3-7-2019

On the Conditional and Unconditional Type I Error Rates and Power of Tests in Linear Models with Heteroscedastic Errors

Patrick J. Rosopa
Clemson University, prosopa@clemson.edu

Alice M. Brawley
Gettysburg College

Theresa P. Atkinson
Allstate

Stephen A. Robertson
Clemson University

Recommended Citation

On the Conditional and Unconditional Type I Error Rates and Power of Tests in Linear Models with Heteroscedastic Errors

Cover Page Footnote

An earlier version of this paper was presented at the 79th Annual Meeting of the Psychometric Society in Madison, WI, and the second author was the recipient of the Psychometric Society's Graduate Student Travel Award based on this paper.

בס"ד

On the Conditional and Unconditional Type I Error Rates and Power of Tests in Linear Models with Heteroscedastic Errors

Patrick J. Rosopa
Clemson University
Clemson, SC

Alice M. Brawley
Gettysburg College
Gettysburg, PA

Theresa P. Atkinson
Allstate
Dallas, TX

Stephen A. Robertson
Clemson University
Clemson, SC

Preliminary tests for homoscedasticity may be unnecessary in general linear models. Based on Monte Carlo simulations, results suggest that when testing for differences between independent slopes, the unconditional use of weighted least squares regression and HC4 regression performed the best across a wide range of conditions.

Keywords: Heterogeneity of variance, heteroscedasticity, linear models, Behrens-Fisher problem

In the behavioral and social sciences, researchers and practitioners employ statistical analyses to test theories, accumulate knowledge, and improve practice. In education, psychology, sociology, and related fields, some of the most frequently used statistical procedures involve linear models (e.g., analysis of variance, linear regression; Stone-Romero, Weaver, & Glenar, 1995) and the F and t statistics. These statistical tests typically require that various assumptions must be satisfied, including homoscedasticity (Box, 1954; Fox, 2008; Glass, Peckham, & Sanders, 1972). In other words, in linear models, the error term is assumed to be homoscedastic (Rencher, 2000). Some research, however, suggests that the practice of checking the homoscedasticity assumption may be outdated and, in the case of the two independent sample t test, unnecessary (Sawilowsky, 2002; Zimmerman,

2004), particularly with the availability of general solutions (Long & Ervin, 2000). The present study expands on existing research by examining tests for slope differences and evaluating not only the resulting Type I error rates, but also statistical power of tests on slope differences.

Introduction

Below, we discuss testing for the equality of independent slopes and the homoscedasticity assumption. Unique aspects of the present study are also noted, including our examination of unconditional and conditional Type I error and power.

Testing for the Equality of Independent Slopes

Testing for slope differences between k independent samples is quite common in the behavioral and social sciences. In the present study, the slope for a continuous predictor (x) when predicting a continuous response (y) may differ across a categorical predictor (z) where z is sometimes labeled a moderator (Saunders, 1956; Shieh, 2009). For example, the association between job involvement (x) and organizational citizenship behaviors (y) has been found to differ across sexes (z), such that higher job involvement is associated with higher levels of organizational citizenship behaviors for females, but lower levels of organizational citizenship behaviors for males (Diefendorff, Brown, Kamin, & Lord, 2002). In a 30-year review of three premier applied psychology and management journals, there were 636 tests for the equality of independent slopes, demonstrating the "pervasive interest in moderators" (Aguinis, Beaty, Boik, & Pierce, 2005, p. 94) in these fields. Note that testing the equality of k independent slopes is also referred to as moderated multiple regression with a categorical moderator (Aguinis, 2004) or a test of interaction (Fox, 2008).

For two independent groups (i.e., $k = 2$) and treating z as a dummy variable (e.g., 1 = group 1; 0 = group 2), the model for the ith observation can be expressed as

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 z_i + \beta_3 \left(x_i \cdot z_i \right) + \varepsilon_i \tag{1}$$

where $i = 1, 2, \ldots, N$, N = total number of observations, and ε_i is a population error. Further, the model in equation (1) assumes that ε_i has an expected value of 0 and a constant variance of σ^2. Note that population parameters are denoted by Greek

symbols (e.g., β_3) and estimates of these parameters that appear with a diacritic are typically obtained using ordinary least squares (OLS) estimation.

It deserves noting that the model in equation (1) does not require that the population errors follow a normal distribution (Rencher, 2000). However, for statistical inferences – such as hypothesis tests on the equality of independent slopes – the normality assumption is required. To test for the equality of two independent slopes, the test statistic, under the null hypothesis (i.e., H_0: $\beta_3 = 0$), is distributed as a t random variable with degrees of freedom $(df) = N - 4$ or, equivalently, as an F random variable with $df_1 = 1$ and $df_2 = N - 4$.

Between-Groups Heteroscedasticity

When conducting hypothesis tests on the equality of independent slopes, researchers and practitioners typically first conduct diagnostic tests to assess whether the homoscedasticity assumption is satisfied (Aguinis, 2004). In the present study, we focus on the scenario where the population error variance may differ systematically across two independent groups. That is, instead of the population errors (i.e., ε_i) having a constant variance of σ^2 (i.e., homoscedasticity), the error variance differs between groups (hereinafter, referred to as between-groups heteroscedasticity). In the jth group, where $j = 1$ or 2, the population error variance $\left(\sigma_j^2\right)$ can be expressed as

$$\sigma_j^2 = \sigma_{y(j)}^2 \left(1 - \rho_{yx(j)}^2\right) \tag{2}$$

where $\sigma_{y(j)}^2 = $ population variance of y in the jth group and $\rho_{yx(j)} = $ population correlation coefficient between y and x in the jth group.

To diagnose whether between-groups heteroscedasticity exists, traditionally, diagnostic procedures such as Levene's (1960) or Bartlett's test (Bartlett & Fowler, 1937) would precede Student's t test for two independent slopes. If the diagnostic test is not statistically significant (suggesting homoscedasticity), then a researcher would proceed with the interpretation of the conventional Student's t (or F statistic) to test the equality of independent slopes. However, if the diagnostic test is statistically significant, then this would signal that the homoscedasticity assumption was violated (i.e., heteroscedasticity exists) and that some ameliorative procedure should be applied (e.g., transformation of the dependent variable, weighted least squares regression; Fox, 2008).

Two diagnostic tests for homoscedasticity are examined in the present study. The first is Levene's test for equality of variances (Levene, 1960), which is simply a one-way analysis of variance (ANOVA) on the absolute value of the residuals. Not only is Levene's test generally well-known and computationally simple, it is available in all major statistical software packages (Rosopa, Schaffer, & Schroeder, 2013). The second diagnostic test examined in the present study is the score test (Breusch & Pagan, 1979; Cook & Weisberg, 1982), which is calculated using a two-step process that can be conducted using a number of major statistical software packages (Rosopa et al., 2013). In the first step, the model of interest is fitted using OLS regression. Then, the squared residuals from the first model are regressed on variables believed to be causing the heteroscedasticity (e.g., predictors or fitted values) to test whether the residuals are related to the focal variables. Under the null hypothesis, the test statistic is asymptotically distributed as χ^2 with *df* equal to the number of predictor variables used in the second step. The score test provides a flexible alternative to Levene's test because it can be used in more complex analyses, including testing interactions and polynomial regression (Rosopa et al., 2013). It is important to note, however, that like many tests for homoscedasticity, both the score test and Levene's test are sensitive to nonnormality (Rosopa et al., 2013).

