British Journal of Mathematical and Statistical Psychology (2012), 65, 282–301

The British Psychological Society

www.wileyonlinelibrary.com

Bootstrap methods for comparing independent regression slopes

Marie Ng[1]* and Rand R. Wilcox[2]

[1]Faculty of Education, The University of Hong Kong
[2]Department of Psychology, University of Southern California, USA

In this study, we explore the effects of non-normality and heteroscedasticity when testing the hypothesis that the regression lines associated with multiple independent groups have the same slopes. The conventional approach involving the *F*-test and the *t*-test (*F*/*t* approach) is examined. In addition, we introduce two robust methods which allow simultaneous testing of regression slopes. Our results suggest that the *F*/*t* approach is extremely sensitive to violations of assumptions and tends to yield misleading conclusions. The new robust alternatives are recommended for general use.

1. Introduction

In social and biological research, it is often of interest to examine whether the relationship between two variables differs across groups – for example, whether the relationship between the risk of obesity and socioeconomic status differs across races, or whether the relationship between the risk of lung cancer and smoking differs across genders. A common way of addressing this issue is to include a multiplicative term in a regression model, estimate the coefficients using ordinary least squares, and then test the coefficient corresponding to the interaction term using classic methods such as the *F*-test and the *t*-test (e.g., Cohen, Cohen, West, & Aiken, 2003; Montgomery, Peck, & Vining, 2006).

To elaborate, consider four independent groups and the linear regression model

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 D_{1i} + \beta_3 X_i D_{1i} + \beta_4 D_{2i} + \beta_5 X_i D_{2i} + \beta_6 D_{3i} + \beta_7 X_i D_{3i} + \epsilon_i, \; i = 1, \ldots, n, \tag{1}$$

where the β_k, $k = 0, 1, \ldots, 7$, are the unknown coefficients, X_i is a continuous explanatory variable and ϵ_i is the error term. The $D_j, j = 1, 2, 3$, are dummy variables used to represent the groups. This model is often referred to as a moderated multiple

*Correspondence should be addressed to Marie Ng, Faculty of Education, The University of Hong Kong, Pok Fu Lam Road, Hong Kong (e-mail: marieng@u.washington.edu)

DOI:10.1111/j.2044-8317.2011.02024.x

Table 1. Example of a dummy coding scheme

	C_1	C_2	C_3
Group 1	0	0	0
Group 2	1	0	0
Group 3	0	1	0
Group 4	0	0	1

regression model (MMR) (e.g., Saunders, 1956; Morris, Sherman, & Mansfield, 1986; Dawson & Richter, 2006). For J independent groups, $J - 1$ dummy variables are used. Table 1 shows an example of the coding specification. Note that under this specification, group 1 is set to be the *base group*. Other coding systems such as unweighted effect codes, weighted effect codes and contrast codes may also be applied. Here, however, we focus our attention on the dummy coding system, which is the most popular approach in applied research.

The regression models for the four groups are: for group 1,

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i; \tag{2}$$

for group 2,

$$Y_i = (\beta_0 + \beta_2) + (\beta_1 + \beta_3) X_i + \epsilon_i; \tag{3}$$

for group 3,

$$Y_i = (\beta_0 + \beta_4) + (\beta_1 + \beta_5) X_i + \epsilon_i; \tag{4}$$

and for group 4,

$$Y_i = (\beta_0 + \beta_6) + (\beta_1 + \beta_7) X_i + \epsilon_i. \tag{5}$$

To determine whether the association between X and Y differs among the groups, it is common to first perform an omnibus test (H_0: $\beta_3 = \beta_5 = \beta_7 = 0$) using the F-test. If H_0 is rejected, one would then test H_0: $\beta_k = 0$, $k = 3, 5, 7$, using the t-test (West, Aiken, & Krull, 1996). Researchers may also perform follow-up tests of any pairwise comparison of interest using contrast coding. If H_0 is rejected for any k, this implies that an interaction exists – that is, the slopes of the groups differ. In this paper, we shall denote this stepwise procedure for comparing pairs of slopes as the F/t approach.

We note that another classic method for comparing regression slopes is the Johnson–Neyman technique (Johnson & Neyman, 1936). Various extensions of this technique have been introduced (Potthoff, 1964; Wilcox, 1987). The Johnson–Neyman technique is particularly useful for computing a region of significance which indicates a set of X values for which the relationship between Y and X differs significantly. However, if the goal is simply to determine whether the groups differ in general, the F/t approach is often preferred (Rogosa, 1980). In this study, we shall focus on the F/t approach.

The effectiveness of the F/t approach for testing interaction has long been under scrutiny. Numerous studies have discussed practical issues, such as sample sizes, multicollinearity and measurement errors, which affect the performance of the test

(e.g., Evans, 1985; Jaccard & Wan, 1995; Paunonen & Jackson, 1988). Many of these studies focused on the interaction between two continuous variables or that between a continuous variable and a categorical variable representing two groups. In this paper, we extend this earlier work and examine the appropriateness of the *F*/*t* approach for testing the interaction between a continuous variable and multiple categorical variables representing more than two groups.

One of the major issues we address is the impact of non-normality and heteroscedasticity on the performance of the *F*/*t* approach. The *F*/*t* approach relies on three assumptions:

1. $E(\epsilon_i) = 0$.
2. $\text{Var}(\epsilon_i) = \sigma^2$ (homoscedasticity).
3. The ϵ_i are independent and identically distributed (i.i.d.) and have a normal distribution.

With respect to the assumption of homoscedasticity, there are three ways in which it can be violated: when the variance of ϵ depends on D in equation (1) but not X; when the variance of ϵ depends on X but not D; when the variance of ϵ depends on both D and X. The first situation is where the groups have unequal variances (i.e. heterogeneity). The second situation is where the groups have equal variances (i.e. homogeneity) yet ϵ is heteroscedastic. In this paper, we use *between-groups (BG) heteroscedasticity* to describe situations where the variance of ϵ is dependent on G and *within-groups (WG) heteroscedasticity* to describe situations where the variance of ϵ is dependent on X.

Most studies examining the properties of the *F*/*t* approach have focused mainly on the violation of the BG homoscedasticity assumption and in settings with two independent groups (e.g., Aguinis & Pierce, 1998; Alexander & DeShon, 1994; Overton, 2001). Little attention has been paid to violations of the WG homoscedasticity and normality assumptions and in settings with multiple groups. Violations of the WG homoscedasticity and normality assumptions are common problems in applied research data. If not handled properly, the validity of statistical results can be severely undermined. For example, when the assumption of WG homoscedasticity is violated, the estimate of the standard error for the ordinary least squares coefficients based on $\text{Var}(\hat{\beta}) = \hat{\sigma}^2(X'X)^{-1}$ is not valid (White, 1980). As a result, standard tests such as the *t*-test no longer control the Type I error probability. With non-normality, the performance of the *t*-test can deteriorate (Wilcox, 1996).

One strategy to tackle violations of assumptions is to apply a test statistic which is relatively robust – for instance, replacing the *t*-test with the quasi-*t*-test based on a heteroscedastic consistent covariance estimator. By substituting the usual estimate of the standard error in the classic *t*-test with a heteroscedastic consistent covariance matrix estimator (HCCM), the quasi-*t*-test offers protection against heteroscedasticity. Several versions of the HCCM have been developed that provide a consistent and an unbiased estimate of the variance of coefficients under heteroscedasticity, including the Huber–White sandwich estimator, HC0 (Huber, 1967; White, 1980); HC1 proposed by Hinkley (1977); HC2 proposed by MacKinnon and White (1985); HC3 proposed by Davidson and MacKinnon (1993); HC4 proposed by Cribari-Neto (2004), and most recently, HC5 proposed by Cribari-Neto, Souza, and Vasconcellos (2007). A nice summary of the various HCCMs can be found in Hayes and Cai (2007).

Studies have suggested that the quasi-*t*-test with HC4 performs satisfactorily under circumstances in which ϵ is heteroscedastic (Cribari-Neto, 2004; Ng & Wilcox, 2010). A wild bootstrap extension of the HCCM-based quasi-*t*-test has also been proposed.

