Testing Average Effects in Regression Models with Interactions

Rolf Steyer, Felix Flory

Friedrich Schiller University, Jena, Germany

Andreas Klein

University of Illinois at Champaign

Ivailo Partchev, Safir Yousfi, Marc Müller and Ulf Kröhne

Friedrich Schiller University, Jena, Germany

Abstract

Testing the *average effect* of a variable X on an outcome variable Y in a regression model with an interaction term $X \cdot Z$ (and perhaps more complicated terms such as $X \cdot Z^2$) plays an important role, because testing this *average effect* means testing the "main effect" of X in the presence of interaction effects between X and a covariate Z that may be correlated with X. If Z is a *stochastic regressor* – in this case, the sample mean and the distribution in the sample will be different from one sample to the next – we show that the classical linear model method of testing the average effect as well as the analysis of covariance can cause inflated α-errors. The error rate increases the larger the interaction effect becomes. Analysis of covariance shows only inflated type I errors if group sizes (groups defined by the values of a discrete treatment variable X) differ. We also show that the Wald test with specific nonlinear constraints as implemented in LISREL and MPlus can be applied in the case of a *linear* effect function yielding a valid test of significance and valid standard errors for the estimate of the average treatment effect even for sample sizes as small as $N = 25$. In another simulation study we show that the sample sizes should be at least 100 in order to guarantee a valid significance test and valid standard errors for cases with very large interaction effects and a *quadratic* effect function.

Keywords: Average effects, average treatment effects, moderator models, stochastic regressors, interaction effects in regression models, Wald test, likelihood-ratio test, general linear model, analysis of covariance

Testing Average Effects in Regression Models with Interactions

Many hypotheses in the social and behavioral sciences postulate that the effects of a variable *X* on another variable *Y* depend on the values of a third variable *Z*. The simplest example is a dichotomous treatment variable *X* with values 0 (for *control*) and 1 (for *treatment*), a dichotomous variable *Z* with values 0 (e.g., *low need* for the treatment) and 1 (e.g., *high need* for the treatment) and a continuous outcome variable *Y* measuring the *success* of the treatment. In such a case

$$E(Y \mid Z, X) = \gamma_{00} + \gamma_{01} Z + \gamma_{10} X + \gamma_{11} X \cdot Z \tag{1}$$

is a saturated parameterization of the regression of *Y* on *X* and *Z*. (The reasons for this unusual choice of the indices will become obvious in the following paragraphs.)

Conditional Effects

Equation 1 may also be written

$$E(Y \mid Z, X) = (\gamma_{00} + \gamma_{01} Z) + (\gamma_{10} + \gamma_{11} Z) \cdot X. \tag{2}$$

Considering the two conditional regressions of *Y* on *X* given $Z = z$ yields:

$$E_{Z=0}(Y \mid X) = \gamma_{00} + \gamma_{10} X \tag{3}$$

$$E_{Z=1}(Y \mid X) = (\gamma_{00} + \gamma_{10}) + (\gamma_{10} + \gamma_{11}) \cdot X\,. \tag{4}$$

The interpretation of the regression coefficients is simple: γ_{10} is the *conditional effect of X* if $Z = 0$ and $\gamma_{10} + \gamma_{11}$ is the *conditional effect of X* if $Z = 1$. Furthermore, γ_{00} is the intercept of the

conditional regression of Y on X given $Z = 0$ and $\gamma_{00} + \gamma_{10}$ is the intercept of the conditional regression of Y on X given $Z = 1$. The parameter γ_{11} is called the *interaction parameter*: If $\gamma_{11} = 0$, then the conditional effects of X on Y given Z do *not* depend on the values z of Z, otherwise the conditional effects of X are modified by Z. Hence, we call Z a *modifier* and $\gamma_{10} + \gamma_{11} Z$ the *effect function*. Synonyms for these terms are *moderator* and *moderator functions* (see, e.g., Saunders, 1956 or Baron & Kenny, 1986). Classical linear model procedures of estimating and testing hypotheses about (linear combinations of) the parameters in Equation (1) can be applied. All this is well-known (see, e.g., Moosbrugger, 1981; Gosslee & Lucas, 1965) and documented by Aiken and West (1991) or Cohen, Cohen, West, & Aiken, 2003), for example.

Generalizations

Generalizations of these regression models with interaction terms are *straightforward*: First, if Z is not dichotomous, the functions $g_0(Z) := \gamma_{00} + \gamma_{01} Z$ and $g_1(Z) := \gamma_{10} + \gamma_{11} Z$ in Equation (2) do not necessarily yield a saturated parameterization for the regression $E(Y| X, Z)$. In this case, these linear functions in Equation (2) may have to be replaced by more general functions, e.g., by polynomial functions of a higher degree, orthogonal polynomials, or dummy variables, for instance. Hence, the most general equation for the regression $E(Y| X, Z)$ with a dichotomous X is

$$E(Y| X, Z) = g_0(Z) + g_1(Z) \cdot X, \tag{5}$$

with (an unknown) *intercept function* $g_0(Z)$ and (an unknown) *effect function* $g_1(Z)$. Interaction refers to all *constellations*, in which the effect function $g_1(Z)$ is not a constant, i.e., in which the effect of X on Y depends on the values of Z. Note that Equation (5) also applies if Z is a vector of random variables $Z_1, \ldots, Z_Q$.

