CHEN, XIDAN, M.A. Simple Slopes are Not as Simple as You Think. (2013)
Directed by Dr. Douglas W. Levine. 44pp.

Simple slopes analysis is commonly used to evaluate moderator or interaction effects in multiple linear regression models. In usual practice, the moderator is treated as a fixed value when the standard error of simple slopes is estimated. The usual method used for choosing the conditional value of moderator (i.e., at one sample SD below, one SD above, and at the mean) makes the moderator a random variable and therefore renders the standard error suspect. In this study I examined whether the standard error used in post hoc probing for interaction effect is a biased estimator of the population variance when moderator is a random variable. I conducted Monte Carlo simulations to evaluate the variance of the simple slope under a variety of conditions corresponding to a 5 (sample size, N) x 5 (variance of focal predictor, *x*) x 5 (variance of moderator, *z*) x 4 (levels of *r*, the correlation between *x* and *z*) x 5(model fit, R^2) x 4 (population slope for interaction, b_{xz}) factorial design. I present circumstances under which usual practice yields an “almost” unbiased estimator and conditions when the estimator is more severely biased and less so.

SIMPLE SLOPES ARE NOT AS SIMPLE AS YOU THINK

by

Xidan Chen

A Thesis Submitted to
the Faculty of The Graduate School at
The University of North Carolina at Greensboro
in Partial Fulfillment
of the Requirements for the Degree of
Master of Arts

Greensboro
2013

Approved by

Committee Chair

APPROVAL PAGE

This thesis has been approved by the following committee of the Faculty of The Graduate School at The University of North Carolina at Greensboro.

Committee Chair ____

Committee Members ____

____

____
Date of Acceptance by Committee

____
Date of Final Oral Examination

ACKNOWLEDGEMENTS

I would firstly like to thank my advisor, Dr. Douglas Levine, for his guidance, encouragement and good advice. I am exceedingly grateful to him for taking me into the quantitative psychology study, which gave me the opportunity to challenge the possibilities I might otherwise never think of. Without his supervision, I would not have been able to finish this thesis. Many thanks must go to him. I also would like to thank my committee members, Dr. Faldowski and Dr. Marcovitch for their help and advice. Finally, I thank the Department of Psychology, UNC-Greensboro, for financially supporting me for my Master’s degree.

TABLE OF CONTENTS

Page

CHAPTER

CHAPTER I

INTRODUCTION

Simple Slopes Analysis

Researchers in psychology, education, and other social science disciplines have a long tradition of testing for moderator effects. Theories often hypothesize that the effect of a predictor (focal predictor) on an outcome will vary as a function of a third variable (moderator) (Baron & Kenny, 1986). To evaluate the interaction of two continuous variables, a common practice is to conduct a simple slopes analysis using multiple regression (Aiken & West, 1991). The purpose of the present study is to evaluate whether the commonly used estimator of the population variance of the simple slope is biased, and if so, whether this bias yields estimates that are too small or too large.

A common approach to examining significant interaction effects, or moderator effects, between two continuous variables is to employ simple slopes analysis. For simplicity, assume that an outcome *Y*, is predicted by two continuous variables *x* and *z*. The sample regression model can be expressed as:

$$Y = b_0 + b_x x + b_z z + b_{xz} x\mathrm{z} + e, \tag{1}$$

where b_0, b_x, b_z, and b_{xz} are the usual ordinary least squares (OLS) estimates of the regression parameters; *x* and *z* are in deviation (centered) form and the interaction form is constructed from the centered *x* and *z*. To evaluate whether there is an interaction effect, the practice in psychology is to examine the conditional relationships. The conditional relationship is often discussed as the relationship of variable *x* (focal predictor) to variable *Y* (outcome) “holding constant”, or “controlling for,” some variable *z* (moderator). Alternatively, a conditional relationship can be described as the relationship of variable *x* to variable *Y* as a function of values of moderator *z* (Aiken & West, 1991). One conditional relationship in the example regression model is the effect of *x* on *Y* for a particular value of *z*. Examination of conditional effects is often done by creating “simple slopes” which are just the estimated conditional effect of *x* on *Y* when *z* is set to a conditional value Z_{CV}. Specifically, the “simple slope” of the regression of *Y* on *x* conditional on $z = Z_{CV}$ is:

$$b_s = b_x + b_{xz} Z_{CV}. \tag{2}$$

The simple slope b_s is obtained using the OLS estimates of the regression parameters b_x and b_{xz} obtained from the regression of *Y* on *x*, *z*, and *xz* and then specifying Z_{CV}, at a specific value for *z*. A convention has arisen to choose three conditional values of *z*. Cohen and Cohen (1983, p. 323) have suggested using a “low” *z* score (one sample SD below the mean), an “average” z score (at the sample mean), and a “high” z score (one sample SD above the mean). These three values are herein denoted as Z_L, Z_M, and Z_H, respectively. The regression lines obtained using these conditional values of z yield three representative members of the simple slope family of lines.

For example, in Aiken and West’s book (Aiken & West, 1991, p.13), they gave the numerical example, Y= 2.54+ 1.14x + 3.58z + 2.58xz. Here b_0 = 2.54, b_x=1.14, b_z= 3.58, b_{xz}= 2.58, and b_s= 1.14 + 2.58 Z_{CV} . To obtain the simple slopes, Aiken & West chose Z_{CV} so that: Z_L = -2.20 (one standard deviation below the mean), Z_M = 0 (at the mean), Z_H = 2.20 (one standard deviation above the mean). By substituting these values of Z_{CV} (i.e., -2.20, 0, and 2.20) into equation (2), three simple slopes were generated: b_H = 6.82, b_M = 1.14, and b_L = -4.54. For instance, using b_s from above, b_H = 1.14 + 2.58(2.20) = 6.82.

Common practice is to examine whether a simple slope is statistically significant by testing the hypothesis that a simple slope differs from zero. This test is simply the t-test in which the simple slope (e.g., b_s) is divided by its standard error with (n-k-1) degrees of freedom (where n is the number of cases and k is the number of predictors). To understand the problem associated with using this test of the simple slope, it is important to appreciate the assumptions underlying the derivation of the standard error of the simple slope. To do this consider the variance of b_s, $V(b_s)$. This variance has been shown in many places (e.g., Aiken & West, 1991, p.16; Friedrich, 1982, p. 810; and Jaccard & Turrisi, 2003, p.26), to be:

$$V(b_s) = V(b_x) + 2Z_{CV}Cov(b_x, b_{xz}) + V\left(b_{xz}Z_{CV}\right) = V(b_x) + 2Z_{CV}Cov(b_x, b_{xz}) + (Z_{CV})^2 V(b_{xz}). \quad (3)$$

$V(b_s)$ is obtained by using algebra of expectations by assuming Z_{CV} is a fixed value. In practice, $V(b_s)$ is estimated by:

$$v(b_s) = s^2(b_x) + 2Z_{CV}s(b_x, b_{xz}) + (Z_{CV})^2 s^2(b_{xz}), \quad (4)$$

where $s^2(b_x)$ and $s^2(b_{xz})$ are the sample variances of b_x and b_{xz}, respectively; $s(b_x,b_{xz})$ is the sample covariance of b_x and b_{xz}. The estimated standard error of the simple slope is the square root of this quantity.