It has been shown to work better than the non-bootstrap version, particularly with small sample sizes (Davidson & Flachaire, 2008; Godfrey, 2006). The application of the wild bootstrap quasi-*t*-test for comparing two independent regression slopes was recently studied (Ng & Wilcox, 2011) and was found to perform well in a wide range of scenarios. However, it has not been extended to settings involving multiple comparisons of more than two groups. In this study, we explore the generalization of the wild bootstrap quasi-*t*-test for pairwise comparisons of regression slopes.

There are three major goals in this paper. The first is to demonstrate the performance of the *F*/*t* approach under varying degrees of WG heteroscedasticity and non-normality. The second is to introduce two alternative methods which apply the wild bootstrap and an HCCM-based covariance matrix estimator to compare multiple regression slopes. The first method is an omnibus test, whereas the second method is a multiple comparison procedure. The final goal is to evaluate the statistical power of the conventional and the new methods. This paper is organized as follows. In Section 2 we provide more detailed description of the the two new bootstrap methods for testing multiple slopes. In Section 3 we delineate the general design of the simulation experiments. In Section 4 we present the results from the simulation experiments. Finally, in Section 5 we provide a conceptual discussion of the implications of the results.

2. Two new HC4-based wild bootstrap methods for comparing multiple regression slopes

2.1. An omnibus test (HC4WOM)

In this section, we describe a new omnibus test of interaction. There are three major features to the new method. First, it involves simultaneous comparison of all possible pairs of regression slopes. Second, it utilizes a generalization of the studentized maximum modulus to control for inflated Type I error rates resulting from multiple comparisons. Finally, the wild bootstrap and an HCCM-based quasi-*t* statistic are applied to tackle heteroscedasticity and non-normality. The details of the method are as follows.

Consider J independent groups. For the jth group, the regression model is

$$Y_{ij} = \beta_{0j} + \beta_{1j} X_{ij} + \epsilon_{ij},\ i = 1, \ldots, n_j. \tag{6}$$

The goal is to test

$$H_0 : \beta_{11} = \ldots = \beta_{1J}. \tag{7}$$

1. For the jth group, estimate regression coefficients using ordinary least squares, yielding $\hat{\beta}_{0j}$ and $\hat{\beta}_{1j}$. Compute the residuals r_{ij} and calculate the HC4 covariance estimate of the coefficient ($\ddot{V}_j$), which is defined as

$$\ddot{V}_j = (X_j' X_j)^{-1} X_j' \text{ diag} \left[\frac{r_{ij}^2}{(1 - h_{ii})^{\delta_i}} \right] X_j (X_j' X_j)^{-1},$$

where X_j is the design matrix of the jth group and the ith row of X_j is denoted by x_{ij}. Let

$$h_{ii} = x_{ij}(X'_j X_j)^{-1} x'_{ij}$$

and

$$\delta_i = \min\left\{4, \frac{h_{ii}}{\bar{h}}\right\} = \min\left\{4, \frac{nh_{ii}}{N_p}\right\},$$

where N_p is the number of regressors, in this case $N_p = 2$. Note that δ_i downweights the influence of high-leverage observations, which tend to have small $r^2{}_i$. The standard error of $\hat{\beta}_{1j}$ is $\ddot{V}_{j22}$, which is the second element along the diagonal of the matrix $\ddot{V}_j$.

2. Compute the quasi-t statistic,

$$T_{jk} = \frac{\hat{\beta}_{1j} - \hat{\beta}_{1k}}{\sqrt{\ddot{V}_{j22} + \ddot{V}_{k22}}}, \quad \text{for all} \quad j < k, \tag{8}$$

then compute the maximum absolute quasi-t statistic,

$$T_{\max} = \max(|T_{jk}|), \quad \text{for all} \quad j < k. \tag{9}$$

3. Construct a bootstrap sample $Y^*_{ij} = \beta_{0j} + \beta_{1j}X_{ij} + a_{ij}r_{ij}$, $i = 1, \ldots, n_j$ and $j = 1, \ldots, J$, where a_{ij} is typically generated from a two-point (lattice) distribution (Liu, 1988),

$$a_{ij} = \begin{cases} -1, & \text{with probability } 0.5, \\ 1, & \text{with probability } 0.5. \end{cases}$$

4. Compute the ordinary least squares estimates $\hat{\beta}^*_{0j}$ and $\hat{\beta}^*_{1j}$, as well as the HC4 covariance estimate for each group ($\ddot{V}^*_j$) based on the bootstrap sample. Compute the bootstrap quasi-t-test statistics (T^*_{jk}) for all pairs of j and k groups, where $j < k$:

$$T^*_{jk} = \frac{(\hat{\beta}^*_{1j} - \hat{\beta}^*_{1k}) - (\hat{\beta}_{1j} - \hat{\beta}_{1k})}{\sqrt{\ddot{V}^*_{j22} + \ddot{V}^*_{k22}}}, \ \forall\ j < k. \tag{10}$$

Then calculate the bootstrap maximum absolute quasi-t statistics:

$$T^*_{max} = \max(|T^*_{jk}|), \quad \text{for all} \quad j < k. \tag{11}$$

5. Repeat steps 3 and 4 B times yielding $T^*_{\max,b}$, $b = 1, \ldots, B$. In the current study, $B = 599$. The choice of $B = 599$ is based on Hall (1986), who suggested that B should be chosen such that $(B + 1)^{-1}$ is a multiple of $1 - \alpha$. From our experience, $B = 599$ is generally sufficient to obtain stable results.

6. The p-value is

$$p = \frac{\#\{T^*_{\max,\ b} \geq T_{\max}\}}{B}.$$

Reject H_0 (7) if $p \leq \alpha$.

2.2. A multiple comparison procedure (HC4WMCP)

Researchers are often interested in not just whether *any* slopes differ, but more specifically which pairs of slopes differ – that is, in testing H_0: $\beta_{1j} = \beta_{1k}$, for all $j < k$. HC4WOM can be extended to a multiple comparison procedure. Essentially, generate the bootstrap distribution of the maximum absolute quasi-t statistics ($T^*_{\max,b}$, $b = 1, \ldots, B$) following steps 1–5 described above. Place $T^*_{\max,b}$ in ascending order, $T^*_{\max,(1)} \leq \ldots \leq T*_{\max,(B)}$. Let $c = (1 - \alpha)B$, rounded to the nearest integer. The confidence interval for $\hat{\beta}_{1j} - \hat{\beta}_{1k}$ is

$$\hat{\beta}_{1j} - \hat{\beta}_{1k} \pm T^*_{\max,(C)}\sqrt{\ddot{V}_{j22} + \ddot{V}_{k22}}. \tag{12}$$

Alternatively, using T_{jk} computed in (8), the p-value is

$$p = \frac{\#\{|T_{jk}| \geq T_{\max}\}}{B}.$$

Reject $H_0 : \hat{\beta}_{1j} = \hat{\beta}_{1k}$ if $p \leq \alpha$.

3. Simulation design

Our simulation experiment consists of three parts. In the first part we compare the omnibus tests, F-test and HC4WOM, in terms of Type I error. In the second part we compare the multiple comparisons procedures, F/t approach and HC4WMCP. Finally, we examine the statistical power of the methods.

3.1. Omnibus tests

The goal of this part of simulation is to identify the impact of non-normality and WG heteroscedasticity on the performance of the omnibus tests, F-test and HC4WOM, in terms of Type I errors. BW homoscedasticity is preserved and sample sizes of the groups are equal. We focus on situations with 4, 6 and 8 independent groups. Data for each group are generated based on the following model:

$$Y_{ij} = \beta_{1j} X_{ij} + \tau(X_{ij})\epsilon_{ij}, \quad i = 1, \ldots, n_j;\ j = 1, \ldots, J, \tag{13}$$

where $J = 4, 6, 8$, X_{ij} is the regressor, ϵ_{ij} is the error term, and τ is a function of X_{ij} used to model the degree of WG heteroscedasticity.