We end this introduction with two important notes about our study design. First, some previous studies (e.g., Markowski & Markowski, 1990; Moser, Stevens, & Watts, 1989) have used Hartley's (1950) *F* as the preliminary test for equality of variances before testing for differences between independent means. However, research by Box (1953) indicates that Hartley's *F* should not be used as a preliminary test for the equality of variances; thus, we did not examine it in the present study. For a discussion and evaluation of other available diagnostic tests for homoscedasticity, we refer the reader to Ng and Wilcox (2011), Rosopa et al. (2013), and Sharma and Kibria (2013).

Second, it deserves noting that, with between-groups heteroscedasticity [see equation (2)], the test for the equality of two independent slopes can also be conducted using another approach that independently estimates two models. Namely, for each group, an OLS regression is conducted, regressing *y* on *x*. By allowing for heterogeneous variances across groups, a *t* test can be calculated using the two estimated slopes and their respective estimated standard errors. However, Satterthwaite's (1946) approximation for the *df* would be needed when using this approach. We thank an anonymous reviewer for bringing this to our attention. Because this analytic method requires conducting *k* separate regression analyses (i.e., one for each group), and the Satterthwaite approximation can be

computationally intensive, we chose to focus on the model depicted in equation (1), particularly, because of its frequent use by researchers and practitioners.

Research on Preliminary Tests for Heteroscedasticity

Although preliminary tests for heteroscedasticity are consistent with recommendations typically described in statistical textbooks, Zimmerman (2004) argued that "[n]othing is gained by the preliminary Levene test... The same line of reasoning applies to any preliminary test for the equality of variances that might be advised" (p. 179). Based on simulation results for tests of mean differences, Zimmerman concluded that the process of selecting test statistics based on preliminary tests for heteroscedasticity could not improve on the Type I error rate of unconditionally used (i.e., without a preliminary test for homoscedasticity) general procedures such as Welch's (1938) *t* test. That is, Zimmerman concluded that there is no need to conduct a preliminary test to assess whether the homoscedasticity assumption is violated when testing for independent mean differences, and Welch's *t* test should supplant Student's *t* test as a general procedure.

As demonstrated by Zimmerman (2004), the unconditional use of Welch's *t* – that is, using the test without conducting preliminary tests of equality of variances – maintains accurate Type I error rates when testing mean differences even when population variances and sample sizes are unequal across groups. Additional research further supports this conclusion with data generated from both normal and skewed distributions (Hayes & Cai, 2007) and when examining estimates of statistical power (Rasch, Kubinger, & Moder, 2011).

Some research involving independent means appears to generalize to instances involving slopes. For example, using preliminary tests of homoscedasticity in testing for differences between non-independent OLS regression estimates results in poorly controlled Type I error rates (Ng & Wilcox, 2011). However, it remains unclear whether this conclusion generalizes to tests on the equality of independent slopes, which as noted above is frequently conducted in the behavioral and social sciences (Aguinis et al., 2005).

Alternatives to OLS Regression when Homoscedasticity Assumption is Violated

As noted above, OLS regression is typically used for the estimation of parameters in equation (1). Although OLS regression-based parameter estimates remain unbiased in the presence of heteroscedasticity, standard errors will be incorrect

(Long & Ervin, 2000). Because standard errors are used in statistical inference, including hypothesis testing and interval estimation, such inferences will also be incorrect. Two alternatives to OLS regression are described below.

Weighted least squares (WLS) regression can be used instead of OLS regression to mitigate the effects of heteroscedasticity (Fox, 2008; Rosopa et al., 2013). With WLS regression, each estimated weight is equal to the reciprocal of a variance estimate (typically, using the OLS-based residuals). Note that OLS regression is a special case of WLS regression where the weights in OLS regression can be viewed as being equal to unity (Neter, Kutner, Nachtsheim, & Wasserman, 1996; Rosopa et al., 2013). However, in practice, the form of heteroscedasticity may be unknown, making the assignment of weights used in WLS regression an impractical procedure (Long & Ervin, 2000; Rosopa et al., 2013). That is, if the form of heteroscedasticity is not accurately identified, the WLS approach may not perform well relative to other alternative procedures. Note that nonparametric approaches for estimating the WLS weights are also available, such as nonparametric smoothing (Hart, 1997) and tree-based approaches (Su, Tsai, & Yan, 2006).

One general remedy for accurate statistical inference, even when the form of heteroscedasticity is unknown, is the use of heteroscedasticity consistent covariance matrices (HCCM). The first variation of HCCM, HC0 (White, 1980), is the most commonly used form, but it has been found to perform poorly with sample sizes less than 250 (Long & Ervin, 2000). Later variations of HCCMs – HC1, HC2, and HC3 – were developed to address issues encountered with small sample sizes (MacKinnon & White, 1985). In particular, HC3 has been shown to perform well when testing for differences between independent slopes under violations of homoscedasticity (Hayes & Agler, 2014).

A more recent variation of HCCM, HC4 (Cribari-Neto, 2004), is especially effective in the presence of outliers, and is calculated as follows:

$$\text{HC4} = \left(\mathbf{X}'\mathbf{X}\right)^{-1} \text{diag}\left[\frac{e_i^2}{\left(1-h_{ii}\right)^{\delta_i}}\right]\mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1} \tag{3}$$

where $\mathbf{X}$ is an $N \times (p + 1)$ model matrix of N observations with p predictors that also includes a leading column vector of 1s, e_i is the ith residual (an estimate of the ith population error, ε_i), h_{ii} is the ith leverage, and $\delta_i = \min\{4, Nh_{ii} / (p + 1)\}$. HC4 and other HCCMs are used for calculating standard errors. Thus, for the test statistic, the estimated slopes remain the same (i.e., OLS-based estimates). However, the

standard errors [i.e., square root of the diag(HC4)] will differ from those found using OLS estimation. Based on simulation research, HC4 performs the best compared to existing HCCMs in terms of control over Type I error rates and statistical power (Ng & Wilcox, 2009, 2011). Thus, HCCMs like HC4 might be expected to perform well across a broad range of heteroscedastic models relative to other alternative procedures because the form of heteroscedasticity need not be known when using HC4. Hereinafter, HC4 regression refers to an OLS regression where the covariance matrix among the estimated regression coefficients is based on equation (3).

The Present Study

To build on extant research, the present study evaluated both Type I error (i.e., size) and power of three tests on the difference between two independent slopes – OLS regression, WLS regression, and HC4 regression – using two preliminary tests for homoscedasticity – Levene's test and the score test. Previous research has thus far established that preliminary tests for homoscedasticity fail to improve Type I error rates for tests of mean differences (Hayes & Cai, 2007; Rasch et al., 2011; Zimmerman, 2004) and non-independent slope differences (Ng & Wilcox, 2011). This study uniquely examined whether commonly used tests for the equality of independent slopes (cf. Aguinis et al., 2005) are affected by the conditional use of statistical tests, and further examined the conditional statistical power of tests for slope differences.