Second, if (the treatment variable) X is not dichotomous, we can generalize Equation (2) to

$$E(Y|\, X, Z) = g_0(Z) + g_1(Z) \cdot I_{X=1} + \ldots + g_J(Z) \cdot I_{X=J}\,. \qquad (6)$$

where the indicator (dummy) variables $I_{X=1}$,…, $I_{X=J}$ indicate with 1 and 0 whether or not the observational unit is assigned to treatment j, j = 1, …, J. If the unit is assigned to the control condition, then $I_{X=1} = 0, \ldots, I_{X=J} = 0$.

The functions $g_0(Z)$, $g_1(Z)$, …, $g_m(Z)$ in Equation (6) can be *any* functions of Z. Oftentimes, they can be parameterized as polynomial functions or as indicator variables of the type explained in the previous paragraph for the treatment variable. Again, these generalizations are well-known and explained in sufficient detail by Aiken and West (1991) or Steyer (2003), for example.

Average Effects

It is straightforward to define the average effect of X on Y as the expected value of the effect function, i.e., $E[g_1(Z)]$. Since the values $g_1(z)$ are the conditional effects of X on Y, $E[g_1(Z)]$ is the *average of the conditional effects* of X on Y. It can be interpreted as the *main effect* of X. Note that this definition of a main effect is unique even if X and Z are correlated, and that it also applies to a multivariate vector Z consisting of several univariate variables $Z_1, \ldots, Z_q$. Although, if there is interaction, it is certainly more informative to consider the effect function and the conditional effects, researchers may often additionally be interested in whether or not the *average effect* of the treatment variable X (i.e., $E[g_1(Z)]$) differs from zero. If there are more than two treatment conditions, there will be a (possibly different) average effect $E[g_j(Z)]$, j = 1, …, J, for each of the J comparisons of a treatment j to the control.

Also note that the difference $E(Y|\, X = 1) - E(Y \mid X = 0)$ (called the *prima facie effect* by Holland, 1986) between the values of the regression $E(Y \mid X)$ is, in general, *not* the average effect of X. Only if the equation $E[g_1(Z) \mid X] = E[g_1(Z)]$ holds, the difference $E(Y \mid X = 1) - E(Y \mid X = 0)$ would be identical with $E[g_1(Z)]$. In other cases the difference $E(Y \mid X = 1) - E(Y \mid X = 0)$ will be the "effect" of X in the simple regression $E(Y \mid X)$, but this difference will not be the average

of the ($Z = z$)-conditional effects [the values of $g_1(Z)$] of X on Y (see also the Simpson paradox presented, for example, in Steyer, et al. 2000). However, it is this average of the conditional effects in which we are usually interested, not the "effect" $\alpha_1 = E(Y \mid X = 1) - E(Y \mid X = 0)$ of X in the simple regression

$$E(Y \mid X) = \alpha_0 + \alpha_1 X. \tag{7}$$

Some statististians advocate not to interprete main effects in orthogonal analysis of variance if there are interactions. Since main effects in orthogonal analysis of variance are also average effects, this view would also oppose analyzing average effects in regression models with interaction. However, all ANOVA programs compute tests for main effects, and therefore also for average effects, even if there are interactions. Also in the literature on non-orthogonal analysis of variance much effort has been spent to find tests of main effects and a number of different types of partitioning the sums of squares have been proposed and can be routinely computed in standard software packages such as SPSS, Systat, or SAS, for instance. Hence, striving for a test of the main or average effects even when there are interaction effects has been considered important by many statisticians. Finally, also in the literature on the analysis of causal effects (see also Steyer, 2005), much effort is spent on testing the average causal effect (see, e.g., Rosenbaum, 1984, 1995; Gelman & Meng, 2004). We will show below that $E[g_1(Z)]$, and, in general not $E(Y \mid X = 1) - E(Y \mid X = 0)$, is the average causal effect as defined in this tradition, provided some assumptions (specified below) hold.

Before we touch these details consider the following example. Wolchik et al., 1993) studied the effects of a psychological intervention with mothers (X) on the degree of *behavioral problems* of their children (Y). One of the covariates was a *pretest of the behavioral problems* (Z). Their interest in this example was not only to learn how the effect of the intervention

depends on the behavioral problems of the children assessed in the pretest, but also if there was an overall or average treatment effect.