Table 2. Some properties of the *g*-and-*h* distribution. κ_1 and κ_2 denote skewness and kurtosis. $\kappa_1 = \mu_{[3]}/\mu_{[2]}^{1.5}$ and $\kappa_2 = \mu_{[4]}/\mu_{[2]}^{2}$, where $\mu_{[k]} = E(X - \mu)^k$

g	h	μ	σ^2	κ_1	κ_2	$\hat{\kappa}_1$	$\hat{\kappa}_2$
0	0	0	1	0	3.0	0	3.0
0	0.5	0	∞	0	undefined	0	11,986.2
0.5	0	0.2653	1.4588	1.75	8.9	1.81	9.7
0.5	0.5	0.8033	∞	undefined	undefined	120.1	18,393.6

X_{ij} and ϵ_{ij} are generated from the *g*-and-*h* distribution (Hoaglin, 1983). Let Z be a random variable generated from a standard normal distribution, in which case

$$X = \left(\frac{\exp(gZ) - 1}{g}\right) \exp\left(\frac{hZ^2}{2}\right) \tag{14}$$

has a *g*-and-*h* distribution. When $g = 0$, this last equation is taken to be $X = Z \exp(hZ^2/2)$. When $g = 0$ and $h = 0$, X has a standard normal distribution. Skewness and heavy-tailedness of the *g*-and-*h* distribution are determined by the values of g and h, respectively: as the value of g increases, the distribution becomes more skewed; as the value of h increases, the distribution becomes more heavy-tailed. Four types of distributions are considered for X_i: standard normal ($g = 0$, $h = 0$), asymmetric light-tailed ($g = 0.5$, $h = 0$), symmetric heavy-tailed ($g = 0$, $h = 0.5$) and asymmetric heavy-tailed($g = 0.5$, $h = 0.5$). More properties of the *g*-and-*h* distribution can be found in Table 2. The error term (ϵ_{ij}) is also randomly generated based on one of these four *g*-and-*h* distributions. When *g*-and-*h* distributions are asymmetric ($g = 0.5$), the mean is non-zero. Therefore, ϵ_{ij} generated from these distributions are recentred to have mean zero.

To simulate WG heteroscedasticity, five choices for $\tau(X_{ij})$ are considered: $\tau(X_{ij}) = 1$, $\tau(X_{ij}) = \sqrt{|X_{ij}|}$, $\tau(X_{ij}) = |X_{ij}|$, $\tau(X_{ij}) = 1 + 2/(|X_{ij}| + 1)$ and $\tau(X_{ij}) = |X_{ij} + 1|$. We refer to these variance pattern (VP) functions as VP1, VP2, VP3, VP4 and VP5, respectively. VP1 represents homoscedasticity. VP2, ..., VP5 represent a particular pattern of variability in Y_{ij} based upon the value of X_{ij}.

All possible pairs of X_{ij} and ϵ_{ij} distributions are considered, resulting in a total of 16 sets of distributions. All five variance patterns are used for each set of distributions. Hence, a total of 80 simulated conditions are considered. The probability of a Type I error is based on 10,000 replications, with 20 observations in each group. The actual Type I error probability, when testing at $\alpha = .05$, is estimated as $\hat{\alpha}$, the proportion of *p*-values less than or equal to .05.

As one of the reviewers pointed out, for the current findings to be relevant to everyday data analysis, it is crucial that the simulation scenarios are representative of real-life data. In terms of skewness, while the magnitude in a quarter of the simulation data may be more extreme than real-life data, the magnitude in the rest of the simulation data is more moderate than many real-life data. For instance, the estimated skewness of the skewed light-tailed distribution ($g = 0$, $h = 0.5$) is 1.81. It has been suggested that many random variables in actual data have estimated skewness greater than 3 (Wilcox, 1990).

With regard to kurtosis, the estimated kurtosis of the simulation data ranges from 3 to 18,393.6. Although in actual data the degree of kurtosis may not be as extreme as in some of the simulation scenarios considered, it can be severe. For example in an HIV

sentinel surveillance data set obtained from India's National AIDS Control Organisation,[1] the estimated kurtosis of certain variables is as high as 24.24, which is 8 times that of a normal distribution. Severe kurtosis is therefore an issue of practical concern.

Finally, the degree of heteroscedasticity in the current simulation is also consistent with real-life data. Using the ratio between maximum and minimum conditional variances of Y given X (i.e. ratio = $\frac{\max\{\mathrm{Var}(Y|X)\}}{\min\{\mathrm{Var}(Y|X)\}}$), we compared the degree of heteroscedasticity between the simulation data and two real-life data sets. The ratios for the various simulation data range from 3.886 to 88.894. In contrast, in a data set from a reading study (data available upon request), the ratio is as high as 154.9. Moreover, in the HIV sentinel surveillance data mentioned earlier, the ratio is as high as 313.237. It appears that the severity of heteroscedasticity in actual data may even exceed that in the current simulation.

We acknowledge that some of the scenarios may appear to be extreme. Nevertheless, if a method performs well under apparently extreme conditions, this provides some reassurance that it will perform well in realistic situations, including situations where departures from normality and other standard assumptions are more severe than might be anticipated.

3.2. Multiple comparisons

In this part of the simulation, we compare the performance of the *F/t* approach and HC4WMCP when testing H_0: $\beta_{11} = \beta_{1j}$, for all $j \neq 1$, under WG heteroscedasticity and non-normality. We focus on this test, which compares the base group with the other groups, rather than H_0: $\beta_{1j} = \beta_{1k}$, for all $j \neq k$, because the former hypothesis is immediately tested by the dummy coding system considered here for the *F/t* approach.

Data are generated in the same manner as described above. BW homoscedasticity is preserved and sample sizes of the groups are equal. Scenarios with 4, 6 and 8 independent groups are considered. Since multiple hypotheses are being tested simultaneously, the two methods are evaluated based on *familywise error rate* (FWE), which is the probability of at least one Type I error. A total of 10,000 replications are carried out. FWE is estimated by $\widehat{FWE}$, the proportion of times at least one of the three hypotheses is rejected.

3.3. Statistical power

In this part of simulation, we examine the performance of the *F*-test and HC4WOM in terms of their statistical power. We consider the scenario where all the assumptions are met. That is, X_{ij} and ϵ_{ij} are randomly generated from a standard normal distribution ($g = h = 0$), VP1, and sample sizes are equal among the groups. We focus on six independent groups and a situation where group 1, the base group, has a positive slope ($\beta_{11} \in [0, 2]$), and groups 2–6 have zero slopes ($\beta_{1j} = 0, j = 2, \ldots, 6$). The maximum difference between the slopes of *any* two groups ($M_{\mathrm{diff}} = \max\{\beta_{1j} - \beta_{1k}\}$) ranges from 0 to 2. We are interested in the statistical power of the methods as a function of M_{diff} when testing H_0: $\beta_{11} = \ldots = \beta_{16} = 0$.

We repeat the same simulated scenarios as described above and compare the statistical power of the *F/t* approach and HC4WMCP. In this case, we are interested in the power of the methods when testing the hypotheses H_0: $\beta_{11} = \beta_{1j}, j = 2, \ldots, 6$.