Additionally, in the present study, the two general procedures (WLS regression and HC4 regression) were conducted both conditionally – i.e., the choice to use a general procedure instead of conventional OLS regression was based on the results of each of the diagnostic tests for homoscedasticity – and unconditionally – i.e., the general procedure was used in all cases, without any diagnostic test for homoscedasticity. Conducting particular statistical tests conditional on the results of a preliminary test of homoscedasticity is consistent with data analysis recommendations typically found in statistics textbooks (Fox, 2008; King, Rosopa, & Minium, 2010), and likely mirrors the data-analytic decision-making process in practice. Furthermore, while previous studies focused on conditional and unconditional Type I error rates (Hayes & Cai, 2007; Ng & Wilcox, 2011; Rasch et al., 2011; Zimmerman, 2004), given the importance of statistical power in designing research studies (Cohen, 1988; Liu, 2014; Shadish, Cook, & Campbell, 2002), the present study also examined the empirical power of these tests (for an exception, see Rasch et al.'s (2011) study of tests on mean

differences). Thus, our simulation estimated conditional and unconditional Type I error rates as well as conditional and unconditional power for tests of independent slope differences.

Methodology

A Monte Carlo simulation was conducted in R (R Core Team, 2012) involving two independent groups to evaluate the performance of several inferential tests of slope differences under various conditions including heteroscedasticity. A 6 (N) × 5 (n_1:n_2) × 8 (heteroscedasticity) × 2 (pairing type) × 7 (effect size) research design was used, resulting in 3,360 conditions. The nominal Type I error rate for all tests (including the tests for homoscedasticity) was .05.

Below, the manipulated variables in our simulation are described. To maximize the utility of our simulation, we selected levels of the manipulated variables that we felt mimicked prototypical conditions that might be encountered in the behavioral and social sciences.

Manipulated Variables

Total Sample Size

Six levels of N (i.e., total sample size) were used in the present study: 30, 60, 120, 180, 240, and 300. These values have been used in previous research (DeShon & Alexander, 1996) and bracket typical Ns encountered in the behavioral (Aguinis et al., 2005; Butler, Chapman, Forman, & Beck, 2006; Shen et al., 2011) and social sciences (Wallander, 2009).

Subgroup Sample Size

Unequal subgroup sample sizes are not uncommon. For example, attrition can result in unbalanced groups (Shadish et al., 2002). In test validation, there may exist unequal subgroups across a focal characteristic of interest (e.g., gender or race; see Hattrup & Schmitt, 1990; Hunter, Schmidt, & Hunter, 1979). Thus, the size of groups within samples was manipulated to include five ratios of n_1:n_2: (a) 1:1, (b) 1:2, (c) 1:3, (d) 1:4, and (e) 1:5. For example, when $N = 120$, the two independent subgroup sample sizes (based on the five ratios) were (a) $n_1 = n_2 = 60$, (b) $n_1 = 40$ and $n_2 = 80$, (c) $n_1 = 30$ and $n_2 = 90$, (d) $n_1 = 24$ and $n_2 = 96$, and (e) $n_1 = 20$ and $n_2 = 100$.

Between-Groups Heteroscedasticity

Between-groups heteroscedasticity assumed eight levels, which was defined as the ratio of the population error variance in group 2 to group 1. The ratios of population error variances were (a) 1:1, (b) 1.25:1, (c) 1.5:1, (d) 1.75:1, (e) 2:1, (f) 2.5:1, (g) 3:1, and (h) 4:1. The 1:1 ratio represents homoscedasticity, as the error variance between the two independent groups was equal.

Type of Pairing

When error variance and subgroup sample sizes are indirectly paired (i.e., largest σ_j^2 paired with the smallest n_j), tests for slope differences can show erroneous Type I error rates or reduced statistical power (DeShon & Alexander, 1996; Ng & Wilcox, 2010; Overton, 2001; Shieh, 2009). As such, we manipulated indirect and direct pairing to thoroughly assess the Type I error and power of these tests. We manipulated the combinations of subgroup sample sizes and error variances to include both indirect pairing and direct pairing (i.e., largest σ_j^2 paired with the largest n_j; see DeShon & Alexander, 1996; Overton, 2001; Zimmerman, 2004).

Effect Size

Effect size was also manipulated in this study. We used the modified effect size f^2 by Aguinis et al. (2005). Based on the effect size formulae, the Solver function in Microsoft Excel was used to solve for the slope of the second independent group that corresponded to a given effect size (i.e., f^2). The slope of one group (viz., group 1) remained fixed (see also Data Generation below). Although Cohen (1977) established $f^2 = .02$ as a small effect, Aguinis et al.'s (2005) 30-year review of research involving slope differences with a categorical moderator in applied psychology identified a median effect size of .002. Therefore, in addition to bracketing Cohen's guideline for a small effect, effect size was manipulated to bracket this median value and included seven levels, with $f^2 = 0, .001, .002, .005, .01, .02,$ and .03.

Data Generation

For x and y, n_1 pairs of random numbers were generated from a bivariate normal distribution where (a) the population means were equal to 0, (b) the population variance of x was equal to 1.5, (c) the population slope for group 1 (β_1) was equal to 0.5, and (d) the population error variance for group 1 was equal to 1. With these values fixed, the population correlation between x and y in group 1, and the

10

population variance of y in group 1 were determined. Similarly, for group 2, n_2 pairs of random numbers were generated from a bivariate normal distribution where (a) the population means were equal to 0, (b) the population variance of x was equal to 1.5, and (c) the population error variance equaled one of the eight values ranging from 1 (i.e., homoscedastic) to 4. Using the Solver function in Microsoft Excel, we calculated the population slope for group 2 that corresponded to one of the manipulated effect sizes. With these values fixed, the population correlation between x and y in group 2, and the population variance of y in group 2 were determined.

After the data were generated for each condition, seven procedures for testing for differences between independent slopes were conducted. All tests – OLS, WLS, and HC4 regression – were first conducted unconditionally; that is, these three procedures were each conducted without considering preliminary tests for homoscedasticity. Four additional procedures were conducted conditionally; that is, the use of the two general solutions – WLS regression and HC4 regression – was decided based on results of each of the two preliminary tests for homoscedasticity. For example, if Levene's test was nonsignificant – suggesting homoscedasticity – then the conventional procedure, OLS regression, was used. If Levene's test was statistically significant – suggesting heteroscedasticity – then a general procedure was used. All possible combinations of preliminary tests (i.e., Levene's and the score test) and the general tests (i.e., WLS regression and HC4 regression) were conducted in this manner, resulting in four additional procedures: (a) WLS regression conditional on Levene's test, (b) WLS regression conditional on the score test, (c) HC4 regression conditional on Levene's test, and (d) HC4 regression conditional on the score test.

Number of Replications

For each condition, 5,000 replications were conducted, and the number of rejections of the null hypothesis out of the 5,000 replications was recorded. For conditions where the true effect size (f^2) was zero, the proportion of times out of the 5,000 replications that the null hypothesis was rejected provided an estimate of Type I error; for conditions where f^2 was not zero, this proportion provided an estimate of statistical power. Note that in addition to multiple checks of the accuracy of our data generation method, our results for unconditional tests involving OLS regression were found to be comparable to similar conditions examined in DeShon and Alexander's (1996) study.

11

Results

The number of times out of the 5,000 replications that each test rejected the null hypothesis was recorded for all 3,360 conditions, resulting in estimates of Type I error and statistical power. Due to space limitations, a representative subset of the results is presented here. The complete set of results and R code can be obtained from the first and second authors.

Type I Error Rate

Empirical Type I error rates for tests of slope differences are reported in Tables 1 and 2. As shown in Table 1, when $N = 60$, 120, and 240 and when subgroup sample sizes were equal, all seven procedures resulted in empirical rejection rates within Serlin's (2000) criterion for robustness.