For Aiken and West's numerical example, $s(b_x) = 2.35$, $s(b_{xz}) = 0.40$, and $s(b_x,b_{xz}) = -0.08$. Substituting $Z_H = 2.20$ into equation (4) yields $V(b_H) = (2.35)^2 + 2(2.20)(-0.08) + (2.20)^2(0.40)^2 = 3.98$. The statistical test for simple slope b_H is: $t = b_H/s(b_H) = (6.82)/\ (3.98)^{.5} = 3.45$. Similar substitutions of the other two conditional values of z (i.e., Z_M, and Z_H) give the corresponding estimated variances of b_M and b_H which are summarized by Aiken & West (1991, p. 17) in their Table 2.3.

Statement of the Problem

In estimating the variance of simple slopes, the conditional value of *z* (i.e., the moderator) is assumed, as shown above, to be a constant. That is, the moderator is assumed to be a fixed value, *not* a random variable. Cohen et al. (2003, p.274) recognized this when they stated "in comparing the simple regression of *Y* on *x* at a value of *z* across different samples, the value of *z* must be held constant (fixed) across samples." As an example of a constant conditional value, consider a study in which the Body Mass Index (BMI) is used as a moderator. In computing the simple slopes, it would be reasonable to fix BMI equal to some CDC (Centers for Disease Control and Prevention) defined cut-point, such as 18.5 (underweight) or 30 (obese). In this case, these choices would be constant (i.e., fixed) across all studies. Suppose instead that we did not have a preexisting cut-point for “underweight” or “obese”, but defined these as 1 sample SD below the mean (underweight) and 1 sample SD above the mean (obese). In this situation the conditional values associated with “underweight” and “obese” will vary across samples because both the sample means and sample variances likely differ from sample to sample. Thus, the conditional value would be a random variable. Consequently, the equation (3) would yield an incorrect estimator of the variance given the conditional value is a random variable.

Choosing a random variable to serve as the conditional value is usual practice. Referring back to the example taken from Aiken and West's book, the simple slopes were computed by choosing conditional values of the moderator at its sample mean, at one sample standard deviation below its sample mean and at one sample standard deviation above its sample mean. Thus, the conditional values change across samples so these are, by definition, random variables and are not fixed. This is true of most studies using simple slopes methodology because researchers usually do not have meaningful cut-points for the conditional values of moderator that could be used across studies. A main argument underlying this study is that, by adopting the convention of basing the conditional values on the sample mean and variance, the chosen conditional values of moderator are random variables and are *not* fixed as the underlying derivations assume. It is crucial to understand that when Z_{CV} is a random variable, equation (3) does not yield the correct variance because this equation does not properly account for the variance and covariance of the product of random variables. For example, in equation (3) the term $V\left(b_{xz}Z_{CV}\right)$ is given as $(Z_{CV})^2V\left(b_{xz}\right)$ by the authors cited above when they present the variance of the simple slope. When Z_{CV} is a constant these two terms are equivalent. They are not, however, equivalent when Z_{CV} is a random variable. Goodman (1960) first provided a general derivation of the exact variance of the products of two random variables *x* and *y*. Bohrnstedt and Goldberger (1969, p.1439, equation (5)) re-expressed it as:

$$\begin{aligned}V(xy) = {} & E^2(x)V(y) + E^2(y)V(x) + E\left\{(\Delta x)^2(\Delta y)^2\right\} + 2E(x)E\left\{(\Delta x)(\Delta y)^2\right\} + \\ & 2E(y)E\left\{(\Delta x)^2(\Delta y)\right\} + 2E(x)E(y)C(x,y) - C^2(x,y),\end{aligned} \tag{5}$$

where *x* and *y* are jointly distributed random variables; $E(x)$, $E(y)$ are the expected values of *x, y* respectively; $C(x,y)$ is the covariance of *x* and *y*;

$\Delta x = x - E(x),\ \Delta y = y - E(y);\ \ V(x) = E(\Delta x)^2,\ \ V(y) = E(\Delta y)^2.$

From equation (5) we can generate the correct variance for the term $V(b_{xz}Z_{CV})$ when Z_{CV} is properly treated as a random variable:

$$\begin{aligned}V(b_{xz}Z_{CV}) = {} & E^2(b_{xz})V(Z_{CV}) + E^2(Z_{CV})V(b_{xz}) + E\left\{(\Delta b_{xz})^2(\Delta Z_{CV})^2\right\} + \\ & 2E(b_{xz})E\left\{(\Delta b_{xz})(\Delta Z_{CV})^2\right\} + 2E(Z_{CV})E\left\{(\Delta b_{xz})^2(\Delta Z_{CV})\right\} + \\ & 2E(b_{xz})E(Z_{CV})C(b_{xz},Z_{CV}) - C^2(b_{xz},Z_{CV}),\end{aligned} \tag{6}$$

where $\Delta b_{xz} = b_{xz} - E(b_{xz}),\ \Delta Z_{CV} = Z_{CV} - E(Z_{CV})$

Bohrnstedt and Goldberger (1969) extended this result in numerous ways. A detailed derivation for the variance of simple slope is provided in Appendix C.

Based on this overview of the derivation of the variance of the simple slopes, we can infer that the estimators in equation (3), and by extension in equation (4), are not unbiased estimators of the population variance of the simple slope, $\sigma^2_{b_s}$. Because equations (3) and (4) do not take account of all sources of variation shown in equation (12, Appendix C), under the common conditions when the covariance is positive, it is likely that $\sigma^2_{b_s}$ is underestimated. Consequently, the test of the null hypothesis that a simple slope equals zero will be rejected more often than it should be (i.e., the test is too liberal).

Despite equations (3) and (4) being technically inappropriate when *z* is a random variable, psychologists routinely use equation (4) to estimate the variance of b_s when testing the null hypothesis that a simple slope equals zero. This practice leads to the question, “Is the bias associated with equation (4) small?” In other words, does the inappropriate use of equation (4) result in variances that are essentially unbiased or are they much smaller than would be obtained had the appropriate estimator been applied?

Current Study

In this study, my interest centers on the estimation of the variance of the simple slope in standard regression models. I am aware of no prior work that has investigated the bias in the variance of simple slope. First goal for this study is to investigate whether the variance used in psychological research is an unbiased estimator, as claimed by proponents of the methodology (e.g., Aiken & West, 1991). Based on statistical theory I surmise that the estimator should be biased, and this bias is likely on the liberal side (i.e., too small). I further investigate under what circumstances equation (4) yields an “almost” unbiased estimator and when the estimator is severely biased. Given the complicated nature of the variance of the products of random variables (Bohrnstedt & Goldberger, 1969; Goodman, 1960), it is quite difficult to make accurate predictions as to all the conditions under which the estimator will perform well. Consequently, I investigate a range of conditions of using the factors indicated by equation (6) that are known to affect the variance of the product. These factors are described below.

CHAPTER II

METHOD

Simulation Studies: General Framework

Bohrnstedt and Goldberger's (1969) derivation of the variance of the product of random variables indicated that the variance of *x, z* and the population slope for interaction term in the regression equation should contribute to the estimation of the population variance of simple slopes. Therefore, those three factors are manipulated in this study. The correlation between *x* and *z* is also manipulated to examine the effects of collinearity on the variance. When simulating regression models, the data often fit the models quite well (i.e., very large R^2), thus I also manipulated the levels of R^2. Finally, the sample size is also manipulated in this study, because any bias may be reduced as sample size increases (Aiken & West, 1991, p. 22). Consequently, I conduct a 5 (Sample size) * 5 (variance of *x*) * 5 (variance of *z*) * 4 (levels of *r* the correlation between *x* and *z*) * 5 (R^2) * 4 (population slope for interaction, b_{xz}) factorial experiment. Within each of the 5*5*5*4*5*4 experimental conditions, 10,000 simulations are conducted.