[1] http://www.nacoonline.org/Quick_Links/Directory_of_HIV_Data

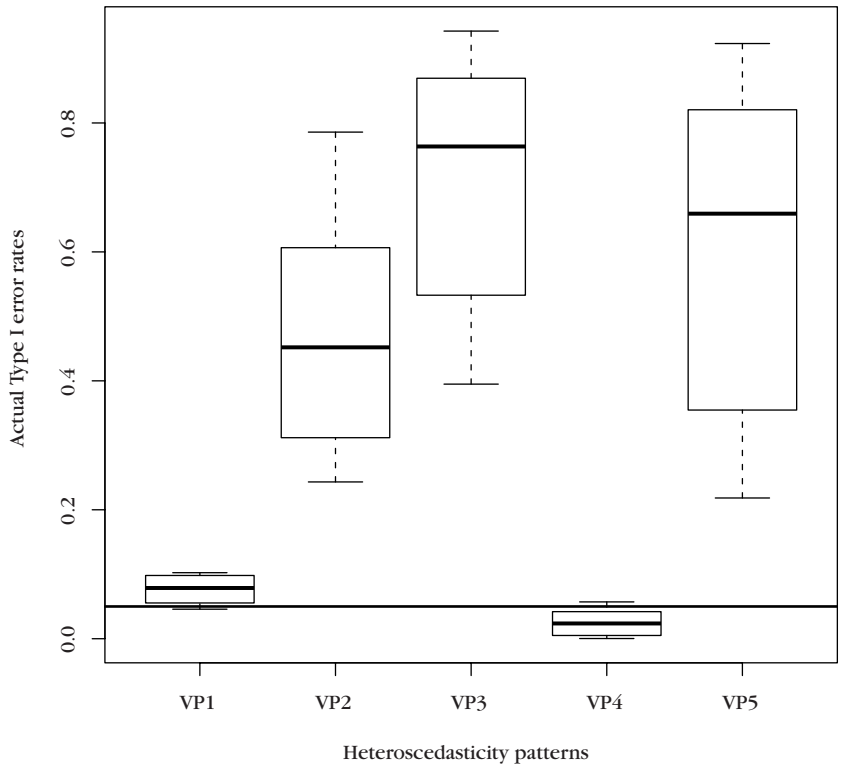

Figure 1. Actual Type I error probabilities for the *F*-test with 6 independent groups across the five variance patterns when testing at $\alpha = .05$. The solid horizontal line indicates $\alpha = .05$.

4. Simulation results

4.1. Omnibus tests

Since the simulation results for 4, 6 and 8 groups are similar, we present only the results for 6 independent groups. In general, the *F*-test performs well when ϵ is homoscedastic and light-tailed ($h = 0$). The actual Type I error rates are close to the nominal level (see Figure 1 and Table 3). However, when the assumption of homoscedasticity is violated and when the distribution of X is heavy-tailed, the actual Type I error rates deviate from the nominal level. For instance, when the distributions of X is skewed and heavy-tailed ($g = 0.5$, $h = 0.5$) and ϵ is normally distributed, under VP3, the actual Type I error rate is .943. On the other hand, under VP4, the Type I error rate is 0.

In contrast, the new bootstrap omnibus test (HC4WOM) performs well in general. As shown in Figure 2 and Table 4, the average actual Type I error rate is .049. However, HC4WOM tends to be slightly liberal under VP5 and slightly conservative under VP4. Under VP5 when X is from a skewed light-tailed distribution ($g = 0.5$, $h = 0$) and ϵ is normally distributed ($g = 0$, $h = 0$) the actual Type I error rate is .066. On the other hand, under VP4 when X and ϵ are from a symmetric heavy-tailed distribution ($g = 0$, $h = 0.5$), the actual Type I error rate is .032. Nevertheless, the actual Type I error rates yielded by the method are within desirable limits in most situations. Overall, the performance of HC4WOM is relatively stable.

Table 3. Actual Type I error rates when testing at $\alpha = .05$, 6 independent groups, $n = 20$, using the *F*-test

X		ϵ		$\hat{\alpha}$				
g	*h*	*g*	*h*	VP1	VP2	VP3	VP4	VP5
0	0	0	0	.050	.316	.533	.007	.298
0	0	0.5	0	.057	.308	.524	.013	.284
0	0	0	0.5	.085	.256	.427	.045	.233
0	0	0.5	0.5	.094	.243	.395	.057	.218
0	0.5	0	0	.054	.745	.926	.001	.895
0	0.5	0.5	0	.072	.679	.905	.005	.864
0	0.5	0	0.5	.101	.491	.803	.035	.720
0	0.5	0.5	0.5	.102	.437	.758	.042	.654
0.5	0	0	0	.050	.471	.769	.006	.665
0.5	0	0.5	0	.059	.439	.732	.011	.623
0.5	0	0	0.5	.091	.327	.586	.042	.463
0.5	0	0.5	0.5	.095	.290	.533	.050	.412
0.5	0.5	0	0	.046	.786	.943	.000	.923
0.5	0.5	0.5	0	.070	.722	.920	.004	.897
0.5	0.5	0	0.5	.101	.534	.834	.037	.777
0.5	0.5	0.5	0.5	.102	.465	.794	.040	.721
		Max		.102	.786	.943	.057	.923
		Min		.046	.243	.395	.000	.218
		Average		.077	.469	.711	.025	.603

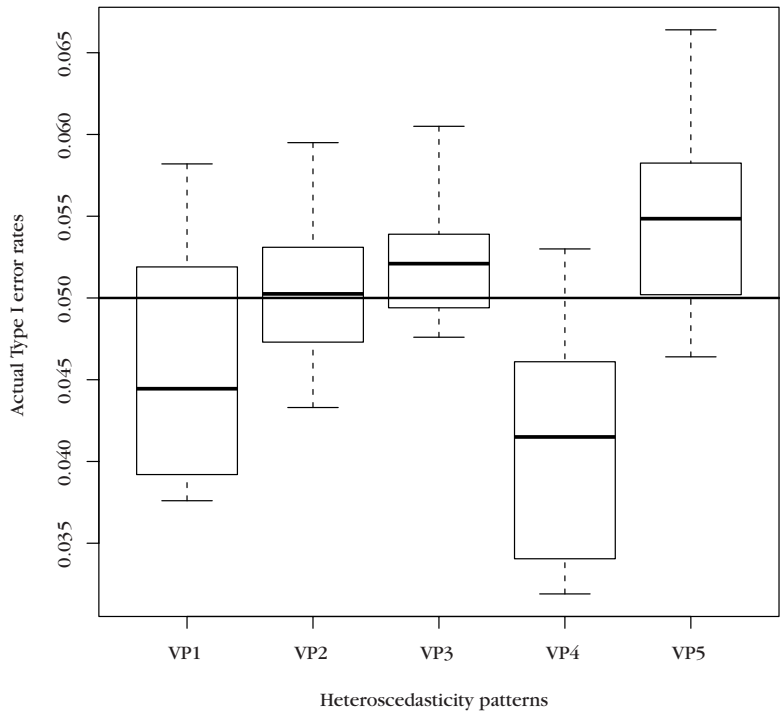

Figure 2. Actual Type I error probabilities for HC4WOM with 6 independent groups across the five variance patterns when testing at $\alpha = .05$. The solid horizontal line indicates $\alpha = .05$.

Table 4. Actual Type I error rates when testing at $\alpha = .05$, 6 independent groups, $n = 20$, using HC4WOM

X		ϵ		$\hat{\alpha}$				
g	h	g	h	VP1	VP2	VP3	VP4	VP5
0	0	0	0	.053	.053	.051	.050	.055
0	0	0.5	0	.049	.051	.053	.045	.050
0	0	0	0.5	.043	.047	.049	.041	.053
0	0	0.5	0.5	.039	.043	.048	.034	.050
0	0.5	0	0	.054	.051	.050	.049	.056
0	0.5	0.5	0	.045	.050	.052	.042	.047
0	0.5	0	0.5	.039	.050	.048	.032	.046
0	0.5	0.5	0.5	.039	.045	.048	.033	.050
0.5	0	0	0	.058	.060	.061	.053	.066
0.5	0	0.5	0	.048	.055	.058	.043	.059
0.5	0	0	0.5	.044	.052	.053	.037	.060
0.5	0	0.5	0.5	.039	.048	.050	.034	.057
0.5	0.5	0	0	.052	.053	.056	.047	.061
0.5	0.5	0.5	0	.052	.053	.055	.045	.055
0.5	0.5	0	0.5	.044	.049	.052	.038	.054
0.5	0.5	0.5	0.5	.038	.047	.053	.032	.055
		Max		.058	.060	.061	.053	.066
		Min		.038	.043	.048	.032	.046
		Average		.046	.050	.052	.041	.055

4.2. Multiple comparisons

In this part of the simulation, we examine the performance of the *F*/*t* procedure and HC4WMCP. As shown in Figure 3 and Table 5, the *F*/*t* procedure performs well under homoscedasticity (VP1) and when ϵ is derived from a light-tailed distribution ($h = 0$). However, when the assumption of homoscedasticity is violated, the actual FWE can substantially deviate from the nominal level. In particular, under VP2, VP3 and VP5, the actual FWEs are substantially higher than the nominal level, whereas under VP4, the actual FWEs are substantially lower than the nominal level. For example, there is a case under VP3 in which $\widehat{FWE}$ is .879 when testing at the $\alpha = .05$ level. On the other hand, under VP4, there is a situation in which $\widehat{FWE}$ is .000. In contrast, the performance of HC4WMCP is fairly satisfactory. The average FWE is close to the nominal level in all situations (see Figure 4 and Table 6).