As shown in Table 2 ($N = 120$), when subgroup sample sizes were unequal, the Type I error rate for all tests except for unconditional OLS regression fell within the acceptable range of Type I error rates. For example, in Table 2, when heteroscedasticity was directly paired, the empirical Type I error rates for unconditional OLS regression ranged between .0032 and .0224, much less than the nominal Type I error rate. Moreover, in the same table, when heteroscedasticity was indirectly paired, the empirical Type I error rates for unconditional OLS regression ranged between .0766 and .1944, much greater than the nominal Type I error rate. Thus, when heteroscedasticity was directly paired or indirectly paired, unconditional OLS regression showed conservative or inflated Type I error rates, respectively, which became increasingly conservative or inflated as the $n_1:n_2$ ratio became more disproportionate. Compared to the unconditional use of WLS regression and HC4 regression, the conditional use of these procedures (i.e., when choice of a test was contingent on the satisfaction of the homoscedasticity assumption) did not show improved control over Type I error rates in any condition. That is, both unconditional WLS regression and unconditional HC4 regression were generally very stable in controlling Type I error rates.

Statistical Power

Empirical statistical power estimates for tests of slope differences are reported in Tables 3 and 4. As shown in Table 3, when $N = 60$, 120, and 240 and when subgroup sample sizes were equal, all tests showed comparable statistical power within conditions and as N increased, the power of all tests increased.

Table 1. Type I error when testing for slope differences with equal subgroup sample sizes (i.e., $n_1 = n_2$)

Heteroscedasticity	N	Unconditional OLS	Unconditional WLS	Unconditional HC4	WLS conditional on score test	WLS conditional on Levene's test	HC4 conditional on score test	HC4 conditional on Levene's test
Small (2:1)	60	0.0508	0.0498	0.0478	0.0502	0.0502	0.0518	0.0502
	120	0.0512	0.0510	0.0514	0.0514	0.0510	0.0528	0.0522
	240	0.0550	0.0550	0.0508	0.0550	0.0552	0.0512	0.0518
Large (4:1)	60	0.0500	0.0512	0.0494	0.0516	0.0514	0.0504	0.0494
	120	0.0534	0.0524	0.0478	0.0524	0.0524	0.0478	0.0478
	240	0.0476	0.0492	0.0488	0.0492	0.0492	0.0488	0.0488

Note: N = total sample size

Table 2. Type I error when testing for slope differences with unequal subgroup sample sizes (i.e., $n_1 \neq n_2$)

Heteroscedasticity	Pairing	$n_1:n_2$ ratio	Unconditional OLS	Unconditional WLS	Unconditional HC4	WLS cond. on score	WLS cond. on Levene	HC4 cond. on score	HC4 cond. on Levene
Small (2:1)	Direct	1:2	0.0224	0.0482	0.0448	0.0430	0.0402	0.0408	0.0386
		1:3	0.0222	0.0546	0.0488	0.0470	0.0452	0.0430	0.0420
		1:4	0.0192	0.0562	0.0516	0.0440	0.0452	0.0416	0.0422
		1:5	0.0146	0.0614	0.0564	0.0440	0.0478	0.0396	0.0436
	Indirect	1:2	0.0766	0.0510	0.0510	0.0538	0.0574	0.0550	0.0592
		1:3	0.1002	0.0538	0.0538	0.0654	0.0702	0.0660	0.0704
		1:4	0.1166	0.0606	0.0546	0.0750	0.0830	0.0736	0.0812
		1:5	0.1210	0.0604	0.0590	0.0808	0.0894	0.0804	0.0876
Large (4:1)	Direct	1:2	0.0160	0.0490	0.0472	0.0490	0.0490	0.0472	0.0472
		1:3	0.0072	0.0484	0.0464	0.0484	0.0484	0.0464	0.0464
		1:4	0.0048	0.0522	0.0440	0.0520	0.0514	0.0438	0.0434
		1:5	0.0032	0.0578	0.0500	0.0562	0.0562	0.0486	0.0484

Note: N = 120; n_1 = sample size in group 1, n_2 = sample size in group 2

Table 2 (continued).

Heteroscedasticity	Pairing	n_1:n_2 ratio	Unconditional OLS	Unconditional WLS	Unconditional HC4	WLS cond. on score	WLS cond. on Levene	HC4 cond. on score	HC4 cond. on Levene
Large (4:1)	Indirect	1:2	0.1094	0.0544	0.0518	0.0544	0.0544	0.0518	0.0518
		1:3	0.1536	0.0546	0.0586	0.0550	0.0564	0.0590	0.0598
		1:4	0.1800	0.0568	0.0536	0.0580	0.0604	0.0546	0.0576
		1:5	0.1944	0.0656	0.0586	0.0684	0.0712	0.0622	0.0660

Note: $N = 120$; n_1 = sample size in group 1, n_2 = sample size in group 2

Table 3. Power when testing for slope differences with equal subgroup sample sizes (i.e., $n_1 = n_2$)

N	Heteroscedasticity	f^2	Unconditional OLS	Unconditional WLS	Unconditional HC4	WLS cond. on score	WLS cond. on Levene	HC4 cond. on score	HC4 cond. on Levene
60	Small (2:1)	0.001	0.0682	0.0672	0.0616	0.0674	0.0676	0.0680	0.0682
		0.002	0.0742	0.0748	0.0640	0.0750	0.0746	0.0710	0.0716
		0.005	0.1170	0.1168	0.1058	0.1158	0.1158	0.1102	0.1110
		0.010	0.1952	0.1966	0.1734	0.1970	0.1966	0.1856	0.1876
		0.020	0.3328	0.3356	0.2952	0.3340	0.3342	0.3180	0.3212
		0.030	0.4582	0.4570	0.4128	0.4576	0.4570	0.4370	0.4408
	Large (4:1)	0.001	0.0672	0.0688	0.0648	0.0688	0.0690	0.0652	0.0656
		0.002	0.0740	0.0756	0.0712	0.0758	0.0752	0.0708	0.0702
		0.005	0.1274	0.1248	0.1176	0.1252	0.1250	0.1184	0.1176
		0.010	0.1944	0.1974	0.1790	0.1976	0.1978	0.1808	0.1826
		0.020	0.3350	0.3398	0.3074	0.3404	0.3406	0.3092	0.3096
		0.030	0.4668	0.4676	0.4288	0.4678	0.4670	0.4318	0.4344

Note: N = total sample size; f^2 = effect size

Table 3 (continued).