All population and estimated values are calculated through a computer program which is written (in FORTRAN) specifically for this study. This program calls several

International Mathematics and Statistics Library (IMSL) subroutines. Pseudorandom numbers are generated from a multivariate normal distribution by using IMSL subroutine RNMVN.

Manipulated Factors

Sample Size. There are five levels of sample size: 100, 400, 1000, 2000, and 10,000 "cases" within each simulation.

Variance of *x* and *z*. There are five levels of population variances of *x* and *z*: 1, 5^2, 10^2, 16^2, and 28^2. I chose these levels, because levels such as 10^2, 16^2, and 28^2 correspond to the variance of some commonly used measures in Psychology. For example, one popular IQ test, the Stanford-Binet test has a standard deviation of 16.

Levels of *r* the correlation between *x* and *z*. There are four levels of the population correlation between *x* and *z* which are used in computing the population covariance of these variables: .00, .20, .40, .70.

Model Fit (R^2). Model fit is restricted to five strata such that the mean R^2 in a strata is in the region: (.20, .30], (.30, .40], (.40, .50], (.50, .60], (.95, .99].

Population Slope for Interaction (b_{xz}). There are four levels for b_{xz}: 1, 3, 5, 7.

Procedure

One purpose of this study is to investigate whether the commonly used estimator for the variance of the simple slope shown in equation (4), i.e., $v(b_s)$, provides an unbiased estimate of the true population variance of the simple slope, σ_b^2, when *z* is a random variable. To investigate this question I conduct a series of Monte Carlo simulations that correspond to the research design above. Because I want to determine whether $v(b_s)$ is an unbiased estimator I need to have a known unbiased estimator with which it can be compared. Fortunately it is well known that the sample variance is such an estimator. Thus, to determine whether $v(b_s)$ is an unbiased estimator I compare it to the sample variance of the simple slopes, i.e., $s^2(b_s)$.

Specifically, the 10,000 simulations within a given experimental condition yield 10,000 estimates of the simple slope, b_s. From these I compute the sample variance of the simple slopes, i.e., $s^2(b_s) = \sum_{i=1}^{10000}\left(b_{si} - \overline{b}_s\right)^2 / 9999$, where $\overline{b}_s$ is the average of the observed sample slopes (b_{si}). If equation (4) yields estimates that are unbiased (or nearly so), then the expected value (i.e., the average) of these estimates, $\overline{v}(b_s) = \sum_{i=1}^{10000} v(b_{si}) / 10000$, should be approximately equal to $\sigma_{b_s}^2$ and also to, $s^2(b_s)$, so that the ratio $\left\{\overline{v}(b_s) / s^2(b_s)\right\}$ should be approximately 1. If, on the other hand, equation (4) yields estimates that are too small when Z_{CV} is a random variable, the ratio $\left\{\overline{v}(b_s) / s^2(b_s)\right\}$ should be less than 1.

To replicate usual practice, Z_{CV} is chosen to be equal to ±1 sample standard deviation of *z*. Note I will center the raw data of *x* and *z* in this simulation study. According to the equation (2), when Z_{CV} is equal to 0, b_s will be the value of b_x which is assumed to be unbiased. Therefore, I am only concerned with evaluating Z_{CV} at the high and low level (but I do report the results “at the mean” to show what happens with an unbiased estimator). I denote the simple slopes obtained with the high and low values of Z_{CV} as b_H and b_L, where *H* and *L* stand for high (i.e., +1SD) and low (i.e., -1SD), respectively. Thus, the high and low values of Z_{CV} change from one simulation to another as the sample standard deviations of *z* change. Two other simple slopes are also computed by defining Z_{CV} as ±1σ, where σ is the population standard deviation obtained from the specified population covariance matrix. Denote the simple slopes obtained with these values of Z_{CV} as b_{Hf} and b_{Lf}, where *f* stands for fixed, *H* and *L* stand for high and low as before. In computing b_{Hf} and b_{Lf}, then, the high and low values of Z_{CV} are constant across all 10,000 simulations. Thus, for each of the 10,000 simulations in a condition, four estimates of $\sigma_{b_s}^2$ are computed using equation (4) corresponding to the four simple slopes b_H, b_L ,b_{Hf} and b_{Lf}; these four sample variances are denoted as $V(b_H)$,$V(b_L)$, $V(b_{Hf})$ and $V(b_{Lf})$.

As noted above, the factorial design has 5 (Sample size) * 5 (variance of *x*) * 5 (variance of *z*) * 4 (levels of *r* the correlation between *x* and *z*) * 5 (R^2) * 4 (population slope for interaction, b_{xz}) = 10,000 conditions. For each condition I conduct 10,000 simulations. Each simulation is conducted with *n* = 100, 400, 1,000, 2,000, and 10,000 cases each with a pair of random numbers *x* and *z* generated from a bivariate normal distribution, with variances and covariance as defined according to the condition. Using an example similar to that used by Aiken and West (1991, p.13) the outcome variable, *Y*, is created by manipulating b_{xz} as follows:

$Y_i = 5 + 1x_i + 5z_i + 1x_iz_i + e_i.$

$Y_i = 5 + 1x_i + 5z_i + 3x_iz_i + e_i.$

$Y_i = 5 + 1x_i + 5z_i + 5x_iz_i + e_i.$

$Y_i = 5 + 1x_i + 5z_i + 7x_iz_i + e_i.$

Following Champoux and Peters (1987), the size of the dependent variable error term, e_i, is systematically varied so that the levels of R^2 can be manipulated as mentioned above. After computing the OLS estimates for *x*, *z* and the interaction, I compute the four simple slopes (b_H, b_L, b_{Hf} and b_{Lf}), as well as $\overline{v}(b_s)$ (i.e., the mean variance using equation

(4)) and $s^2(b_s)$ (i.e., the unbiased variance) for those simple slopes as described above. To evaluate whether the simple slope estimator is unbiased I compute, for each experimental condition, the ratios $\left\{\overline{v}(b_H)/s^2(b_H)\right\}$, $\left\{\overline{v}(b_L)/s^2(b_L)\right\}$, $\left\{\overline{v}(b_{Hf})/s^2(b_{Hf})\right\}$ and $\left\{\overline{v}(b_{Lf})/s^2(b_{Lf})\right\}$. For completeness, I also compute their ratios when is Z_{CV} at the mean.

Index of Underestimation. I report the index of underestimation as final result. To simplify the presentation of results, the value of this index is computed as the average of the ratio $\left\{\overline{v}(b_s)/s^2(b_s)\right\}$ at high (i.e., +1SD) and low (i.e., -1SD) minus one.