4.3. Statistical power

The statistical power of the various methods is examined here. We first compare the omnibus tests, the *F*-test and HC4WOM. We consider the situation where normality and homoscedasticity assumptions hold and sample sizes are equal among the groups. We examine the statistical power of the methods as a function of the maximum difference between the slopes of *any* two groups (M_{diff}). Theoretically, when all assumptions hold, the *F*-test should possess the best statistical power among all the tests.

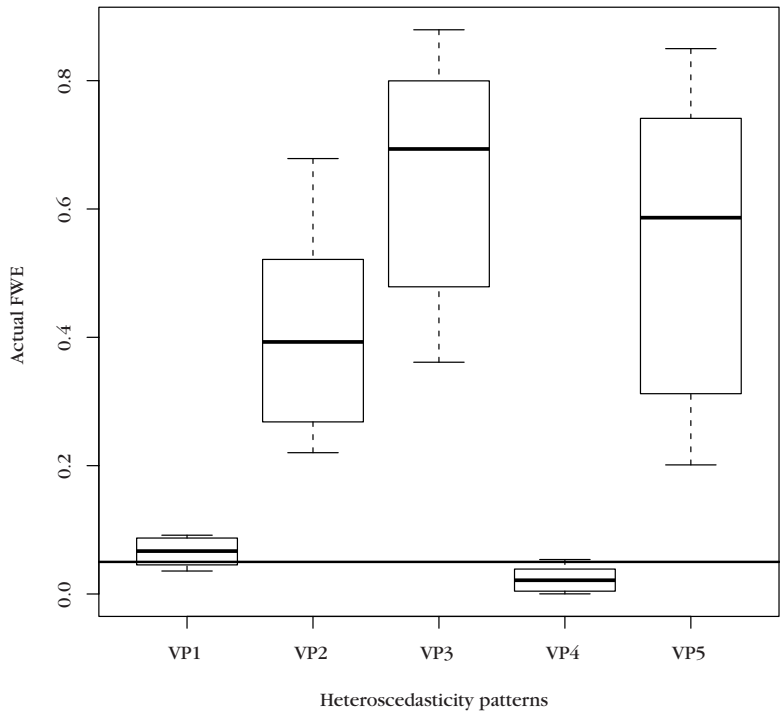

Figure 3. Actual familywise error rates for the *F*/*t* tests with 6 independent groups across the five variance patterns when testing at $\alpha = .05$. The solid horizontal line indicates $\alpha = .05$.

Table 5. Actual familywise error rates when testing at $\alpha = .05$, 6 independent groups, $n = 20$, using the *F*/*t*approach.

X		ϵ		$\hat{\alpha}$				
g	*h*	*g*	*h*	VP1	VP2	VP3	VP4	VP5
0	0	0	0	.041	.269	.472	.006	.249
0	0	0.5	0	.048	.268	.466	.011	.246
0	0	0	0.5	.076	.226	.383	.041	.211
0	0	0.5	0.5	.086	.220	.361	.054	.201
0	0.5	0	0	.042	.640	.869	.001	.823
0	0.5	0.5	0	.058	.584	.843	.004	.790
0	0.5	0	0.5	.088	.421	.732	.032	.640
0	0.5	0.5	0.5	.092	.381	.685	.039	.579
0.5	0	0	0	.040	.403	.702	.005	.594
0.5	0	0.5	0	.049	.383	.672	.009	.556
0.5	0	0	0.5	.083	.291	.537	.039	.420
0.5	0	0.5	0.5	.087	.260	.486	.047	.375
0.5	0.5	0	0	.036	.679	.879	.000	.850
0.5	0.5	0.5	0	.054	.616	.849	.004	.817
0.5	0.5	0	0.5	.088	.459	.756	.032	.692
0.5	0.5	0.5	0.5	.090	.403	.714	.036	.634
		Max		.092	.679	.879	.054	.850
		Min		.036	.220	.361	.000	.201
		Average		.066	.406	.650	.022	.542

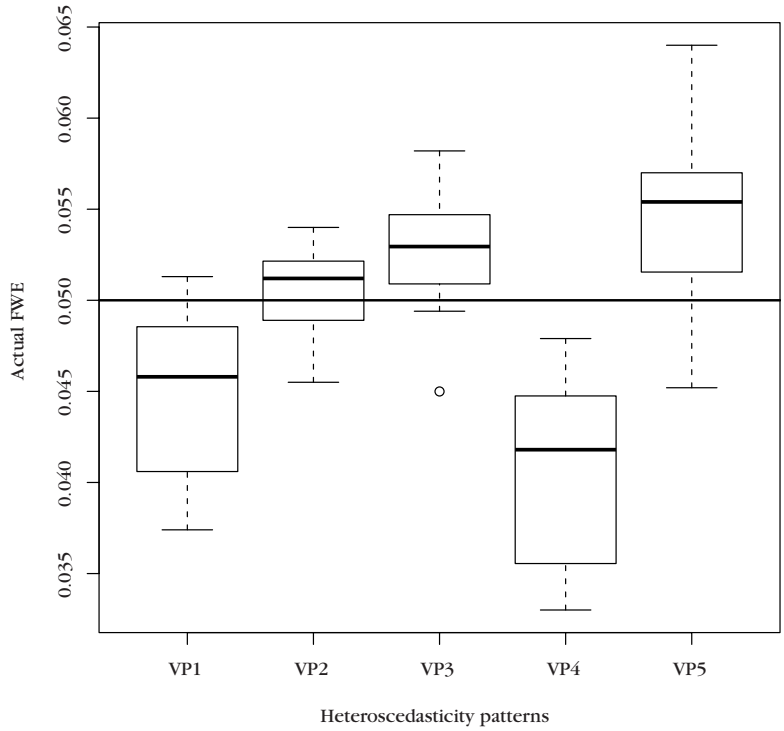

Figure 4. Actual familywise error rates for HC4WMCP with 6 independent groups across the five variance patterns when testing at $\alpha = .05$. The solid horizontal line indicates $\alpha = .05$.

Table 6. Actual familywise error rates when testing at $\alpha = .05$, 6 independent groups, $n = 20$, using HC4WMCP

X		ϵ		$\hat{\alpha}$				
g	b	g	b	VP1	VP2	VP3	VP4	VP5
0	0	0	0	.048	.053	.055	.048	.058
0	0	0.5	0	.050	.051	.051	.047	.051
0	0	0	0.5	.045	.048	.052	.042	.055
0	0	0.5	0.5	.044	.048	.049	.043	.051
0	0.5	0	0	.049	.051	.053	.046	.052
0	0.5	0.5	0	.048	.053	.051	.042	.050
0	0.5	0	0.5	.041	.051	.050	.034	.056
0	0.5	0.5	0.5	.040	.046	.045	.033	.045
0.5	0	0	0	.051	.052	0.056	.048	.061
0.5	0	0.5	0	.048	.052	.055	.044	.059
0.5	0	0	0.5	.047	.054	.058	.040	.064
0.5	0	0.5	0.5	.040	.048	.053	.035	.056
0.5	0.5	0	0	.051	.051	.058	.042	.056
0.5	0.5	0.5	0	.044	.051	.053	.040	.054
0.5	0.5	0	0.5	.038	.051	.055	.036	.055
0.5	0.5	0.5	0.5	.037	.050	.053	.034	.056
		Max		.051	.054	.058	.048	.064
		Min		.037	.046	.045	.033	.046
		Average		.045	.051	.053	.041	.055