As already mentioned above, it can be shown that $E[g_1(Z)]$ is equal to the *average causal effect* (see Rubin, 1974, 1978; Steyer, Nachtigall, Wüthrich-Martone, & Kraus, 2002 for details). The assumptions implying that

(a) the values $g_1(z)$ are the *conditional causal effects* and

(b) $E[g_1(Z)]$ is the *average causal effect* of a dichotomous treatment variable X with values 0 and 1,

can be easily stated, if we additionally introduce the observational-unit variable U, the value of which is the observational unit sampled in the following random experiment: Sample a unit u from a population (a set) of units, observe its covariate value z, assign or register the assignment of the unit to one of the treatment conditions x, and register the outcome y. (This is also what all previous regressions refer to.)

In this framework the assumptions implying (a) and (b) are:

$$E(Y|X,U,Z) = E(Y|X,U) \tag{8}$$

$$E[\, f_0(U)|X,Z] = E[\, f_0(U)|Z], \tag{9}$$

$$E[\, f_1(U)|X,Z] = E[\, f_1(U)|Z], \tag{10}$$

Where

$$E(Y|X,U) = f_0(U) + f_1(U)\cdot X, \tag{11}$$

is always true if X is dichotomous. In the last equation, a value $f_0(u)$ of the function $f_0(U)$ is the *expected outcome* $E(Y|X=0,U=u)$ of the unit u *under control* ($X=0$) and a value $f_1(u)$ of the

function $f_1(U)$ is the *individual causal effect* $E(Y \mid X = 1, U = u) - E(Y \mid X = 0, U = u)$ of X on Y for the unit u (see Steyer, 2005, for more details).

Equation (8) is a requirement for the covariate which is trivially true if Z is a deterministic function $f(U)$ of the observational-unit variable U, such as gender, race, or any other "organismic variable". Equation (8) will also be true, however, if Z is the sum of a deterministic function of the observational-unit variable U and a measurement error variable which has no additional effects on Y. An example in case is a fallible pretest Z assessed before the treatment.

Equations (9) and (10) state that, given Z, the expected values of the expected-outcome and individual-effects functions, $f_0(U)$ and $f_1(U)$, do not depend on X. This may sound as if X would not affect the individual effects. But this is *not* what the Equations (9) and (10) mean. Both functions, $f_0(U)$ and $f_1(U)$, represent (unknown) properties of the observational units. Hence, if treatment assignment does not depend on the units (given Z), it also will not depend on their properties represented by $f_0(U)$ and $f_1(U)$. This implies that the conditional expectations of these functions do not depend on X (given Z) (see Appendix A for the mathematical details).

To summarize: without assumptions other than X being dichotomous with values 0 and 1, the values $g_1(z)$ are the conditional "effects" of X on Y given $Z = z$ in the sense of a regression slope and the conditional mean difference $E(Y \mid X = 1, Z = z) - E(Y \mid X = 0, Z = z)$. Furthermore, $E[g_1(Z)]$ is the average of these conditional "effects" in the population. If assumptions (8) to (10) are added, a value $g_1(z)$ is also the conditional causal effect (i.e., the average $E[f_1(U) \mid Z = z]$ of the individual causal effects given $Z = z$), and $E[g_1(Z)]$ is also the average causal effect (i.e., the average $E[f_1(U)]$ of the individual causal effects $f_1(u)$ in the population). The difference $E(Y \mid X = 1) - E(Y \mid X = 0)$ (the prima facie effect) is not of interest in an analysis of treatment effects unless $E[g_1(Z) \mid X] = E[g_1(Z)]$ holds which implies $E(Y \mid X = 1) - E(Y \mid X = 0) = E[g_1(Z)]$. Randomization implies this equation, according to which the prima facie effect is equal to the average causal effect.

Null Hypothesis in Case of a Linear Effect Function

In the special case of a *linear effect function* $g_1(Z) = \gamma_{10} + \gamma_{11} Z$ the null hypothesis that the average effect is zero,

$$H_0: \quad E[g_1(Z)] = E(\gamma_{10} + \gamma_{11} Z) = \gamma_{10} + \gamma_{11} E(Z) = 0, \tag{12}$$

is equivalent to the hypothesis that the conditional effect of X on Y given the expected value of Z is 0. The null hypothesis (12) is a linear hypothesis, if $E(Z)$ is a known parameter.

If the intercept function is also linear, i.e., $g_0(Z) = \gamma_{00} + \gamma_{01} Z$, the following regression model results:

$$\begin{aligned} E(Y \mid X, Z) &= g_0(Z) + g_1(Z) \cdot X \\ &= (\gamma_{00} + \gamma_{01} Z) + (\gamma_{10} + \gamma_{11} Z) \cdot X \\ &= \gamma_{00} + \gamma_{01} Z + \gamma_{10} X + \gamma_{11} Z \cdot X. \end{aligned} \tag{13}$$

Three Procedures to Test the Null Hypothesis

Aiken and West (1991), for example, recommend to center Z about its sample mean and then test the null hypothesis H_0: $\gamma_{10} = 0$ instead of (12). Although the two hypotheses are equivalent indeed, we will show that this *GLM procedure* yields inflated α-errors, if the interaction effects are large. The reason is that centering Z about its sample mean – and not about its population mean – induces that the wrong hypothesis (i.e. a hypothesis that is not equivalent to (12)) is tested whenever the sample mean of Z differs from the population mean.