N	Heteroscedasticity	f^2	Unconditional OLS	Unconditional WLS	Unconditional HC4	WLS cond. on score	WLS cond. on Levene	HC4 cond. on score	HC4 cond. on Levene
120	Small (2:1)	0.001	0.0858	0.0856	0.0808	0.0856	0.0852	0.0812	0.0822
		0.002	0.1050	0.1046	0.0970	0.1048	0.1042	0.1002	0.0994
		0.005	0.2010	0.2008	0.1916	0.2008	0.2010	0.1922	0.1948
		0.010	0.3524	0.3516	0.3350	0.3524	0.3526	0.3386	0.3386
		0.020	0.6170	0.6194	0.5928	0.6186	0.6186	0.5984	0.5998
		0.030	0.7798	0.7812	0.7580	0.7814	0.7816	0.7614	0.7624
	Large (4:1)	0.001	0.0776	0.0804	0.0774	0.0804	0.0804	0.0774	0.0774
		0.002	0.1110	0.1096	0.1024	0.1096	0.1096	0.1024	0.1024
		0.005	0.2066	0.2102	0.1982	0.2102	0.2102	0.1982	0.1982
		0.010	0.3478	0.3504	0.3332	0.3504	0.3504	0.3332	0.3332
		0.020	0.6110	0.6116	0.5940	0.6116	0.6116	0.5940	0.5940
		0.030	0.7812	0.7830	0.7628	0.7830	0.7832	0.7628	0.7628
240	Small (2:1)	0.001	0.0858	0.0856	0.0808	0.0856	0.0852	0.0812	0.0822
		0.002	0.1050	0.1046	0.0970	0.1048	0.1042	0.1002	0.0994
		0.005	0.2010	0.2008	0.1916	0.2008	0.2010	0.1922	0.1948
		0.010	0.3524	0.3516	0.3350	0.3524	0.3526	0.3386	0.3386
		0.020	0.6170	0.6194	0.5928	0.6186	0.6186	0.5984	0.5998
		0.030	0.7798	0.7812	0.7580	0.7814	0.7816	0.7614	0.7624
	Large (4:1)	0.001	0.0776	0.0804	0.0774	0.0804	0.0804	0.0774	0.0774
		0.002	0.1110	0.1096	0.1024	0.1096	0.1096	0.1024	0.1024
		0.005	0.2066	0.2102	0.1982	0.2102	0.2102	0.1982	0.1982
		0.010	0.3478	0.3504	0.3332	0.3504	0.3504	0.3332	0.3332
		0.020	0.6110	0.6116	0.5940	0.6116	0.6116	0.5940	0.5940
		0.030	0.7812	0.7830	0.7628	0.7830	0.7832	0.7628	0.7628

Note: N = total sample size; f^2 = effect size

Table 4. Power when testing for slope differences with unequal subgroup sample sizes (i.e., $n_1 \neq n_2$)

Heteroscedasticity	Pairing	$n_1:n_2$ ratio	f^2	Uncond. OLS	Uncond. WLS	Uncond. HC4	WLS cond. on score	WLS cond. on Levene	HC4 cond. on score	HC4 cond. on Levene
Small (2:1)	Direct	1:2	0.001	0.0508	0.0870	0.0848	0.0794	0.0788	0.0784	0.0770
		1:2	0.002	0.0848	0.1308	0.1182	0.1206	0.1186	0.1126	0.1126
		1:2	0.005	0.1772	0.2476	0.2246	0.2346	0.2294	0.2210	0.2188
		1:2	0.010	0.3378	0.4272	0.4040	0.4106	0.4054	0.3960	0.3922
		1:2	0.020	0.6170	0.7054	0.6790	0.6888	0.6840	0.6706	0.6672
		1:2	0.030	0.7986	0.8614	0.8396	0.8486	0.8456	0.8356	0.8340
	Indirect	1:2	0.001	0.1074	0.0742	0.0714	0.0800	0.0824	0.0784	0.0794
		1:2	0.002	0.1344	0.0908	0.0832	0.0984	0.1030	0.0932	0.0992
		1:2	0.005	0.2358	0.1802	0.1670	0.1884	0.1948	0.1806	0.1858
		1:2	0.010	0.3794	0.2972	0.2842	0.3086	0.3200	0.3014	0.3108
		1:2	0.020	0.5994	0.5180	0.4942	0.5308	0.5370	0.5126	0.5218
		1:2	0.030	0.7544	0.6844	0.6476	0.6954	0.7014	0.6684	0.6794
	Direct	1:3	0.001	0.0394	0.0856	0.0798	0.0730	0.0738	0.0704	0.0702
		1:3	0.002	0.0598	0.1274	0.1166	0.1092	0.1084	0.1022	0.1016
		1:3	0.005	0.1700	0.2678	0.2488	0.2426	0.2386	0.2308	0.2282
		1:3	0.010	0.3220	0.4664	0.4308	0.4228	0.4214	0.4050	0.4054
		1:3	0.020	0.6200	0.7454	0.7106	0.7110	0.7088	0.6908	0.6892
		1:3	0.030	0.8006	0.8884	0.8646	0.8612	0.8620	0.8500	0.8510
	Indirect	1:3	0.001	0.1224	0.0744	0.0680	0.0854	0.0922	0.0810	0.0882
		1:3	0.002	0.1632	0.0986	0.0936	0.1110	0.1188	0.1098	0.1176
		1:3	0.005	0.2442	0.1612	0.1508	0.1768	0.1902	0.1738	0.1874
		1:3	0.010	0.3824	0.2848	0.2574	0.3000	0.3106	0.2848	0.2992
		1:3	0.020	0.5876	0.4750	0.4406	0.4966	0.5088	0.4744	0.4902
		1:3	0.030	0.7266	0.6316	0.5842	0.6500	0.6624	0.6190	0.6344

Note: $N = 120$; n_1 = sample size in group 1, n_2 = sample size in group 2, f^2 = effect size

Table 4 (continued).

Heteroscedasticity	Pairing	$n_1 : n_2$ ratio	f^2	Uncond. OLS	Uncond. WLS	Uncond. HC4	WLS cond. on score	WLS cond. on Levene	HC4 cond. on score	HC4 cond. on Levene
Small (2:1)	Direct	1:4	0.001	0.0342	0.0960	0.0864	0.0738	0.0740	0.0690	0.0696
		1:4	0.002	0.0646	0.1558	0.1384	0.1208	0.1212	0.1102	0.1108
		1:4	0.005	0.1502	0.2864	0.2564	0.2296	0.2348	0.2166	0.2216
		1:4	0.010	0.3216	0.4920	0.4438	0.4210	0.4196	0.3996	0.4036
		1:4	0.020	0.6190	0.7760	0.7274	0.7104	0.7162	0.6912	0.6976
		1:4	0.030	0.8112	0.9074	0.8720	0.8690	0.8722	0.8560	0.8570
	Indirect	1:4	0.001	0.1390	0.0792	0.0738	0.0922	0.1014	0.0908	0.0998
		1:4	0.002	0.1754	0.1074	0.1004	0.1238	0.1330	0.1222	0.1302
		1:4	0.005	0.2590	0.1670	0.1484	0.1880	0.2010	0.1790	0.1932
		1:4	0.010	0.3830	0.2692	0.2416	0.2944	0.3122	0.2818	0.2986
		1:4	0.020	0.5932	0.4714	0.4230	0.4928	0.5078	0.4674	0.4868
		1:4	0.030	0.7270	0.5990	0.5406	0.6220	0.6414	0.5892	0.6106
	Direct	1:5	0.001	0.0386	0.1112	0.0992	0.0800	0.0844	0.0740	0.0796
		1:5	0.002	0.0558	0.1512	0.1324	0.1096	0.1148	0.1008	0.1062
		1:5	0.005	0.1408	0.2996	0.2654	0.2246	0.2350	0.2080	0.2170
		1:5	0.010	0.3152	0.5176	0.4612	0.4184	0.4314	0.4032	0.4116
		1:5	0.020	0.6126	0.7938	0.7300	0.7074	0.7194	0.6902	0.6960
		1:5	0.030	0.7988	0.9092	0.8678	0.8558	0.8626	0.8460	0.8528
	Indirect	1:5	0.001	0.1510	0.0848	0.0764	0.1048	0.1160	0.1026	0.1136
		1:5	0.002	0.1808	0.1058	0.1010	0.1300	0.1418	0.1310	0.1404
		1:5	0.005	0.2594	0.1640	0.1412	0.1898	0.2060	0.1834	0.2010
		1:5	0.010	0.3854	0.2644	0.2302	0.2966	0.3152	0.2824	0.3030
		1:5	0.020	0.5794	0.4440	0.3888	0.4780	0.4996	0.4474	0.4724
		1:5	0.030	0.7328	0.6086	0.5428	0.6336	0.6538	0.5986	0.6244

Note: $N = 120$; n_1 = sample size in group 1, n_2 = sample size in group 2, f^2 = effect size

Table 4 (continued).