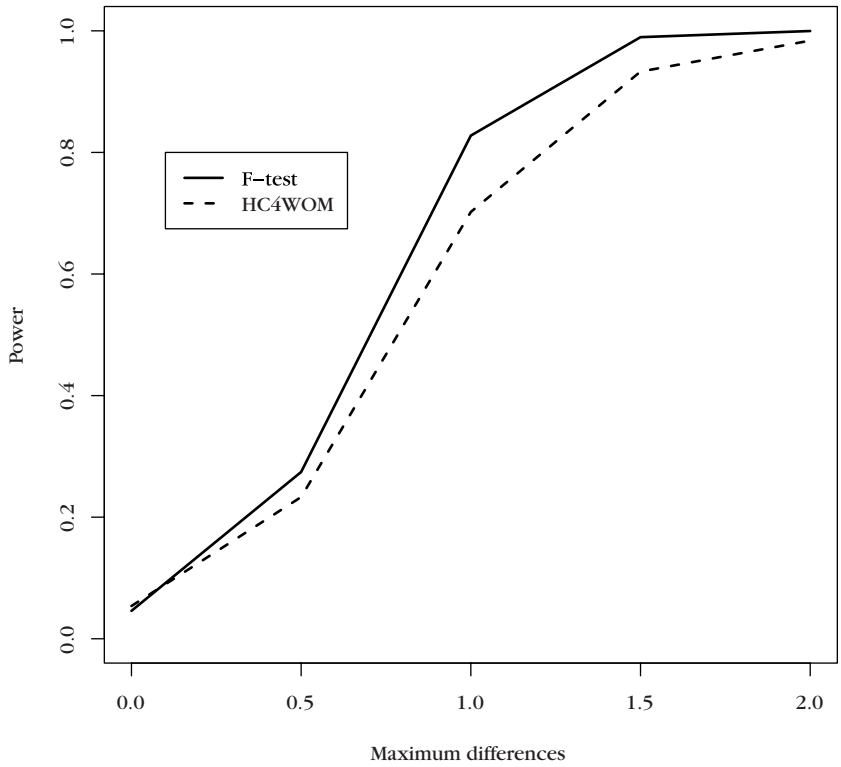

Figure 5. Statistical power as a function of the maximum difference ($M_{\text{diff}} \in [0, 2]$) between any two groups for the F-test and HC4WOM when testing H_0: $\beta_{11} = \ldots = \beta_{16}$. X and ϵ are normally distributed and homoscedasticity is present.

As shown in Figure 5, in the situation where group 1 has a positive slope and groups 2–6 have zero slopes, the F-test has slightly better power than HC4WOM. Nonetheless, HC4WOMC still performs reasonably well.

We focus on HC4WOM and consider the situation where X has a skewed heavy-tailed distribution ($g = h = 0.5$), ϵ has a symmetric heavy-tailed distribution ($g = 0$, $h = 0.5$) and WG heteroscedasticity (VP4) is present. Sample sizes of the groups are equal. This is a situation in which HC4WOM appears to be slightly conservative in terms of control over Type I errors. We are interested in how this affects the statistical power of the method and how quickly statistical power can be recovered with increased sample sizes. We consider the case in which Group 1 has a positive slope and groups 2–6 have zero slopes. As shown in Figure 6, when the assumptions of normality and homoscedasticity are violated, the statistical power is relatively low with small M_{diff} value and small sample sizes. With $n_j = 20$, the statistical power remains below .5 even when $M_{\text{diff}} = 2$. Nevertheless, statistical power increases readily with increased sample sizes. When $M_{\text{diff}} = 2$, power increases from .38 for $n_j = 20$ to .67 for $n_j = 40$ and is over .8 with $n_j = 60$.

Next, we compare the statistical power of the F/t approach and HC4WMCP when testing H_0: $\beta_{11} = \beta_{1j}, j = 2, \ldots, 6$. As shown in Figure 7, in the situation where group 1 has a positive slope and groups 2–6 have zero slopes, the statistical power of the F/t approach is slightly superior to HC4WMCP. However, HC4WMCP still performs satisfactorily.

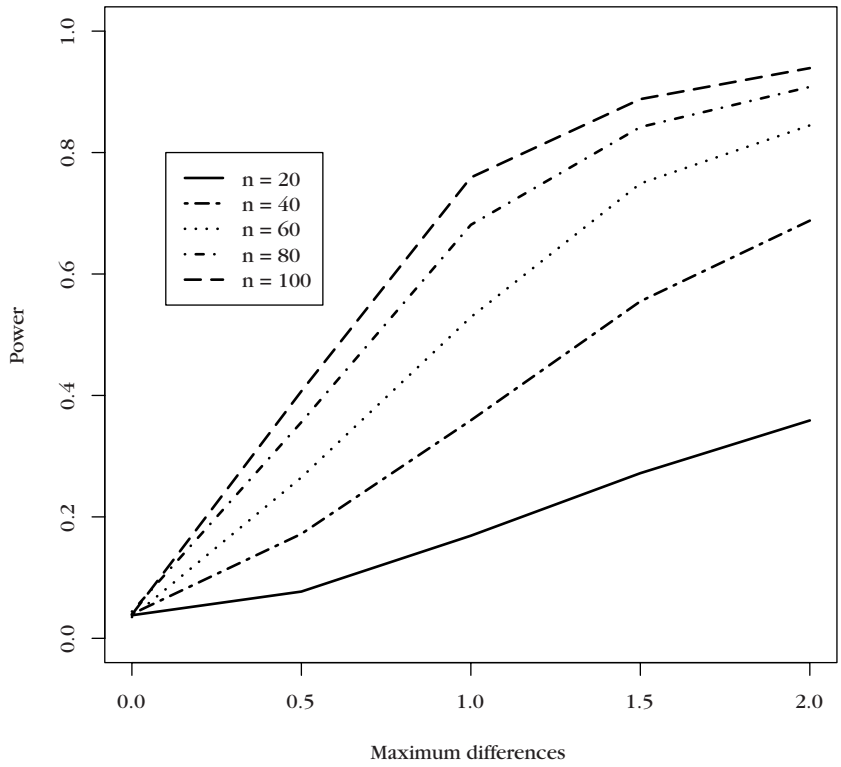

Figure 6. Statistical power as a function of the maximum difference ($M_{\text{diff}} \in [0, 2]$) between any two groups for HC4WOM when testing H_0: $\beta_{11} = \ldots = \beta_{16}$. X is generated from an asymmetric distribution ($g = h = 0.5$) and ϵ is generated from a symmetric distribution ($g = 0$, $h = 0.5$) with VP4.

While the comparison of slopes in the F/t approach is limited by the particular coding scheme used, the comparison of slopes in HC4WMCP is not. HC4WMCP is designed to allow simultaneous comparisons of all pairs of slopes. We examine the statistical power of HC4WMCP for detecting *any* pairwise difference (i.e. when testing H_0: $\beta_{1j} = \beta_{1k}$, for all $j \neq k$). We modify the simulation scenario slightly, such that groups 1 and 2 have non-zero and unequal slopes and groups 3–6 have zero slopes. The maximum difference between any two slopes remains between 0 to 2. As shown in Figure 8, HC4WMCP has considerable statistical power for detecting any pairwise differences across all sample sizes.

5. Discussion

Tests of interaction are important for detecting whether the association between variables differs among groups of interest. The F/t approach, which consists of an F-test and multiple t-tests, is widely applied in current social and biological sciences research. However, its effectiveness and reliability under violations of the homoscedasticity and normality assumptions have not been carefully examined. In this study, we have assessed the impact of heteroscedasticity and non-normality on the performance of the F/t test of interaction. Furthermore, we have introduced two alternative methods for comparing multiple regression slopes.