An alternative to the GLM procedure is to estimate the average effect $E[g_1(Z)]$ and its standard error by maximum likelihood and test the resulting Wald-statistic against zero. We will refer to this procedure as the *Wald test*. This procedure is asymptotically equivalent to a likelihood-ratio test.

If $\gamma_{11} = 0$, the regression model described by Equation (13) is a simple analysis of covariance (ANCOVA) model. One might be tempted to rely on the robustness of the ANCOVA and test the average treatment effect by omitting the interaction term from the model and testing $\gamma_{10} = 0$.

Indeed, several Monte Carlo simulations have shown, that the average effect of *X* on *Y* could be tested with this *ANCOVA procedure* even for nonparallel within-group regressions (i.e., for $\gamma_{11} \neq 0$, see e.g., Glass, Peckham, & Sanders, 1972; Hamilton, 1976) and nonnormal *Z* (Levy, 1980). Marked deviations from the nominal type I error have only been reported, if group sizes are unequal and γ_{11} differs markedly from zero. Rogosa (1980) criticized these simulation studies because of two reasons: (1) The definition of an overall (main) treatment effect in the presence of an interaction of *X* and *Z* was unclear. (2) The bias of the ANCOVA estimate of the average effect of *X* on *Y* depends not only on the difference in group sizes but also on the difference of the within-group variance of *Z*, i.e., a smaller group size would be compensated by a greater within-group variance. He showed analytically that the ANCOVA procedure is only seriously biased if all of the following conditions hold: (a) there is interaction of *Z* and *X*, (b) the group means of *Z* are not the same and (c) the products of the group size and the corrected sum of squares of *Z* are not the same across groups.

Plan for the Simulation Studies

In the present study we investigate via Monte Carlo simulations whether or not the three different procedures for testing the average effect of *X* on *Y* are appropriate if *Z* is a stochastic regressor (with varying means between samples). The different procedures are: the Wald test, the ANCOVA and the GLM procedure. In the first simulation study, both the intercept function and the effect function are linear. In a second simulation study we investigate, if the Wald procedure can be applied for a quadratic effect function $g_1(Z) = \gamma_{10} + \gamma_{11} Z + \gamma_{12} Z^2$ as well. Finally, we discuss how to extend our results to other models and identify open questions.

Method

Data Generation

We generated different samples of (varying) size N by repeating the following procedure N times: First, draw a standard-normally distributed random number, the covariate Z_i, i.e., $Z_i \sim N(0, 1)$. In order to generate a dependency of the dichotomous (treatment variable) X (with values 0 and 1) on Z, X was randomly drawn from a Bernoulli distribution, where the probability of $X = 1$was calculated by the logistic function, i.e.:

$$P\left(X_i = 1 \middle| Z_i = z_i\right) = \frac{1}{1 + e^{-(\lambda + \theta \cdot z_i)}}. \tag{14}$$

A value of the outcome variable Y_i was then computed by inserting Z_i into the functions $g_0(Z_i)$ and $g_1(Z_i)$, inserting X_i and adding a normally distributed error component with expectation zero and variance $\sigma^2 = 4$, i.e.:

$$Y_i = g_0(Z_i) + g_1(Z_i) \cdot X_i + \varepsilon_i . \tag{15}$$

For each parameter constellation (see Table 1) *Nsim* = 10,000 replications were produced.

Insert Table 1 about here

In simulation studies 1a and b both the intercept function $g_0(Z)$ and the effect function $g_1(Z)$ were linear, whereas in studies 2a and b, a quadratic effect function was specified (see Table 1). The parameters were chosen such that the average effect is zero, i.e., $E[g_1(Z)] = 0$. In all parameter constellations chosen, the expected value $E[g_1(Z)]$ is a function of the parameters γ_{10}, γ_{11}, γ_{12} and the first and second moment of the distribution of Z, i.e.:

$$E[g_1(Z)] = E(\gamma_{10} + \gamma_{11} Z + \gamma_{12} Z^2) = \gamma_{10} + \gamma_{11} E(Z) + \gamma_{12} E(Z^2). \quad (16)$$

Since Z has a standard normal distribution, Z^2 is χ^2-distributed with $df = 1$. Consequently, $E(Z) = 0$ and $E(Z^2) = 1$.

In studies 1a and 2a, $\lambda = 0$ and $\theta = 0.5$, which results on average in equal group size (for $X = 0$ and $X = 1$), equal variances of the covariate Z within treatment groups [$Var(Z \mid X = 0) = Var(Z \mid X = 1)$] and a moderate correlation of .236 between the covariate and the dichotomous treatment variable in the population. In studies 1b and 2b, $\lambda = 0.5$ and $\theta = 0.5$, which yields on average a greater proportion for an assignment to the treatment group [$P(X = 1) = .616$] and also a negligible difference between the two within-group variances of the covariate. In study 1b and 2b the true correlation between the covariate and the dichotomous treatment variable is .230, which slightly differs from study 1a and 2a.