Using the logic from above, note that if equation (4) yields estimates that are unbiased (or nearly so), then $\overline{v}(b_s)$ should be approximately equal to $s^2(b_s)$ so that the ratios $\left\{\overline{v}(b_H)/s^2(b_H)\right\}$ and $\left\{\overline{v}(b_L)/s^2(b_L)\right\}$ should each be approximately 1, while the corresponding indices of underestimation should be approximately 0. If, on the other hand, equation (4) yields estimates that are too small when Z_{CV} is a random variable, the ratios $\left\{\overline{v}(b_H)/s^2(b_H)\right\}$ and $\left\{\overline{v}(b_L)/s^2(b_L)\right\}$ should each be less than 1, the indices of underestimation should be negative. When Z_{CV} is fixed, statistical theory indicates that the ratios $\left\{\overline{v}(b_{Hf})/s^2(b_{Hf})\right\}$ and $\left\{\overline{v}(b_{Lf})/s^2(b_{Lf})\right\}$ should each be approximately 1 and the indices of underestimation should be approximately 0.

CHAPTER III

RESULTS

Mean Comparisons

Table 1 provides the results for each factor level when the moderator is fixed, at the mean, and random. Overall, in this table I present the index of underestimation averaged across other conditions (i.e., there are results for the main effects). If $\bar{v}(b_s)$ yields estimates that are too large, the index will be indicated by a positive value. If the estimates are too small, the index will be indicated by a negative value. If the estimate is close to the unbiased value then the tabled value will be near zero. For example, when n=100, the index is -.2346 for the random condition, indicating that, on average, the variance obtained using equation (4) is 23.46% below that obtained using the unbiased estimator, an indication that equation (4) yields estimates that are too liberal. On the other hand, when the conditional value is fixed, not random, for n=100, the index is -.0005 indicating that the variance estimator is unbiased.

As expected, when moderator is fixed and at the mean, the underestimation for each factor level is approximately 0. However, when the moderator is random, results are all negative over the levels of the factors. For each factor level condition, there is approximately a 20% under estimation. The underestimation across different factor levels is plotted in figure 1. It is seen that for the fixed condition underestimation index is consistently around 0. As expected, this is also true for at the mean condition. These findings indicate that, across a wide variety of manipulated factor levels, when moderator is fixed or at the mean, equation (4) performs well. Whereas, by adopting the convention of using the moderator at one standard deviation above the mean and one standard deviation below the mean, equation (4) yields a liberal estimator of the variance.

Trend Analysis

An examination of Table 1 reveals that, with the exception of sample size and correlation, there appears to be a trend for the underestimation to consistently increase as the levels of the factors increase. To investigate this, trend analyses across the levels for each factor were conducted. As shown in Table 2, for the random simple slopes, trends were statistically significant for the factors: R^2, the variance of *x*, the variance of *z*, and the slope of the interaction. For these factors the trends were such that the bias (i.e., the underestimation) became more pronounced as the value of the factor level increased. This was most pronounced for R^2. The effect was less pronounced for the variance of *x*, the slope of the interaction, and the variance of *z*. For these factors, there were no statistically significant trends when the moderator was fixed or “at the mean”.

For the correlation coefficient a trend was observed when the moderator was a random variable, such that the bias became more extreme as collinearity was reduced. There were also statistically significant trends observed when the moderator was fixed and “at the mean”. While these latter two trends were statistically significant, there was no clear bias evident. For example, as shown in Table 1, when the moderator was fixed, the index of underestimation changed over the four levels of correlation is as follows: -0.0007, 0.0000, 0.0001, 0.0001. Thus, while there was an increasing trend there is not apparent evidence of bias.

Finally, the sample size was studied because Aiken and West (1991, p. 22) suggested that any bias would be reduced with sample size. Interestingly, then, I did not observe a decrease in the bias when the moderator was a random variable as sample size increased from 100 cases to 10,000 cases. There was no statistically significant linear ($F(1, 9995) = 0.10$, $p = .75$) or nonlinear trends ($F(3, 9995) = 0.21$, $p = .89$). The index of underestimation averaged -0.24 across all sample sizes. On the other hand, there were statistically significant effects for both the fixed and “at the mean” simple slopes. While these effects were statistically significant, the “trends” were not apparent when looking at the index of underestimation across the levels of sample size. For the fixed moderator, over the five levels of sample size (from 100 to 10,000), the index of underestimation changed as follows: -0.0005, -0.0004, 0.0014, -0.0017, 0.0006. The values are all close to zero and there is no systematic under or overestimation evident across the levels of sample size.

Effect of Factors

To examine the effect of factors on the underestimation index, I conducted a five-way ANOVA. Sample size was excluded from the analysis as it had no significant effect with the random moderator. Excluding this factor also permitted a full factorial analysis. Otherwise, in a six-way ANOVA, examining the highest order interaction would not be possible as there would only be 1 case per cell. ANOVA results are presented in Table 3. Note with 10,000 "cells" (i.e., conditions), all effects and interactions are statistically significant, so p-values are not reported, and effect sizes are given.

All main effects are large as expected (Table 3). Results are sorted from highest to lowest η^2. The largest main effect is R^2 ($\eta^2 = .999$), smallest is the variance of *z* ($\eta^2 = .758$). Overall, underestimation ranges from 5% to 82%, with a mean of 24%. That is all factors I manipulated here are influential in estimating the population variance of simple slopes for the random conditions.

While all interactions with R^2 are statistically significant, R^2 continues to have the largest effect on bias even within the interactions. Figure 2 shows the 2-way interactions with R^2. The interaction effects appear very slight and the general pattern resembles the form of the R^2 main effect. For example, for the interaction with the variance of *x*, the bias is lowest when the variance of *x* equals one. The biases associated with other levels of *x* (*x* =5^2, 10^2, 16^2, and 28^2) are roughly all the same as indicated by lines that are mostly coincident. The same pattern also holds for the variance of *z*. Similarly the bias is lowest when the variance of *z* equals one. These patterns also hold for the three-way interaction between R^2, the variance of *x* and b_{xz} as seen in Figure 3.

Figure 4 shows the two-way interactions between the variance of *x* and the slope b_{xz}, as well as the variance of *z* and the slope b_{xz}. Here again the picture resembles two main effects. The interaction effect is quite small and mainly appear to be evident in the slope of the line connecting the first two levels of the variance of *x* or *z* within each level of b_{xz}. Put simply, these figures closely resemble those of the main effect for the variance of *x* and the variance of *z*. But now there are four additional almost parallel lines reflecting the levels of b_{xz}. There does not appear to be a new story told by these analyses. The bias is less when the variance of *x* or the variance of *z* equals one. The bias is also less when the slope b_{xz} is one. The other levels of the variance of *x* and *z* are quite similar when $b_{xz} > 1$ as indicated by the coincident lines.

CHAPTER IV

DISCUSSION

Summary of Current Study

This study evaluates the estimator of the variance of the simple slope used in simple slopes analysis. These simulations provide results that have long been ignored in the methodological literature. When moderator values are fixed, the usual practice using equation (4) will yield unbiased estimators. However, when the moderator is a random variable, but treated as a fixed value and using equation (4), the population variance of the simple slope is underestimated. Therefore, the test of the moderator effect will result in variances that are too small (liberal). I also showed how sample size, the variance of the predictor (*x*) and the moderator (*z*), the correlation between those two (*r*), model fit (R^2), and the interaction slope (b_{xz}) all affect the estimate of the population variance of simple slopes and have illustrated that the variance is severely underestimated when R^2 exceeds .90. In contrast to statements made in the literature (e.g., Aiken & West, 1991; Cohen, et.al, 2003), increasing sample size has no effect on the bias.

To summarize the findings, equation (4) is a biased estimator of the variance of the simple slope if the moderator is a random variable. In this situation, the bias becomes worse with better model fit (increasing R^2), less collinearity (smaller *r*), larger variance of *x* and *z*, and does not improve with sample size. These are undesirable characteristics for a statistical estimator.