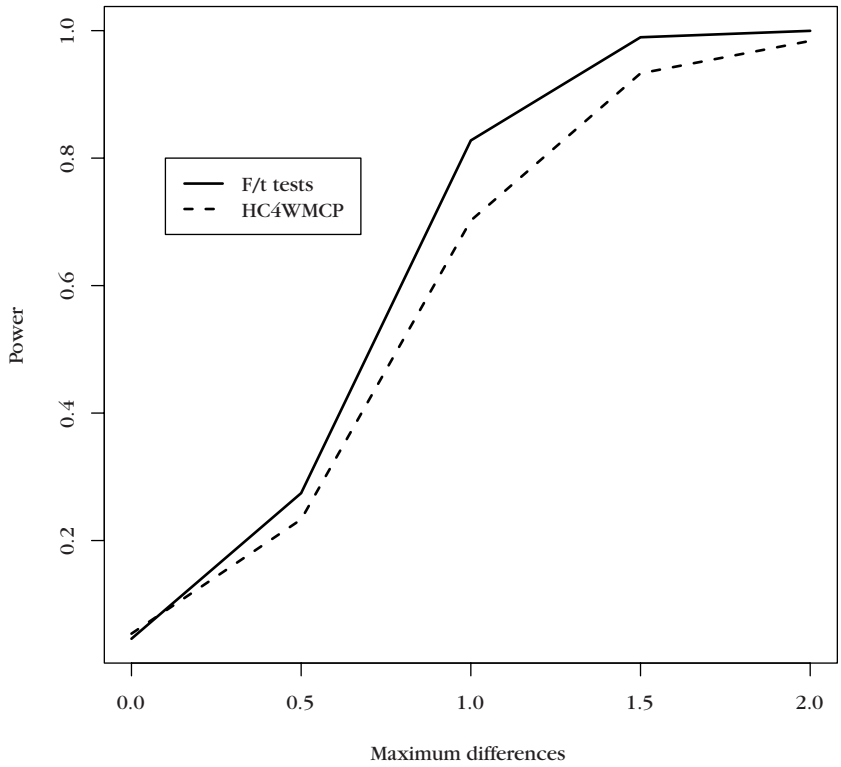

Figure 7. Statistical power as a function of the maximum difference ($M_{\text{diff}} \in [0, 2]$) between any two groups for the *F/t* tests and HC4WMCP when testing H_0: $\beta_{11} = \beta_{1j}, j = 2, \ldots, 6$. X and ϵ are normally distributed and homoscedasticity is present.

Two omnibus tests for comparing multiple regression slopes, namely the *F*-test and HC4WOM, were compared in terms of their Type I error probability and statistical power. The *F*-test appears to be extremely sensitive to violation of the homoscedasticity assumption. It offers little control over the probability of a Type I error when the assumption is violated. The actual Type I error rates when testing at $\alpha = .05$ vary widely. Their performances are further complicated by the presence of non-normality in ϵ. Specifically, when ϵ deviates from a light-tailed distribution, the actual Type I error rates increase. In contrast, the new bootstrap omnibus test (HC4WOM) offers satisfactory control over Type I errors and possesses decent statistical power.

Two tests for comparing pairwise differences between regression slopes, namely the *F/t* approach and HC4WMCP, were compared in terms of their familywise error rates and statistical power. Similarly to the previous results, the performance of the *F/t* approach in terms of control over FWE is severely hampered by heteroscedasticity and non-normality. In particular, the actual FWEs tend to be substantially higher than the nominal level. On the other hand, HC4WMCP appears to be robust against violations of assumptions and performs well under a wide range of situations. HC4WMCP also displays good statistical power for detecting all pairwise differences.

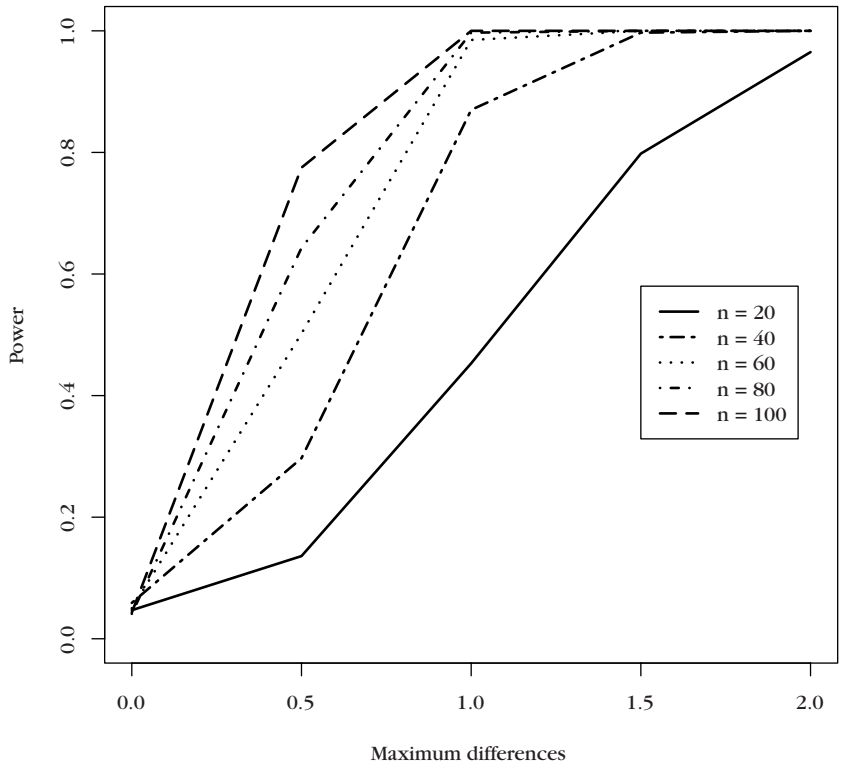

Figure 8. Statistical power as a function of the maximum difference ($M_{\text{diff}} \in [0, 2]$) between any two groups for HC4WMCP when testing H_0: $\beta_{1j} = \beta_{1k}$, for all $j \neq k$. X and ϵ are normally distributed and homoscedasticity is present.

It has been well documented in the literature that tests of interaction in multiple regression may not possess sufficient power (Aguinis, 1995; Aguinis & Stone-Romero, 1997; McClelland & Judd, 1993; Shieh, 2009). The detection of a significant interaction effect is often found to be difficult. The results from the current study indicate otherwise. The power analysis suggests that the *F*-test and the *F*/*t* approach possess decent statistical power for detecting an interaction effect when all assumptions hold. When assumptions are violated, these methods have a strong tendency to conclude a significant effect when there is none. Given that violations of assumptions are prevalent in applied research data, the findings here imply that some previously found significant interaction effects could potentially be spurious.

In summary, two practical implications follow from this study. First, given that heteroscedasticity and non-normality are prevalent in applied research data, using the conventional *F*-test and *t*-test for testing interactions may not be appropriate. It has a strong tendency to yield an erroneous conclusion of significant interaction when none exists. Second, the new methods HC4WOM and HC4WMCP are recommended for general use. They provide desirable control over Type I errors and offer satisfactory statistical power. In addition, unlike the *F*/*t* approach, HC4WMCP allows simultaneous comparisons of all pairs of slopes. An R function for HC4WOM and HC4WMCP is available; more details can be found in the Appendix.

References

Aguinis, H. (1995). Statistical power problems with moderated multiple regression in management research. *Journal of Management*, *21*, 1141–1158. doi:10.1177/014920639502100607

Aguinis, H., & Pierce, C. A. (1998). Heterogeneity of error variance and the assessment of moderating effects of categorical variable: A conceptual review. *Organizational Research Methods*, *1*, 296–314.

Aguinis, H., & Stone-Romero, E. F. (1997). Methodological artifacts in moderated multiple regression and their effects on statistical power. *Journal of Applied Psychology*, *82*, 192–206. doi:10.1037/0021-9010.82.1.192

Alexander, R. A., & DeShon, R. P. (1994). Effect of error variance heterogeneity on the power of tests for regression slope differences. *Psychological Bulletin*, *2*, 308–314. doi:10.1037/0033-2909.115.2.308

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences*. London: Lawrence Erlbaum Associates.