Heteroscedasticity	Pairing	$n_1:n_2$ ratio	f^2	Uncond. OLS	Uncond. WLS	Uncond. HC4	WLS cond. on score	WLS cond. on Levene	HC4 cond. on score	HC4 cond. on Levene
Large (4:1)	Direct	1:2	0.001	0.0358	0.0888	0.0812	0.0888	0.0888	0.0812	0.0814
		1:2	0.002	0.0650	0.1404	0.1300	0.1404	0.1402	0.1298	0.1296
		1:2	0.005	0.1664	0.2796	0.2620	0.2796	0.2796	0.2620	0.2620
		1:2	0.010	0.3228	0.4980	0.4732	0.4976	0.4974	0.4728	0.4726
		1:2	0.020	0.6334	0.7894	0.7680	0.7894	0.7890	0.7680	0.7680
		1:2	0.030	0.8244	0.9226	0.9090	0.9224	0.9224	0.9088	0.9088
	Indirect	1:2	0.001	0.1428	0.0792	0.0724	0.0792	0.0794	0.0726	0.0728
		1:2	0.002	0.1664	0.0950	0.0904	0.0952	0.0954	0.0906	0.0908
		1:2	0.005	0.2522	0.1614	0.1506	0.1614	0.1616	0.1508	0.1510
		1:2	0.010	0.3932	0.2622	0.2430	0.2624	0.2622	0.2430	0.2436
		1:2	0.020	0.5850	0.4482	0.4164	0.4482	0.4482	0.4164	0.4166
		1:2	0.030	0.7422	0.6166	0.5834	0.6166	0.6172	0.5836	0.5838
	Direct	1:3	0.001	0.0218	0.1078	0.0994	0.1074	0.1068	0.0990	0.0986
		1:3	0.002	0.0436	0.1562	0.1430	0.1554	0.1552	0.1424	0.1426
		1:3	0.005	0.1216	0.3326	0.3062	0.3318	0.3304	0.3062	0.3046
		1:3	0.010	0.3022	0.5906	0.5556	0.5898	0.5888	0.5550	0.5544
		1:3	0.020	0.6376	0.8588	0.8320	0.8582	0.8576	0.8316	0.8312
		1:3	0.030	0.8344	0.9592	0.9464	0.9592	0.9588	0.9464	0.9466
	Indirect	1:3	0.001	0.1796	0.0766	0.0702	0.0770	0.0784	0.0708	0.0718
		1:3	0.002	0.1984	0.0856	0.0784	0.0862	0.0876	0.0792	0.0810
		1:3	0.005	0.2834	0.1302	0.1180	0.1304	0.1322	0.1188	0.1200
		1:3	0.010	0.3988	0.2266	0.2130	0.2274	0.2282	0.2140	0.2154
		1:3	0.020	0.5780	0.3660	0.3360	0.3664	0.3676	0.3358	0.3368
		1:3	0.030	0.7162	0.5314	0.4860	0.5320	0.5342	0.4864	0.4884

Note: $N = 120$; n_1 = sample size in group 1, n_2 = sample size in group 2, f^2 = effect size

Table 4 (continued).

Heteroscedasticity	Pairing	n_1:n_2 ratio	f^2	Uncond. OLS	Uncond. WLS	Uncond. HC4	WLS cond. on score	WLS cond. on Levene	HC4 cond. on score	HC4 cond. on Levene
Large (4:1)	Direct	1:4	0.001	0.0174	0.1200	0.1056	0.1186	0.1184	0.1050	0.1050
		1:4	0.002	0.0340	0.1762	0.1588	0.1744	0.1726	0.1570	0.1560
		1:4	0.005	0.1196	0.3830	0.3474	0.3806	0.3770	0.3456	0.3430
		1:4	0.010	0.2938	0.6322	0.5858	0.6286	0.6270	0.5828	0.5830
		1:4	0.020	0.6318	0.8940	0.8628	0.8916	0.8900	0.8612	0.8600
		1:4	0.030	0.8442	0.9734	0.9588	0.9716	0.9712	0.9582	0.9580
	Indirect	1:4	0.001	0.2104	0.0764	0.0790	0.0774	0.0798	0.0808	0.0820
		1:4	0.002	0.2316	0.0896	0.0786	0.0912	0.0934	0.0800	0.0828
		1:4	0.005	0.2988	0.1320	0.1208	0.1330	0.1366	0.1220	0.1256
		1:4	0.010	0.4180	0.2176	0.2012	0.2198	0.2232	0.2050	0.2086
		1:4	0.020	0.5804	0.3506	0.3122	0.3516	0.3548	0.3136	0.3170
		1:4	0.030	0.6948	0.4640	0.4098	0.4652	0.4676	0.4122	0.4152
	Direct	1:5	0.001	0.0142	0.1264	0.1136	0.1238	0.1236	0.1114	0.1120
		1:5	0.002	0.0276	0.1984	0.1698	0.1940	0.1934	0.1664	0.1666
		1:5	0.005	0.1024	0.4000	0.3590	0.3902	0.3886	0.3522	0.3514
		1:5	0.010	0.2836	0.6820	0.6246	0.6690	0.6698	0.6156	0.6152
		1:5	0.020	0.6346	0.9144	0.8796	0.9060	0.9056	0.8736	0.8732
		1:5	0.030	0.8368	0.9784	0.9650	0.9722	0.9726	0.9612	0.9616
	Indirect	1:5	0.001	0.2244	0.0754	0.0754	0.0784	0.0838	0.0788	0.0830
		1:5	0.002	0.2518	0.0906	0.0790	0.0934	0.0966	0.0828	0.0872
		1:5	0.005	0.3174	0.1368	0.1200	0.1394	0.1456	0.1232	0.1292
		1:5	0.010	0.4128	0.1962	0.1794	0.1982	0.2018	0.1816	0.1864
		1:5	0.020	0.5766	0.3330	0.2914	0.3368	0.3440	0.2966	0.3046
		1:5	0.030	0.6840	0.4506	0.3876	0.4534	0.4578	0.3926	0.3986

Note: $N = 120$; n_1 = sample size in group 1, n_2 = sample size in group 2, f^2 = effect size

19

Table 4 ($N = 120$) presents empirical power estimates when subgroup sample sizes were unequal. When heteroscedasticity was directly paired, WLS regression had the greatest statistical power. Although conditional WLS regression and conditional HC4 regression had greater statistical power than OLS regression, unconditional WLS regression provided the greatest statistical power.

When heteroscedasticity was indirectly paired, OLS regression showed the greatest statistical power. However, it is important to note that for these similar conditions the Type I error rate of OLS regression was very inflated (see Table 2). Thus, the increased power of OLS regression comes at the expense of very inflated Type I error rates in these conditions. Notably, the power of OLS regression in these conditions was only slightly higher than that of WLS regression and HC4 regression. However, recall that WLS regression and HC4 regression were better able to control Type I error rates at the nominal level (see comparable conditions in Table 2). In addition, conditional WLS regression and conditional HC4 regression did not improve statistical power in any condition when compared to unconditional WLS regression and unconditional HC4 regression.