Implications for the Applied Researcher and Future Directions

Simple slopes analysis has likely had such popularity because it is easy to do. Our results indicate that when the variance of *x* and *z* and the slope b_{xz} are equal to one, the bias is reduced. Consequently, it seems that a possible approach would be for researchers to use standardized solutions in evaluating their models. This is, of course, a tentative suggestion as I have not evaluated whether this transformation will reduce bias when it exists. Further research will evaluate this recommendation. It is also possible that a better estimator than equation (4) can be obtained by using bootstrap methods. I have tried some pilot analyses that are promising but this is also an area of future research.

Simple slopes analysis, while simple to do, does not fully answer questions most researchers are posing. Few researchers start with a question about conditional values of the moderator that are one standard deviation above and below the mean. When there is an interaction between one predictor and one moderator, researchers would be more interested in knowing where the areas of significance lie. This is the type of question answered by using the Johnson-Neyman (J-N) procedure used to evaluate the significant interaction between a continuous and a categorical variable. Recently some research has been presented for using this method with continuous variables (Bauer & Curran, 2005). The J-N technique also provides the test of specific simple slopes. Additionally, it provides two additional indices, regions of significance and confidence bands. Regions of significance provide a range of moderator over which the effect of focal predictor is significant. The computation of regions of significance depends on the selection of type I error rate. Confidence bands indicate the precision of the estimation of the effect of focal predictor over the entire range of moderator. However, it is noted that J-N technique also involves the selection of conditional values of the moderator. The estimation of the variance of simple slopes in J-N technique is same as that is shown in equation (4). It is likely therefore that the when the conditional values of moderator are selected at one SD below the mean and one SD above the mean, the variance of simple slopes at those two values would be underestimated. For example, Bauer and Curran (2005) tested the predictability of child antisocial behavior and hyperactive behavior on child math ability using a sample of N=956 children. In their example, the outcome (*y*) is math ability, focal predictor (*x*) is antisocial behavior, and moderator (*z*) is hyperactive behavior. They first conducted a fixed-effect regression model with a continuous (antisocial behavior) by continuous (hyperactive behavior) interaction. Four additional covariates (age, grade, sex, and minority status) were entered. Regression analysis show that the interaction was statistically significant (*p* = 0.006). To further probe this significant interaction, simple slopes were computed at the mean of hyperactivity, one standard deviation above and one standard deviation below the mean. Only the simple slope of high hyperactivity was found to be statistically significant (*p* = .045). However, when the adopting the convention selecting the conditional value at one standard deviation above and below the mean, our results show that the variance of the simple slopes will be underestimated. Therefore, it would be easier for researchers to obtain a statistically significant result from the test of the simple slopes using equation 4 than using an unbiased estimator. In their example, given a sample size of 956, the average underestimation could be as large as 24% according to our simulation study. If I suppose a 24% underestimation happened, then using their simple slopes analysis, the simple slope of high hyperactivity is also found to be not statistically significant (*t* = 1.8, *p* = 0.072, *df* = 955)[1]. Further research is needed to investigate whether this approach to applying the J-N technique will be useful.

REFERENCES

Aiken, L. S., & West, S. G. (1991). *Multiple regression: Testing and interpreting interactions*. Newbury Park, CA: Sage.

Baron, R. M., & Kenny, D. A. (1986). The moderator-mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology, 51*, 1173–1182.

Bauer, D. J., & Curran, P. J. (2005). Probing interactions in fixed and multilevel regression: inferential and graphical techniques. *Multivariate Behavioral Research, 40*, 373–400.

Bohrnstedt, G. W., & Goldberger, A. S. (1969). On the exact covariance of products of random variables. *Journal of the American Statistical Association, 64*, 1439-1442.

Champoux, J. E. C., & Peters, W. S. (1987). Form, effect size and power in moderated regression analysis. *Journal of Occupational Psychology, 60*, 243-255.

Cohen, J., & Cohen, P. (1983). *Applied multiple regression/correlation analysis for the behavioral sciences (2rd ed.)*. Hillsdale, NJ: Lawerence Erlbaum Associates, Inc.

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences (3rd ed.)*. Mahwah, NJ: Lawerence Erlbaum Associates, Inc.

Friedrich, R. J. (1982, Nov.). In defense of multiplicative terms in multiple regression equations. *American Journal of Political Science, 26*, 797-833. doi: 10.2307/2110973

Goodman, L. A. (1960). On the exact variance of products. *Journal of the American Statistical Association,* 55, 708-713.

Jaccard, J., & Turrisi, R. (2003). *Interaction effects in multiple regression (2nd ed.).* Thousand Oaks, CA: Sage.

FOOTNOTES

1. I can estimate the slope by using rise over run from Bauer & Curran's (2005) Figure 1. That is, at $x = 0$, $y = 39.3$ and at $x = 5$, $y = 42.5$, so the slope would be 3.2/5=.64. Thus, with b =.64 and t=2.007289, SE = .3188; 1.24SE^2 = .126, then the new t=.64/.355=1.802 and p=.072.

APPENDIX A

TABLES

Table 1. Average Index of Underestimation by Factor Level

Factor	Fixed	At mean	Random
Sample Size			
100	-0.0005	-0.0004	-0.2346
400	-0.0004	-0.0002	-0.2402
1000	0.0014	0.0009	-0.2399
2000	-0.0017	-0.0014	-0.2423
10000	0.0006	0.0022	-0.2408
R^2			
(.20,.30]	-0.0002	0.0004	-0.0447
(.30,.40]	0.0000	0.0007	-0.0729
(.40,.50]	0.0000	-0.0001	-0.1092
(.50,.60]	-0.0003	-0.0001	-0.1552
(.95,.99]	-0.0002	0.0000	-0.8159
Variance of *X*			
1	-0.0003	0.0000	-0.1669
5^2	0.0000	0.0001	-0.2473
10^2	-0.0003	-0.0001	-0.2591
16^2	0.0003	0.0008	-0.2614
28^2	-0.0003	0.0002	-0.2632
Variance of *Z*			
1	-0.0004	0.0003	-0.2116
5^2	-0.0004	-0.0004	-0.2443
10^2	0.0001	0.0002	-0.2465
16^2	0.0000	0.0005	-0.2478
28^2	0.0000	0.0005	-0.2476
Interaction Slope			
1	-0.0002	0.0000	-0.1909
3	0.0000	0.0004	-0.2440
5	-0.0001	0.0002	-0.2584
7	-0.0002	0.0002	-0.2650
Correlation			
0.00	-0.0007	0.0004	-0.2602
0.20	0.0000	0.0008	-0.2553
0.40	0.0001	-0.0001	-0.2442
0.70	0.0001	-0.0003	-0.1985

Table 2. Trend Analysis (F-values are reported)

Factor		Fixed	At mean	Random
R^{2a}				
	linear	0.24	1.44	121265.55**
	deviation	0.27	1.24	3431.20**
Variance of x^a				
	Linear	0.10	0.44	36.56**
	deviation	1.78	1.26	38.15**
Variance of z^a				
	linear	1.20	1.62	4.42*
	deviation	0.58	1.03	5.75**
Interaction Slope^b				
	linear	0.02	0.09	77.84**
	deviation	0.43	0.53	8.14**
Correlation^b				
	linear	5.38*	4.89*	58.24**
	deviation	1.94	1.46	3.84*
Sample Size^a				
	linear	7.24**	40.19**	0.10
	deviation	29.11**	9.85**	0.21

*p <.05. **p<.01. *Note:* ^{a}*df* denominator=9995, ^{b}*df* denominator = 9996; linear represents linear trend, deviation is departure from linearity.