Within each study we varied the size of the (linear or quadratic) interaction of the treatment with the covariate (see Note a of Table 1) as well as the sample size (see Note b in Table 1).

Tests of the Null Hypothesis

In studies 1a and 1b three ways of testing the null hypothesis ($E[g_1(Z)] = 0$) were pursued for each generated data set: ANCOVA, the GLM procedure, and the Wald test. In studies 2a and 2b only the Wald test was investigated. The proportion of significant outcomes at $\alpha = .05$ and the (mean of the) estimates of the average effect were recorded.

- An ANCOVA was performed, by testing β_2 for significance in the linear model $Y = \beta_0 + \beta_1 Z + \beta_2 X + \varepsilon$.
- The GLM procedure means centering the covariate at its sample mean and testing $\gamma_{10} = 0$ in a model including a linear interaction term: $Y = \gamma_{00} + \gamma_{01} Z + \gamma_{10} X + \gamma_{11} Z X + \varepsilon$.

- The Wald test was computed with a program for structural equation modeling (LISREL). In this procedure not only the regression parameters were estimated, but also $E(Z)$ (and $E(Z^2)$ in study 2a and 2b), and the hypothesis was specified as a nonlinear constraint between these parameter estimates: The estimator for the average effects is now $\hat{\gamma}_{10} + \hat{\gamma}_{11}\,\overline{Z}$ for a linear effect function $g_1(Z)$ (studies 1a and b) and $\hat{\gamma}_{10} + \hat{\gamma}_{11}\,\overline{Z} + \hat{\gamma}_{12}\,\overline{Z^2}$ for a quadratic effect function $g_1(Z)$ (studies 2a and b) where $\overline{Z}$ and $\overline{Z^2}$ denote the sample means of Z and Z^2, respectively. These estimators were divided by their standard error (computed via the delta method) and the resulting t-value was tested against a Student's t-distribution. The degrees of freedom were determined by substracting the number of estimated parameters [i.e., 8, if $g_0(Z)$ and $g_1(Z)$ are linear functions, 12 if $g_0(Z)$ and $g_1(Z)$ are quadratic functions] from the number of observations. This method will be referred to as the *Wald test*. (The LISREL input file is given in the Appendix B.)

Results

Study 1a

In Study 1a, there were a linear effect function and equal group sizes in the population. All three procedures yielded means of the estimates of the average treatment effect around zero which is in full agreement with the null hypothesis. However, large differences were observed with respect to the percentage of significant tests at the .05-level.

ANCOVA

The results for the ANCOVA are presented in Table 2. According to this table the ANCOVA yielded slightly inflated α-errors if a large interaction concured with a small sample size.

Insert Table 2 about here

GLM procedure

The results for the GLM procedure are shown in Table 3. The α-level was correct only if the interaction parameter γ_{11} was equal or close to zero. Regardless of sample size, the more γ_{11} deviated from 0, the higher the proportion of significant tests at the .05-level. For $\gamma_{11} = 10$ or $\gamma_{11} = -10$, the proportion of significant tests was almost ten times inflated comparised to the nominal significance level.

Insert Table 3 about here

Wald test

The results for the Wald test are shown in Table 4. The Wald test yielded exact and satisfactory results. Even for the extreme cases $\gamma_{11} = 10$ and $\gamma_{11} = -10$ and sample sizes as small as 25 the proportions of significant tests did not exceed the nominal .05 level.

Insert Table 4 about here

Our conclusion from these results is that the GLM procedure fails in testing the null hypothesis [see Equation (12)] if *Z* is stochastic. The results for the Wald test, however, are most promising. Hence, it seems worthwhile to explore if similar results obtain in different constellations.

Study 1b

In Study 1b, we used a linear effect function again. Now, however, group sizes were *unequal* in the population. The GLM and the Wald test yielded unbiased estimates of the average treatment effect under all conditions. The ANCOVA estimates, however, were biased (see Table 5). The magnitude and the direction of the bias depended on the magnitude and direction of the

interaction term. The higher the interaction, the larger was the bias. A positive interaction term yields to a negative bias and vice versa.

Insert Table 5 about here

Marked differences between the three procedures can be observed with respect to the percentage of significant tests. They will be described in the following paragraphs.

ANCOVA

The results of the significance test of the average effect via ANCOVA are presented in Table 6. The ANCOVA yielded extremely inflated α-errors if a large interaction concurs with a large sample size. The inflation of the alpha error was less dramatic if the sample size was small. In the absence of an interaction the ANCOVA yielded valid tests of the average effect regardless of the sample size.

Insert Table 6 about here

GLM procedure

The results for the GLM procedure are shown in Table 7. They differ only negligibly from Study 1a. The α-level was correct only if the interaction parameter γ_{11} was equal or close to zero. Obviously, the more γ_{11} deviated from 0, the higher the proportion of significant tests at the .05 level.

Insert Table 7 about here

Wald test

The results for the Wald test are shown in Table 8. The Wald test again yielded exact and satisfactory results. Even for the extreme cases $\gamma_{11} = 10$ and $\gamma_{11} = -10$ the proportions of significant tests did not exceed the nominal .05-level.

