–	–	–	–
100	–	0.589	–	0.608	–	–	0.622	0.616
500	0.623	0.611	–	0.621	–	0.619	0.630	0.614
2000	0.632	0.633	–	0.631	–	0.632	0.632	0.635

References

Anderson, M. J. (2001). Permutation tests for univariate and multivariate analysis of variance and regression. Canadian Journal of Fisheries and Aquatic Sciences, 58:626–639.

Anderson, M. J., and Robinson, J. (2001). Permutation tests for linear models. Australina & New Zealand Journal of Statistics, 43:75–88.

Bůžková, P., Lumley, T., and Rice, K. (2011). Permutation and parametric bootstrap tests for gene-gene and gene-environment interactions. Annals of Human Genetics, 75:36–45.

Brown, B. M., and Maritz, J. S. (1982). Distribution-free methods in regression. Australian Journal of Statistic, 24:318–331.

Cox, D. R. (1984). Interaction (with discussion. International Statistical Review, 52:1–31.

Cox, D. R., and Hinkley, D. V. (1979). Theoretical Statistics. Boca Raton, FL, USA: CRC Press.

Davison, A. C., and Hinkley, D. V. (1997). Bootstrap Methods and Their Applications. New York, NY, USA: Cambridge University Press.

Ernst, M. D. (2004). Permutation methods: A basis for exact inference. Statistical Science, 19:676–685.

Fisher, R. A. (1935). The Design of Experiments. Edinburgh: Oliver and Boyd.

Freedman, D., and Lane, D. (1983). A nonstochastic interpretation of reported significance levels. Journal of Business and Economic Statistics, 1:292–298.

Higgins, J. J. (2004). An Introduction to Modern Nonparametric Statistics. Pacific Grove, CA: Thomson, Brooks/Cole.

Manly, B. F. J. (1997). Randopmization, Bootstrap and Monte Carlo Methods in Biology. 2nd Edition. London: Chapman & Hall/CRC.

O'Gorman, T. W. (2012). Adaptive Tests of Significance Using Permutations of Residuals with R and SAS. Hoboken, NJ, USA: John Wiley & Sons.

R Core Team (2014). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, http://www.R-project.org/.