Table 3. ANOVA for the Random Conditions (Results ordered by η^2)

Effect and Source	*df*	Sum of Squares	*F*	η^2
R^2	4	505.481	1337535.421	0.999
X	4	8.009	21193.618	0.955
$R^2 * x * \mathrm{b}$	48	5.613	1237.790	0.937
B	3	5.051	17820.981	0.930
$R^2 *x$	16	4.604	3045.854	0.924
R	3	3.536	12473.942	0.903
$R^2 * \mathrm{b}$	12	3.273	2887.110	0.896
$x * \mathrm{b}$	12	2.825	2491.649	0.882
Z	4	1.183	3130.443	0.758
$z * \mathrm{b}$	12	0.944	832.531	0.714
$R^2 * r$	12	0.853	752.346	0.693
$R^2 * z * \mathrm{b}$	48	0.647	142.660	0.631
$R^2 * z$	16	0.461	305.085	0.550
$X * z * \mathrm{b}$	48	0.235	51.795	0.383
$R^2 * x * z * \mathrm{b}$	192	0.138	7.603	0.267
$x * r$	12	0.108	95.333	0.222
$R^2 * x * z * \mathrm{b} * r$	576	0.090	1.659	0.193
$x * z$	16	0.070	46.065	0.156
$R^2 * x * \mathrm{b} * r$	144	0.056	4.097	0.129
$\mathrm{b} * r$	9	0.051	59.726	0.118
$R^2 * x * r$	48	0.050	11.024	0.117
$R^2 * \mathrm{b} * r$	36	0.048	14.211	0.113
$R^2 * z * \mathrm{b} * r$	144	0.041	3.018	0.098
$R^2 * x * z$	64	0.040	6.665	0.096
$R^2 * x * z * r$	192	0.027	1.503	0.067
$x * z * \mathrm{b} * r$	144	0.027	1.964	0.066
$R^2 * z * r$	48	0.025	5.526	0.062
$x * \mathrm{b} * r$	36	0.022	6.555	0.056
$x * z * r$	48	0.022	4.875	0.055
$z * \mathrm{b} * r$	36	0.013	3.795	0.033
$z * r$	12	0.008	6.958	0.020

Note: x denotes variance of *x, z* denotes variance *of z,* b denotes the population slope for interaction, *r* denotes correlation between *x* and *z.*