Conclusion

Researchers have recommended that the procedure of preliminary tests for homoscedasticity be abandoned when testing mean differences and non-independent slope differences, in favor of more general solutions that are robust in the presence of heteroscedasticity (Sawilowsky, 2002; Zimmerman, 2004). We expanded and further supported this recommendation by investigating the impact of abandoning this assumption when testing for independent slope differences, and by examining effects of this recommendation on statistical power. By evaluating the conditional and unconditional Type I error rates and statistical power of particular tests of slope differences under various conditions, our results may provide guidance for researchers and practitioners.

Although power does increase with effect size and N, when subgroup sample sizes were equal, all tests performed equally well when testing for independent slope differences. When subgroup sample sizes were unequal, all tests except for the conventional procedure (i.e., OLS regression) performed equally well. Regardless of whether subgroup sample sizes are equal or not, with the conditional use of statistics based on results of tests for homoscedasticity, there were no incremental improvements in controlling Type I error and there were no incremental increases in power. However, regarding the two diagnostic tests for detecting heteroscedasticity (i.e., score test and Levene's test), although WLS

20

regression and HC4 regression did not necessarily perform better when using the score test vs. Levene's test, it deserves noting that the conditional WLS regression tended to have greater power when compared to the conditional HC4 regression. Overall, the results of our statistical simulation suggest that when testing the equality of independent slopes, researchers and practitioners should (unconditionally) use general statistical procedures such as WLS regression and HC4 regression.

The present study may support the use of conventional OLS regression under some exploratory conditions: our results indicate that conventional tests may result in increased statistical power when effect sizes are small and heteroscedasticity is indirectly paired. However, this increased power is available at the cost of inflated Type I error rates. On the other hand, researchers and practitioners may opt for the use of general procedures (e.g., HC4 regression) that – despite slightly lower statistical power (typically, in the hundredth or thousandth decimal place when sample sizes are relatively large) in the presence of indirectly paired heteroscedasticity – result in more accurate Type I error rates.

It is important to note that we do not argue that our simulation study has examined all conceivable conditions that might be encountered in practice. For example, our simulation study did not include manipulated effect sizes that would be considered large (Cohen, 1988). Because power curves asymptote as effect size increases, especially as N increases, power curves can overlap considerably and examining relative power at such large effect sizes may not provide much practical value. For example, when $f^2 = .03$, $N = 180$, $n_1 = n_2$, and the ratio of population error variances was 4:1, the power of the various procedures we examined ranged between .915 and .924. For this condition, had we included a manipulated $f^2 = .25$ (medium effect size according to Cohen, 1988), power would have equaled 1.0 for all procedures. Because our conditions were designed to represent and bracket circumstances typical in behavioral and social science research, we feel that our findings provide a framework relevant to a variety of conditions likely to mirror those encountered in practice (e.g., comparing two independent groups, small effect sizes and effects sizes near the median).

To facilitate the use of statistical procedures by researchers and practitioners, it can be useful if such procedures are readily accessible and are user friendly. Both WLS regression and HC4 regression procedures can be easily implemented in a number of statistical analysis programs that are commonly used in the behavioral and social sciences (Rosopa et al., 2013). More specifically, WLS regression can be implemented in SPSS, SAS, R, STATA, and SYSTEM. At the time of this writing, HC4 regression can be implemented in all of the same programs except for

SYSTAT and STATA. Additionally, in SPSS and SAS, the PROCESS macro (Hayes, 2013) can conduct tests for slope differences in moderated multiple regression using the HC4 estimator, and the RLM macro (Darlington & Hayes, 2017) can implement the HC4 for several types of regression analyses. Therefore, both WLS regression and HC4 regression are accessible to researchers using a variety of common statistical programs and packages.

In conclusion, the present study extends support for abandoning the process of conducting preliminary tests of homoscedasticity. Our results provide support for this paradigm shift when testing for independent slope differences, and with regard to not only Type I error rates but also statistical power. Consistent with recommendations for tests on mean differences by Sawilowsky (2002) and Zimmerman (2004), our results suggest that preliminary tests for homoscedasticity when testing for slope differences may not be necessary when using a general procedure (e.g., HC4 regression). Therefore, under most research and practice applications, we recommend the unconditional use of a general procedure (e.g., WLS regression or HC4 regression for slope differences) examined here. Results of this study highlight the importance of adequate research design execution and analysis, such as maintaining equal subgroup sample sizes when possible and understanding the type of heteroscedasticity pairing when selecting a statistic to test for slope differences.

References

Aguinis, H. (2004). *Regression analysis for categorical moderators*. New York, NY: Guilford.

Aguinis, H., Beaty, J. C., Boik, R. J., & Pierce, C. A. (2005). Effect size and power in assessing moderating effects of categorical variables using multiple regression: A 30-year review. *Journal of Applied Psychology, 90*(1), 94-107. doi: 10.1037/0021-9010.90.1.94

Bartlett, M. S., & Fowler, R. H. (1937). Properties of sufficiency and statistical tests. *Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences, 160*(901), 268-282. doi: 10.1098/rspa.1937.0109

Box, G. E. P. (1953). Non-normality and tests on variances. *Biometrika, 40*(3/4), 318-335. doi: 10.1093/biomet/40.3-4.318

Box, G. E. P. (1954). Some theorems on quadratic forms applied in the study of analysis of variance problems, I. Effect of inequality of variance in the

one-way classification. *Annals of Mathematical Statistics, 25*(2), 290-302. doi: 10.1214/aoms/1177728786

Breusch, T. S., & Pagan, A. R. (1979). A simple test for heteroscedasticity and random coefficient variation. *Econometrica, 47*(5), 1287-1294. doi: 10.2307/1911963

Butler, A., Chapman, J., Forman, E., & Beck, A. (2006). The empirical status of cognitive behavioral therapy: A review of meta-analyses. *Clinical Psychology Review, 26*(1), 17-31. doi: 10.1016/j.cpr.2005.07.003

Cohen, J. (1977). *Statistical power analysis for the behavioral sciences*. New York, NY: Academic Press. doi: 10.1016/C2013-0-10517-X

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum. doi: 10.4324/9780203771587

Cook, R. D., & Weisberg, S. (1982). *Residuals and influence in regression* (Vol. 5). New York, NY: Chapman and Hall. Retrieved from https://conservancy.umn.edu/handle/11299/37076

Cribari-Neto, F. (2004). Asymptotic inference under heteroskedasticity of unknown form. *Computational Statistics & Data Analysis, 45*(2), 215-233. doi: 10.1016/S0167-9473(02)00366-3

Darlington, R. B., & Hayes, A. F. (2017). *Regression analysis and linear models: Concepts, applications, and implementation* (2nd ed.). New York, NY: Guilford Press.

DeShon, R. P., & Alexander, R. A. (1996). Alternative procedures for testing regression slope homogeneity when group error variances are unequal. *Psychological Methods, 1*(3), 261. doi: 10.1037/1082-989X.1.3.261

Diefendorff, J. M., Brown, D. J., Kamin, A. M., & Lord, R. G. (2002). Examining the roles of job involvement and work centrality in predicting organizational citizenship behaviors and job performance. *Journal of Organizational Behavior, 23*(1), 93-108. doi: 10.1002/job.123

Fox, J. (2008). *Applied regression analysis and generalized linear models* (2nd ed.). Thousand Oaks, CA: Sage.