Cribari-Neto, F. (2004). Asymptotic inference under heteroscedasticity of unknown form. *Computational Statistics & Data Analysis*, *45*, 215–233. doi:10.1016/S0167-9473(02)00366-3

Cribari-Neto, F., Souza, T. C., & Vasconcellos, A. L. P. (2007). Inference under heteroskedasticity and leveraged data. *Communication in Statistics – Theory and Methods*, *36*, 1877–1888. doi:10.1080/03610920601126589

Davidson, R., & Flachaire, E. (2008). The wild bootstrap, tamed at last. *Journal of Econometrics*, *146*, 162–169. doi:10.1016/j.jeconom.2008.08.003

Davidson, R., & MacKinnon, J. G. (1993). *Estimation and inference in econometrics*. New York: Oxford University Press.

Dawson, J. F., & Richter, A. W. (2006). Probing three-way interactions in moderated multiple regression: Development and application of a slope difference test. *Journal of Applied Psychology*, *91*, 917–926. doi:10.1037/0021-9010.91.4.917

Evans, M. G. (1985). A monte carlo study of the effects of correlated method variance in moderated multiple regression anaysis. *Organizational Behavior and Human Decision Processes*, *36*, 305–323. doi:10.1016/0749-5978(85)90002-0

Godfrey, L. G. (2006). Tests for regression models with heteroscedasticity of unknown form. *Computational Statistics & Data Analysis*, *50*, 2715–2733. doi:10.1016/j.csda.2005.04.004

Hall, P. (1986). On the number of bootstrap simulations required to construct a confidence interval. *Annals of Statistics*, *14*, 1453–1462.

Hayes, A. F., & Cai, L. (2007). Using heteroscedasticity-consistent standard error estimators in ols regression: An introduction and software implementation. *Behavior Research Methods*, *39*, 709–722.

Hinkley, D. V. (1977). Jackknifing in unbalanced situations. *Technometrics*, *19*, 285–292.

Hoaglin, D. C. (1983). Encyclopedia of statistical sciences. In S. Kotz & N. L. Johnson (Eds.), (pp. 298–301). New York: Wiley.

Huber, P. J. (1967). Proceedings of the fifth berkeley symposium on mathematical statistics and probability. In L. M. Le Cam & J. Neyman (Eds.), (Chap. The behavior of maximum likelihood estimation under nonstandard conditions). Berkeley, CA: Univeristy of California Press.

Jaccard, J., & Wan, C. K. (1995). Measurement error in the analysis of interaction effects between continuous predictors using multiple regression: Multiple indicator and structural equation approaches. *Psychological Bulletin*, *117*, 348–357. doi:10.1037/0033-2909.117.2.348

Johnson, P. O., & Neyman, J. (1936). Tests of certain linear hypothesses and their applications to some educational problems. *Statistical Research Memoirs*, *1*, 57–93.

Liu, R. Y. (1988). Bootstrap procedures under some non i.i.d. models. *The Annuals of Statistics*, *16*, 1696–1708.

MacKinnon, J. G., & White, H. (1985). Some heteroscedasticity consistent covariance matrix estimators with improved finite sample properties. *Journal of Econometrics*, *29*, 305–325. doi:10.1016/0304-4076(85)90158-7

McClelland, G. H., & Judd, C. M. (1993). Statistical difficulties of detecting interactions and moderator effects. *Psychological Bulletin*, *114*, 376–390. doi:10.1037/0033-2909.114.2.376

Montgomery, D. C., Peck, E. A., & Vining, G. G. (2006). *Introduction to linear regression analysis*. New York: Wiley.

Morris, J. H., Sherman, J. D., & Mansfield, E. R. (1986). Failures to detect moderating effects with ordinary least squares-moderated multiple regression: Some reasons and a remedy. *Psychological Bulletin*, *99*, 282–288. doi:10.1037/0033-2909.99.2.282

Ng, M., & Wilcox, R. R. (2010). Comparing regression slopes of two independent groups. *British Journal of Mathematical and Statistical Psychology*, *63*, 319–340. doi:10.1348/000711009X456845

Ng, M., & Wilcox, R. R. (in press). Level robust methods based on the least squares regression estimator. *Journal of Modern Applied Statistical Methods*.

Overton, R. C. (2001). Moderated multiple regression for interaction involving categorical variables: A statistical control for heterogeneous variance across two groups. *Psychological Methods*, *6*, 218–233. doi:10.1037/1082-989X.6.3.218

Paunonen, S. V., & Jackson, D. N. (1988). Type I error rates for moderated multiple regression analysis. *Journal of Applied Psychology*, *73*, 569–573. doi:10.1037/0021-9010.73.3.569

Potthoff, R. F. (1964). On the Johnson–Neyman technique and some extensions thereof. *Psychometrika*, *29*, 241–256. doi:10.1007/BF02289721

Rogosa, D. (1980). Copmaring nonparallel regression lines. *Psychological Bulletin*, *2*, 307–321. doi:10.1037/0033-2909.88.2.307

Saunders, D. R. (1956). Moderator variables in prediction. *Educational and Psychological Measurement*, *16*, 209–222. doi:10.1177/001316445601600205

Shieh, G. (2009). Detecting interaction effects in moderated multiple regression with continuous variables: Power and sample size considerations. *Organizational Research Methods*, *12*, 510–528. doi:10.1177/1094428108320370

West, S. G., Aiken, L. S., & Krull, J. L. (1996). Experimental personality designs: Analyzing categorical by continuous variable interactions. *Journal of Personality*, *64*, 1–48. doi:10.1111/j.1467-6494.1996.tb00813.x

White, H. (1980). A heteroskedastic-consistent covariance matrix estimator and a direct test of heteroscedasticity. *Econometrica*, *48*, 817–838.

Wilcox, R. R. (1987). Pairwise comparisons of j independent regression lines over a finite interval, simultaneous pairwise comparison of their parameters, and the Johnson–Neyman procedure. *British Journal of Mathematical and Statistical Psychology*, *40*, 80–93.

Wilcox, R. R. (1990). Comparing the means of two independent groups. *Biometrical Journal*, *32*, 771–780. doi:10.1002/bimj.4710320702

Wilcox, R. R. (1996). Confidence intervals for the slope of a regression line when the error term has nonconstant variance. *Computational Statistics & Data Analysis*, *22*, 89–09. doi:10.1016/0167-9473(95)00038-0

Received 2 March 2010; revised version received 5 December 2011

Appendix

An R function for comparing multiple regression slopes

The R function for HC4WOM and HC4WMCP (function: hc4wmc) is stored in the source codes library Rallfun-v14, which can be downloaded from http://www-rcf.usc.edu/⟨0:br⟩⟨/0:br⟩~rwilcox/

Below is an example of the R syntax for applying the method:

```
> source('Rallfun-v14')
> X<-read.table('location/X.txt')
> X
```

	x1	x2	x3	x4
[1,]	−0.388	0.045	0.497	0.478
[2,]	−0.127	−1.563	−1.458	1.065
[3,]	−0.360	−0.874	1.028	−0.760
	...	...	...	...
[18,]	0.136	−0.308	0.951	−2.132
[19,]	0.452	−2.032	1.895	−0.803
[20,]	−1.143	0.676	0.660	0.001

```
> Y<-read.table('location/Y.txt')
> Y
```

	y1	y2	y3	y4
[1,]	−1.622	0.014	0.323	−1.123
[2,]	−0.162	−1.815	−0.151	−2.448
[3,]	−0.691	0.009	−1.013	0.926
	...	...	...	...
[18,]	0.706	−2.222	−1.124	2.679
[19,]	3.577	−3.278	−1.724	0.075
[20,]	−3.647	1.963	−1.406	0.128

```
> hc4wmc(X, Y)
$omnibus.p.value
[1] 0.02671119
$CI
```

	Group	Group	Lower CI	Upper CI
[1,]	1	2	−1.5117466	4.410823
[2,]	1	3	0.4225536	6.245460
[3,]	1	4	0.5538609	7.545452
[4,]	2	3	−0.4883542	4.257291
[5,]	2	4	−0.4616321	5.661868
[6,]	3	4	−2.2979304	3.729230

The output $omnibus.p.value shows the *p*-value of the omnibus test (HC4WOM) when testing H_0: $\beta_{11} = \ldots = \beta_{14}$. The output $CI shows the confidence intervals for the pairwise comparisons (HC4WMCP).