Insert Table 8 about here

Simulation Studies 2a and b

These studies explored whether or not the Wald test can be extended to quadratic effect functions $g_1(Z)$. The following intercept and effect functions were studied:

$$g_0(Z) = \gamma_{00} + \gamma_{01} Z + \gamma_{02} Z^2, \quad (17)$$

with $\gamma_{00} = \gamma_{02} = 0$ and $\gamma_{01} = 1$,

$$g_1(Z) = \gamma_{10} + \gamma_{11} Z + \gamma_{12} Z^2, \quad (18)$$

with $\gamma_{11} = 0$, $\gamma_{10} = -\gamma_{12}$, and γ_{12} varying again between −10 and +10 (see Table 1). All the other parameters were the same as in Study 1a. Note that $E[g_1(Z)] = -\gamma_{12} + 0\ Z + \gamma_{12} Z^2 = 0$, since Z is standard normally distributed in this simulation [implying $E(Z^2) = 1$] .

The LISREL model was specified in such a way that all regression coefficients and expected values of the covariate and its square (i.e., the first and second moment of the distribution of Z) were estimated (see Appendix B for the LISREL input file). Again, the nonlinear hypothesis

$$E[g_1(Z)] = E(\gamma_{10} + \gamma_{11} Z + \gamma_{12} Z^2) = \gamma_{10} + \gamma_{11} E(Z) + \gamma_{12} E(Z^2) = 0$$

was tested via the Wald test described above.

Inspecting Table 9 and Table 10 reveals that the Wald test again yielded valid results for medium and large sample sizes ($N \geq 250$). However, for the extreme cases with small sample

sizes of $N = 25$ or $N = 50$ and $\gamma_{12} = 10$ or $\gamma_{12} = -10$, the proportion of significant *t*-tests exceeded the fixed .05 level. Instead of 5% there were around 8 or 9% significant *t*-values.

Insert Table 9 and Table 10 about here

Summarizing these results, we can recommend the Wald procedure also for quadratic effect functions for sample sizes above $N = 100$. Nevertheless, there seem to be somewhat inflated type-I errors for relatively small sample sizes and extreme effect functions so that better procedures might be developed for these cases.

Discussion

In this paper we argued that testing the hypothesis that the expectation of the effect function is zero ($E[g_1(Z)] = 0$) plays an important role in analyzing the effects of a variable X on an outcome variable Y which are modified by a covariate Z, because testing this *average effect* means testing the *main effect* of X in the presence of an interaction effect. Although it is certainly true that the conditional effects $g_1(z)$ of X on Y given $Z = z$ are much more informative, we believe that the average effect is still a useful additional summary information about the effect of X. The results presented apply to both meanings of an average effect: the average effect in the sense of the average of the conditional effects $g_1(z)$ and the average causal effect in terms of the Neyman-Rubin theory of individual and average causal treatment effects provided that assumptions (8) to (10) hold.

We have shown that, if Z is a *stochastic regressor* and the treatment effect is modified by the covariate, the GLM procedure of testing the average effect yields inflated Type I-errors, the errors increasing the larger the interaction effect becomes. Surprisingly, in the case of equal group sizes, the ANCOVA yielded only slightly inflated α-errors; despite the fact that the interaction is not represented in the ANCOVA model. Only one of the reported simulation studies

about the robustness of the ANCOVA explicitly[1] dealt with a stochastic covariate *Z* (Hamilton, 1977). Hamilton's results also indicate that the ANCOVA does not exceed the nominal type I error rate, unless unequal group sizes are combined with $\gamma_{11} \neq 0$. With regard to the critique of Rogosa (1980), it seems plausible that the ANCOVA outperforms the GLM procedure, unless an interaction coincides with unequal group means of the covariate and the product of the squared group sizes with the respective within group variance differs across the groups. Especially, in case of a randomized experiment with an interaction effect, ANCOVA is likely to be more adequate for testing the average treatment effect than the GLM procedure, because there will be no systematic difference in group means and variances of the covariate (whether the group size differs or not). All these virtues of the ANCOVA are confirmed by our simulation results.

However, if group mean differences of the covariate coincide with unequal group sizes (and essentially equal group variances of the covariate) and interaction like in study 1b, ANCOVA yields not only inflated α-errors, but also biased estimates of the average effect. However, even for equal group sizes and equal variances of the covariate within the groups (like in study 1a), we would not recommend using ANCOVA, whenever there is an interaction of the treatment with the covariate. Even in this case, the ANCOVA yields inflated α-errors, although the deviation from the nominal significance level was much less dramatic than for the GLM procedure. This should be no surprise, because Rogosa (1980) has shown that the *F*-statistic of the ANCOVA follows no central *F*-distribution in case of an interaction of a fixed covariate with the treatment. Additionally, even if the α-errors are only slightly inflated the ANCOVA model is a mis-specification in case of interaction and the interaction itself usually is of great substantive interest and should not be completely ignored (which is true for the ANCOVA procedure).