APPENDIX B

FIGURES

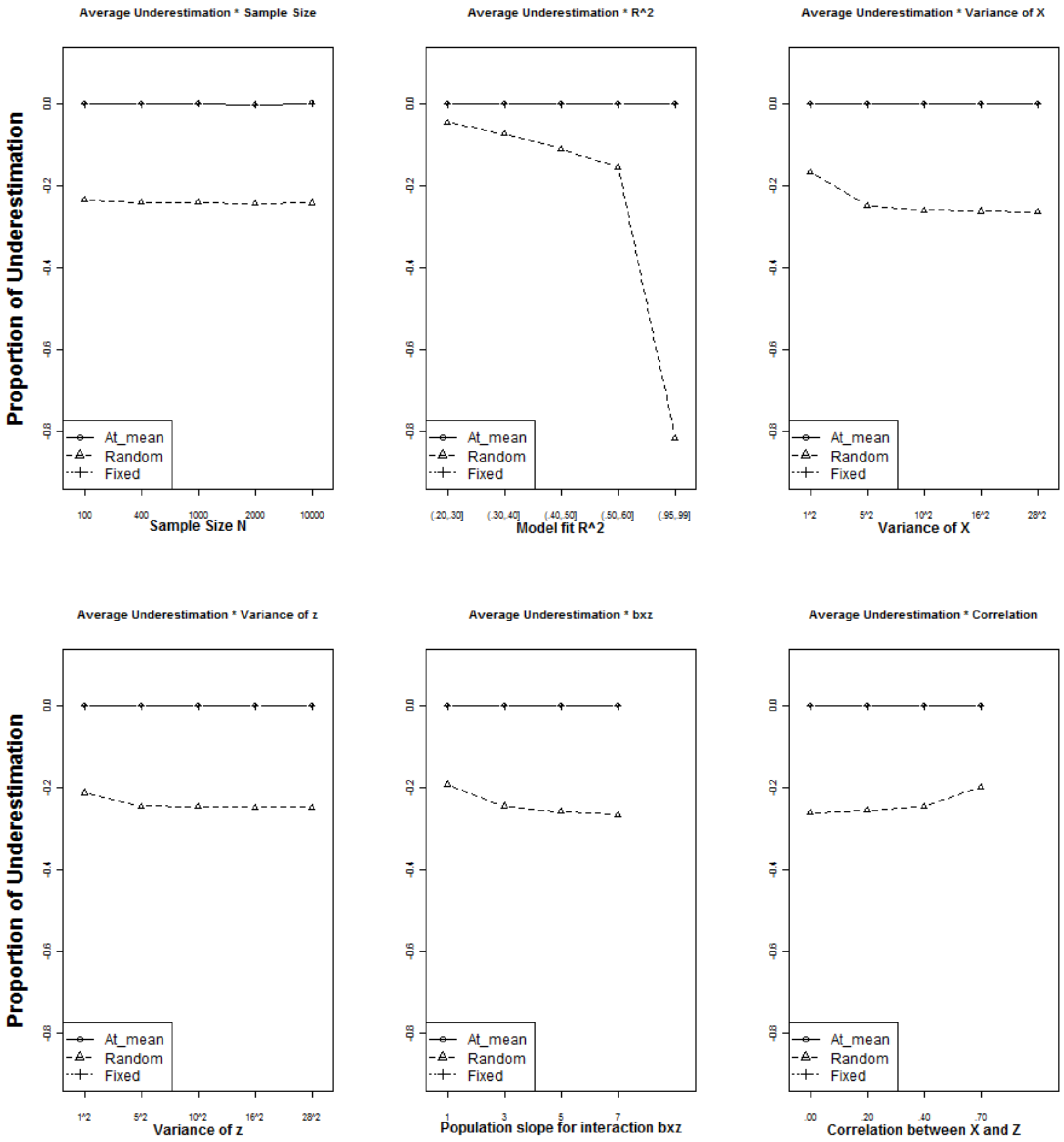

Figure 1. Average Index of Underestimation by Factor Levels

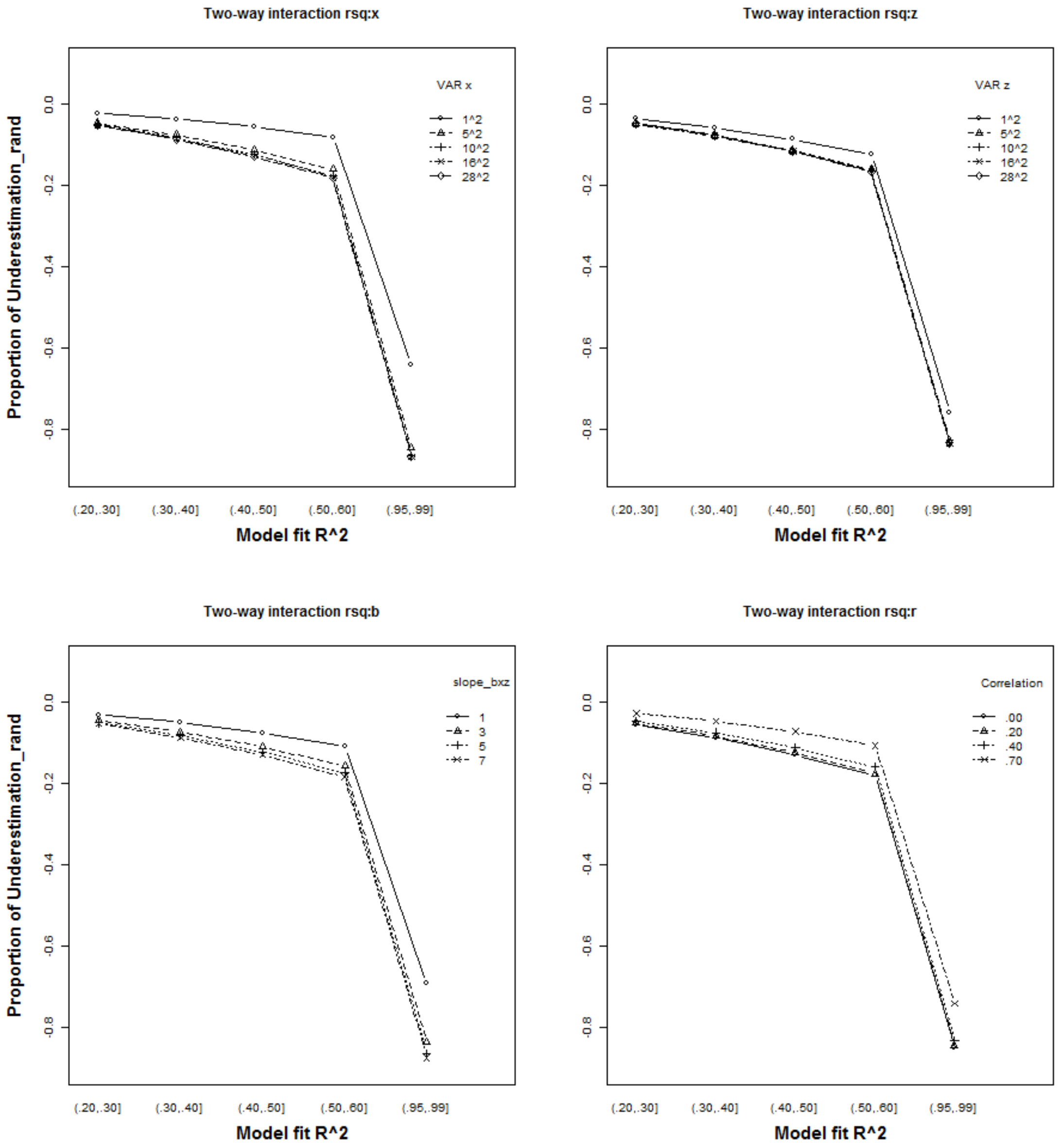

Figure *2*. Two-way interaction: R^2*x, R^2*z, R^2*b, R^2*r

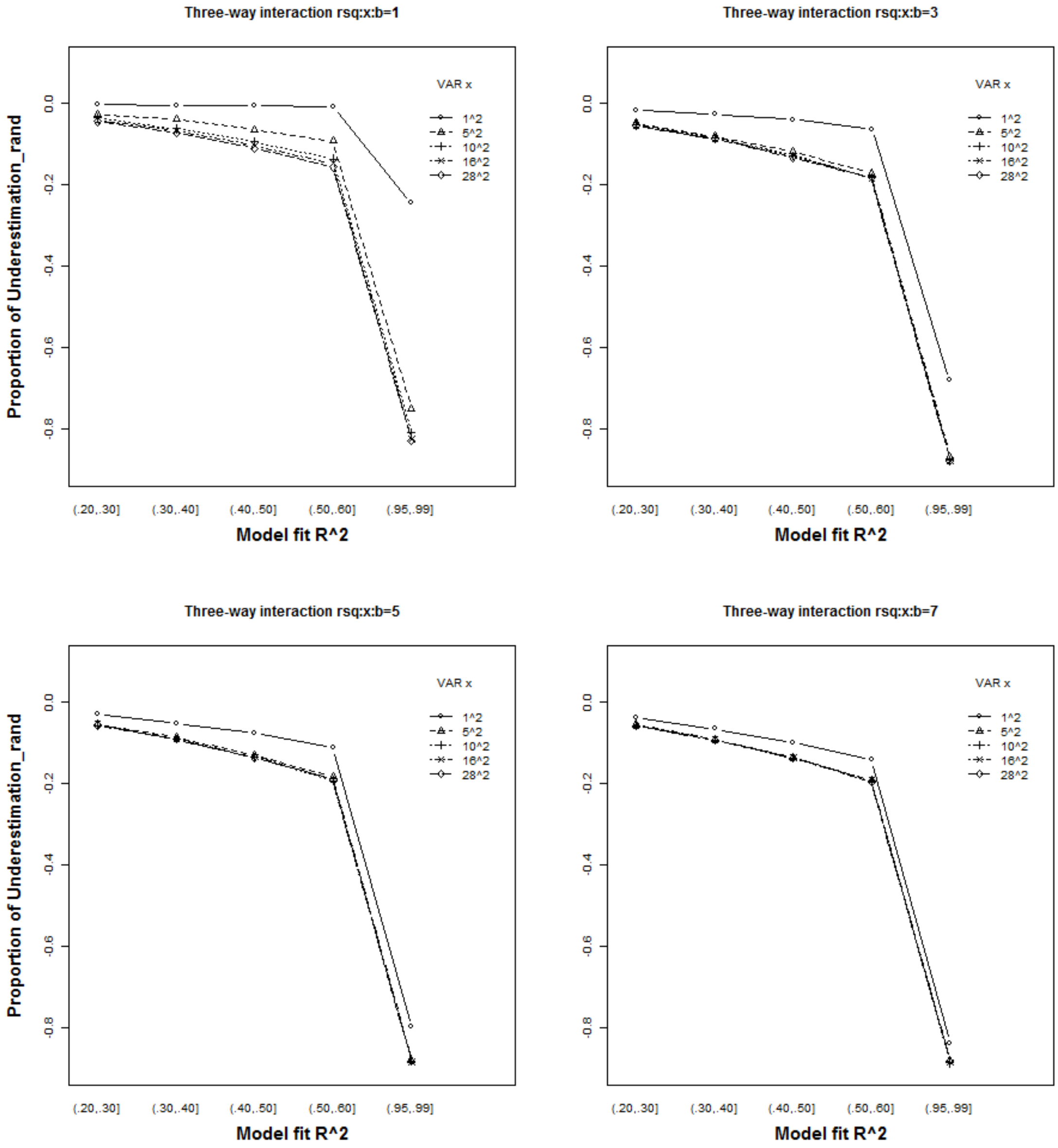

Figure 3. Three-way interaction: R^2*x*b

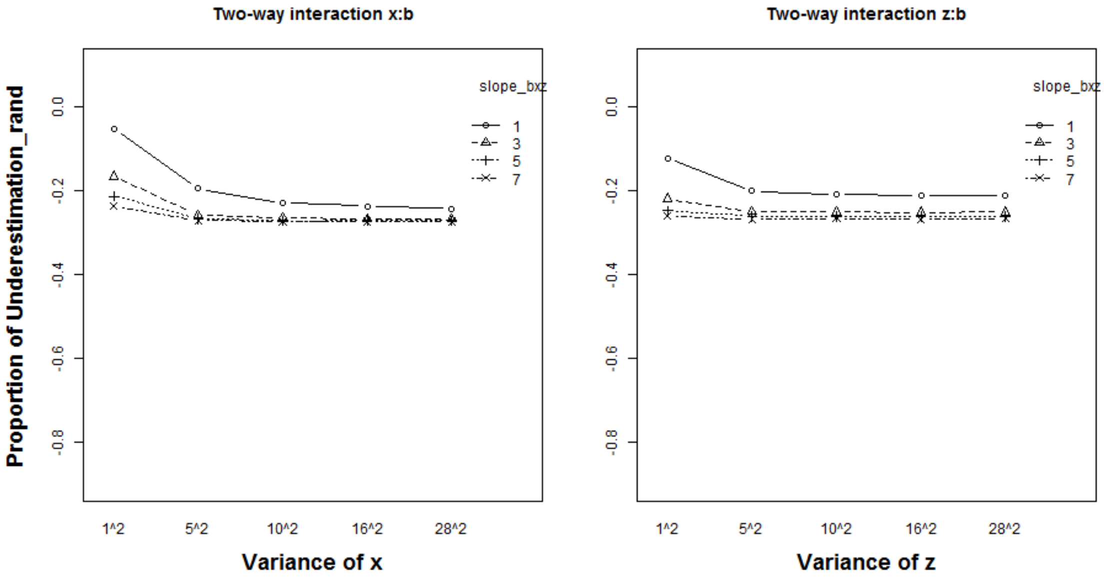

Figure 4: Two-way interaction: *x**b, *z**b

APPENDIX C

DERIVATION OF THE VARIANCE OF SIMPLE SLOPES

Bohrnstedt & Goldberger (1969) extended Goodman's (1960) results in numerous ways. For example, they showed that if *x* and *y* are bivariate normally distributed, then the equation (5) reduces to:

$$\begin{aligned} V(xy) = {} & E^2(x)V(y) + E^2(y)V(x) + E\left\{(\Delta x)^2(\Delta y)^2\right\} + 2E(x)E\left\{(\Delta x)(\Delta y)^2\right\} \\ & + 2E(y)E\left\{(\Delta x)^2(\Delta y)\right\}. \end{aligned} \tag{7}$$

Using Goodman (1960, p.709, equation (2)), this equation can be further simplified to:

$$V(xy) = E^2(x)V(y) + E^2(y)V(x) + V(x)V(y). \tag{8}$$

We can substitute *x*, *y* with b_{xz} and Z_{CV}, then equation (8) can be rewritten to:

$$V(b_{xz}Z_{CV}) = E^2(b_{xz})V(Z_{CV}) + E^2(Z_{CV})V(b_{xz}) + V(b_{xz})V(Z_{CV}) \tag{9}$$

For the covariance of products of two random variables, Bohrnstedt & Goldberger (1969, p.1441, equation (12)) specifies this as:

$$C(xy,v) = E(x)C(y,v) + E(y)C(x,v) + E\left\{(\Delta x)(\Delta y)(\Delta v)\right\}, \tag{10}$$

where *x*, *y* and *v* are independent random variables.

We now turn to our case, plug in $b_{xz} = x$, $Z_{CV} = y$ and $b_x = v$, therefore, the covariance of $b_{xz}Z_{CV}$ and b_x is:

$$C(b_{xz}Z_{CV}, b_x) = E(b_{xz})C(Z_{CV}, b_x) + E(Z_{CV})C(b_{xz}, b_x) + E\left\{(\Delta b_{xz})(\Delta Z_{CV})(\Delta b_x)\right\}. \tag{11}$$

Therefore, the derivation for the variance of simple slope, by assuming b_x, b_{xz} and Z_{CV} are independent, is:

$$\begin{aligned} V(b_s) &= V(b_x) + V\left(b_{xz} Z_{CV}\right) + 2C\left(b_{xz} Z_{CV}, b_x\right) \\ &= V(b_x) + V\left(Z_{CV}\right)\left\{E^2\left(b_{xz}\right) + V\left(b_{xz}\right)\right\} + \\ &E^2\left(Z_{CV}\right)V\left(b_{xz}\right) + 2E\left\{\left(\Delta b_{xz}\right)\left(\Delta Z_{CV}\right)\left(\Delta b_x\right)\right\}. \end{aligned} \tag{12}$$

Since b_x, b_{xz} and Z_{CV} are independent, $C\left(Z_{CV}, b_x\right) = 0$ $C\left(b_{xz}, b_x\right) = 0$. By adopting the convention, taking Z_{CV} at one standard deviation above the mean, below the mean and at the mean, we can infer that when $Z_{CV} = Z_H = S_z + \overline{Z}$ (one standard deviation above the mean), equation (12) can be written as:

$$\begin{aligned} V(b_s) = V(b_x) &+ V\left(S_z + \overline{Z}\right)\left\{E^2\left(b_{xz}\right) + V\left(b_{xz}\right)\right\} + \left\{E(S_z) + E(\overline{Z})\right\}^2 V\left(b_{xz}\right) \\ &+ 2E\left\{\left(\Delta b_{xz}\right)\left(\Delta(S_z + \overline{Z})\right)\left(\Delta b_x\right)\right\}, \end{aligned} \tag{13}$$

where $E\left(\overline{Z}\right) = \mu_z$, $E(S_z) = (4n-4)S_z/(4n-3)$.

If Z is centered, then $\overline{Z} = 0$ equation (13) is:

$$\begin{aligned} V(b_s) = V(b_x) &+ V\left(S_z\right)\left\{E^2\left(b_{xz}\right) + V\left(b_{xz}\right)\right\} + E^2(S_z)V\left(b_{xz}\right) \\ &+ 2E\left\{\left(\Delta b_{xz}\right)\left(\Delta S_z\right)\left(\Delta b_x\right)\right\}. \end{aligned} \tag{14}$$

For the same reason, we can get the variance equation when $Z_{CV} = Z_L = S_z - \overline{Z}$,

$$\begin{aligned} V(b_s) = V(b_x) &+ V\left(S_z - \overline{Z}\right)\left\{E^2\left(b_{xz}\right) + V\left(b_{xz}\right)\right\} + \left\{E(S_z) - E(\overline{Z})\right\}^2 V\left(b_{xz}\right) \\ &+ 2E\left\{\left(\Delta b_{xz}\right)\left(\Delta(S_z - \overline{Z})\right)\left(\Delta b_x\right)\right\}. \end{aligned} \tag{15}$$

When $Z_{CV} = Z_M = \overline{Z}$,

$$
\begin{aligned}
V(b_s) = V(b_x) + V\left(\overline{Z}\right)\left\{E^2\left(b_{xz}\right) + V\left(b_{xz}\right)\right\} + E^2(\overline{Z})V\left(b_{xz}\right) \\
+ 2E\left\{\left(\Delta b_{xz}\right)\left(\Delta(\overline{Z})\right)\left(\Delta b_x\right)\right\}.
\end{aligned}
\tag{16}
$$