Glass, G. V., Peckham, P. D., & Sanders, J. R. (1972). Consequences of failure to meet assumptions underlying the fixed effects analyses of variance and covariance. *Review of Educational Research, 42*(3), 237-288. doi: 10.3102/00346543042003237

Hart, J. D. (1997). *Nonparametric smoothing and lack-of-fit tests*. New York, NY: Springer. doi: 10.1007/978-1-4757-2722-7

Hartley, H. O. (1950). The maximum *F*-ratio as a short-cut test for heterogeneity of variance. *Biometrika, 37*(3-4), 308-312. doi: 10.1093/biomet/37.3-4.308

Hattrup, K., & Schmitt, N. (1990). Prediction of trades apprentices' performance on job sample criteria. *Personnel Psychology, 43*(3), 453-466. doi: 10.1111/j.1744-6570.1990.tb02392.x

Hayes, A. F. (2013). *Introduction to mediation, moderation, and conditional process analysis: A regression-based approach*. New York, NY: Guilford Press.

Hayes, A. F., & Agler, R. A. (2014). On the standard error of the difference between independent regression coefficients in moderation analysis. *Multiple Linear Regression Viewpoints, 40*(2), 16-27.

Hayes, A. F., & Cai, L. (2007). Further evaluating the conditional decision rule for comparing two independent means. *British Journal of Mathematical and Statistical Psychology, 60*(2), 217-244. doi: 10.1348/000711005X62576

Hunter, J. E., Schmidt, F. L., & Hunter, R. (1979). Differential validity of employment tests by race: A comprehensive review and analysis. *Psychological Bulletin, 86*(), 721-735. doi: 10.1037/0033-2909.86.4.721

King, B. M., Rosopa, P. J., & Minium, E. W. (2010). *Statistical reasoning in the behavioral sciences* (6th ed.). Hoboken, NJ: Wiley.

Levene, H. (1960). Robust tests for equality of variances. In I. Olkin (Ed.), *Contributions to probability and statistics: Essays in honor of Harold Hotelling* (pp. 279-292). Stanford, CA: Stanford University Press.

Liu, X. S. (2014). *Statistical power analysis for the social and behavioral sciences: Basic and advanced techniques*. New York, NY: Routledge.

Long, J. S., & Ervin, L. H. (2000). Using heteroscedasticity consistent standard errors in the linear regression model. *The American Statistician, 54*(3), 217-224. doi: 10.1080/00031305.2000.10474549

MacKinnon, J. G., & White, H. (1985). Some heteroskedasticity-consistent covariance matrix estimators with improved finite sample properties. *Journal of Econometrics, 29*(3), 305-325. doi: 10.1016/0304-4076(85)90158-7

Markowski, C. A., & Markowski, E. P. (1990). Conditions for the effectiveness of a preliminary test of variance. *The American Statistician, 44*(4), 322-326. doi: 10.1080/00031305.1990.10475752

Moser, B. K., Stevens, G. R., & Watts, C. L. (1989). The two-sample *t* test versus Satterthwaite's approximate *F* test. *Communications in Statistics – Theory and Methods, 18*(11), 3963-3975. doi: 10.1080/03610928908830135

Neter, J., Kutner, M. H., Nachtsheim, C. J., & Wasserman, W. (1996). *Applied linear regression models* (3rd ed.). Chicago, IL: Irwin.

Ng, M., & Wilcox, R. R. (2009). Level robust methods based on the least squares regression estimator. *Journal of Modern Applied Statistical Methods, 8*(5), 384-395. doi: 10.22237/jmasm/1257033840

Ng, M., & Wilcox, R. R. (2010). Comparing the regression slopes of independent groups. *British Journal of Mathematical and Statistical Psychology, 63*(2), 319-340. doi: 10.1348/000711009X456845

Ng, M., & Wilcox, R. R. (2011). A comparison of two-stage procedures for testing least-squares coefficients under heteroscedasticity. *British Journal of Mathematical and Statistical Psychology, 64*(2), 244–258. doi: 10.1348/000711010X508683

Overton, R. C. (2001). Moderated multiple regression for interactions involving categorical variables: A statistical control for heterogeneous variance across two groups. *Psychological Methods, 6*(3), 218-233. doi: 10.1037/1082-989X.6.3.218

R Core Team. (2012). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Rasch, D., Kubinger, K. D., & Moder, K. (2011). The two-sample *t* test: Pre-testing its assumptions does not pay off. *Statistical Papers, 52*(1), 219-231. doi: 10.1007/s00362-009-0224-x

Rencher, A. C. (2000). *Linear models in statistics*. New York, NY: Wiley.

Rosopa, P. J., Schaffer, M. M., & Schroeder, A. N. (2013). Managing heteroscedasticity in general linear models. *Psychological Methods, 18*(3), 335-351. doi: 10.1037/a0032553

Satterthwaite, F. E. (1946). An approximate distribution of estimates of variance components. *Biometrics Bulletin, 2*(6), 110-114. doi: 10.2307/3002019

Saunders, D. R. (1956). Moderator variables in prediction. *Educational and Psychological Measurement, 16*(2), 209-222. doi: 10.1177/001316445601600205

Sawilowsky, S. S. (2002). Fermat, Schubert, Einstein, and Behrens-Fisher: The probable difference between two means when $\sigma_1^2 \neq \sigma_2^2$. *Journal of Modern Applied Statistical Methods, 1*(2), 461-472. doi: 10.22237/jmasm/1036109940

Serlin, R. C. (2000). Testing for robustness in Monte Carlo studies. *Psychological Methods, 5*(2), 230-240. doi: 10.1037/1082-989X.5.2.230

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston, MA: Houghton-Mifflin.

Sharma, D., & Kibria, B. M. G (2013). On some test statistics for testing homogeneity of variances: A comparative study. *Journal of Statistical Computation and Simulation, 83*(10), 1944-1963. doi: 10.1080/00949655.2012.675336

Shen, W., Kiger, T. B., Davies, S. E., Rasch, R. L., Simon, K. M., & Ones, D. S. (2011). Samples in applied psychology: Over a decade of research in review. *Journal of Applied Psychology, 96*(5), 1055-1064. doi: 10.1037/a0023322

Shieh, G. (2009). Detecting interaction effects in moderated multiple regression with continuous variables power and sample size considerations. *Organizational Research Methods, 12*(3), 510-528. doi: 10.1177/1094428108320370

Stone-Romero, E. F., Weaver, A. E., & Glenar, J. L. (1995). Trends in research design and data analytic strategies in organizational research. *Journal of Management, 21*(1), 141-157. doi: 10.1177/014920639502100109

Su, X., Tsai, C. L., & Yan, X. (2006). Treed variance. *Journal of Computational and Graphical Statistics, 15*(2), 356-371. doi: 10.1198/106186006X113575

Wallander, L. (2009). 25 years of factorial surveys in sociology: A review. *Social Science Research, 38*(3), 505-520. doi: 10.1016/j.ssresearch.2009.03.004

Welch, B. L. (1938). The significance of the difference between two means when the population variances are unequal. *Biometrika, 29*(3/4), 350-362. doi: 10.1093/biomet/29.3-4.350

White, H. (1980). A heteroskedastic-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica, 48*(4), 817-838. doi: 10.2307/1912934

Zimmerman, D. W. (2004). A note on preliminary tests of equality of variances. *British Journal of Mathematical and Statistical Psychology, 57*(1), 173-181. doi: 10.1348/000711004849222