We have shown that the Wald test with nonlinear constraints (as implemented in LISREL and Mplus) seems to be the only method that protects the type I error rate in case of a stochastic *Z*

[1] Hamilton (1976), Levy (1980), and Wu (1984) did not state explicitly, whether or not *Z* was random or if the values of *Z* were held fixed across the simulations as in the study of Glass et al. (1972).

not only for a restricted set of parameter settings. Hence, it seems to be the only advisable method in case of a stochastic covariate *Z*, whereas the GLM procedure is not generally appropriate in case of a stochastic covariate *Z*.

How to decide whether or not *Z* is stochastic? The critical question is: If you would select a sample of the same size again: Would your sample mean of *Z* be exactly the same as in your first sample? If not, you have a sampling scheme that yields a stochastic *Z*. While the distinction between fixed and stochastic regressors does not matter for most purposes (e.g., parameter estimation and tests for the regression of *Y* on *X* and *Z*), the present study showed that it *does* matter for testing the average effect hypothesis.

How to generalize to more than two treatment groups? According to Equation (6) there will be J effect functions $g_j(Z)$ for $J + 1$ treatment conditions. Hence, the overall hypothesis of no treatment effects will be: H_0: $E[g_j(Z)] = 0$, for $j = 1, \ldots, J$. This would necessitate J nonlinear constraints in LISREL and a χ^2-difference test.

What are the open questions? First of all, more simulation studies seem warranted to study the behavior of the χ^2-difference test for more than two treatment conditions and different sample sizes. Second, more studies are necessary also with effect functions other than linear and quadratic. Third, a generalization to latent covariates seems highly desirable, because manifest covariates may not always remove the bias introduced by latent covariates.

References

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

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.

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

Gelman, A. & Meng, X.-L. (2004). *Applied Bayesian modeling and causal inference from incomplete data perspectives*. Chichester: Wiley.

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,* 237-288.

Gosslee, D. G. & Lucas, H. L. (1965). Analysis of variance of disproportionate data when interaction is present. *Biometrics, 21,* 115-133.

Hamilton, B. L. (1976). A Monte-Carlo test of the robustness of parametric and nonparametric analysis of covariance against unequal regression slopes. *Journal of the American Statistical Association, 71,* 864-869.

Hamilton, B. L. (1977). An empirical investigation of the effects of heterogeneous regression slopes in analysis of covariance. *Educational and Psychological Measurement, 37,* 701-712.

Holland, P. (1986). Statistics and causal inference (with comments). *Journal of the American Statistical Association, 81,* 945-970.

Levy, K. J. (1980). A Monte Carlo study of analysis of covariance under violations of the assumptions of normality and equal regression slopes. *Educational and Psychological Measurement, 40,* 835-840.

Moosbrugger, H. (1981). Zur differentiellen Validität bei nichtlinearen Test-Kriterium-Zusammenhängen. *Zeitschrift für Differentielle und Diagnostische Psychologie, 2,* 219-274.

Rogosa, D. (1980). Comparing nonparallel regression lines. *Psychological Bulletin, 88,* 307-321.

Rosenbaum, P. R. (1984). From association to causation in observational studies: The role of tests of strongly ignorable treatment assignment. *Journal of the American Statistical Association, 79,* 41-48.

Rosenbaum, P. R. (1995). *Observational studies*. New York: Springer.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology, 66,* 688-701.

Rubin, D. B. (1978). Bayesian inference for causal effects: The role of randomization. *Annals of Statistics, 6,* 34-58.

Saunders, D. R. (1956). Moderator variables in prediction. *Educational and Psychological Measurement, 16,* 209-222.

Steyer, R. (2005). Analyzing individual and average causal effects via structural equation models. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences, 1,* 39-54.

Steyer, R., Nachtigall, C., Wüthrich-Martone, O., & Kraus, K. (2002). Causal regression models III: Covariates, conditional, and unconditional average causal effects. *Methods of Psychological Research Online, 7,* 41-68.

Wolchik, S. A., West, S. G., Westover, S., Sandler, I. N., Martin, A., Lustig, J. et al. (1993). The children of divorce parenting intervention: Outcome evaluation of an empirically based program. *American Journal of Community Psychology, 21,* 293-331.

Wu, Y.-W. B. (1984). The effects of heterogeneous regression slopes on the robustness of two test statistics in the analysis of covariance. *Educational and Psychological Measurement, 44,* 647-663.

Table 1

Data generation and parameter settings in the simulation studies

General Model	Z	$P\left(X_i = 1 \middle\| Z_i = z_i\right)$		$g_0(Z_i)$			$g_1(Z_i)$			ε
	$N(0, 1)$	$\left(1 + e^{-(\lambda + \theta \cdot z_i)}\right)^{-1}$		$\gamma_{00} + \gamma_{01} Z_i + \gamma_{02} Z_i^2$			$\gamma_{10} + \gamma_{11} Z_i + \gamma_{12} Z_i^2$			$N(0, \sigma^2)$
Parameters		λ	θ	γ_{00}	γ_{01}	γ_{02}	γ_{10}	γ_{11}	γ_{12}	σ^2
Study 1a		0	.5	0	1	0	0	various[a]	0	4
Study 1b		.5	.5	0	1	0	0	various[a]	0	4
Study2a		0	.5	0	1	0	$-\gamma_{12}$	0	various[a]	4
Study2b		.5	.5	0	1	0	$-\gamma_{12}$	0	various[a]	4

Note.
a: various means γ_1 and γ_{12} were set to -10, -5, -2.5, -1, -0.5, -0.25, 0, 0.25, 0.5, 1, 2.5, 5 and 10, respectively.
b: In each study the following sample sizes were used for all parameter constellations: 25, 50, 100, 250, 500, 1000.

Table 2

Percent of significant treatment effects at $\alpha = .05$ *for* ANCOVA *with equal group size in the population*

	Sample size					
γ_{11}	25	50	100	250	500	1000
-10	.086	.069	.067	.065	.065	.064
-5	.074	.069	.064	.059	.055	.059
-2.5	.059	.059	.060	.053	.050	.058
-1	.053	.051	.051	.048	.050	.050
-0.5	.054	.052	.049	.048	.047	.048
-0.25	.049	.049	.048	.048	.051	.049
0	.046	.053	.051	.050	.052	.051
0.25	.049	.050	.052	.054	.049	.053
0.5	.048	.049	.054	.049	.055	.051
1	.048	.054	.052	.048	.051	.052
2.5	.058	.054	.058	.055	.056	.054
5	.075	.069	.063	.059	.056	.058
10	.084	.070	.067	.063	.059	.060

Note. Based on *Nsim* = 10,000 samples.

Table 3

Percent of significant treatment effects at $\alpha = .05$ *for the GLM procedure with equal group size in the population*

	Sample size					
γ_{11}	25	50	100	250	500	1000
-10	.409	.438	.450	.448	.456	.460
-5	.190	.199	.211	.215	.206	.211
-2.5	.089	.094	.097	.091	.092	.099
-1	.058	.058	.056	.053	.055	.055
-0.5	.056	.056	.051	.050	.050	.049
-0.25	.048	.050	.049	.048	.051	.048
0	.048	.052	.050	.050	.053	.051
0.25	.049	.052	.052	.055	.050	.053
0.5	.049	.052	.057	.050	.055	.051
1	.054	.059	.057	.053	.057	.059
2.5	.086	.090	.098	.096	.096	.093
5	.184	.210	.212	.211	.210	.211
10	.410	.440	.442	.461	.449	.460

Note. Based on *Nsim* = 10,000 samples.

Table 4

Percent of significant treatment effects at $\alpha = .05$ *for the Wald test with equal group size in the population*

	Sample size					
γ_{11}	25	50	100	250	500	1000
-10	.046	.048	.050	.047	.055	.055
-5	.049	.051	.053	.052	.047	.050
-2.5	.051	.054	.056	.050	.049	.054
-1	.056	.055	.050	.047	.049	.049
-0.5	.058	.057	.051	.049	.048	.049
-0.25	.052	.054	.051	.048	.052	.048
0	.052	.056	.053	.050	.054	.052
0.25	.052	.055	.054	.055	.050	.053
0.5	.051	.053	.057	.049	.054	.051
1	.052	.056	.053	.047	.051	.054
2.5	.051	.050	.055	.050	.052	.050
5	.049	.051	.051	.051	.046	.049
10	.045	.047	.047	.051	.050	.052

Note. Based on *Nsim* = 10,000 samples.

Table 5

Mean estimates of the average effect at $\alpha = .05$ *for the ANCOVA with unequal group sizes in the population*

	Sample size					
γ_{11}	25	50	100	250	500	1000
-10	1.109	1.138	1.127	1.119	1.103	1.106
-5	0.559	0.575	0.559	0.548	0.558	0.551
-2.5	0.300	0.278	0.276	0.278	0.275	0.278
-1	0.113	0.111	0.111	0.107	0.109	0.108
-0.5	0.062	0.049	0.051	0.055	0.055	0.054
-0.25	0.038	0.024	0.022	0.027	0.025	0.030
0	0.001	-0.007	0.003	0.006	0.001	-0.001
0.25	-0.052	-0.029	-0.025	-0.030	-0.027	-0.028
0.5	-0.059	-0.051	-0.056	-0.058	-0.053	-0.055
1	-0.127	-0.111	-0.111	-0.107	-0.110	-0.109
2.5	-0.275	-0.275	-0.271	-0.272	-0.273	-0.276
5	-0.580	-0.574	-0.564	-0.554	-0.550	-0.546
10	-1.212	-1.132	-1.085	-1.091	-1.110	-1.105

Note. Based on *Nsim* = 10,000 samples.

Table 6

Percent of significant treatment effect at $\alpha = .05$ *for ANCOVA with unequal group sizes in the population*

	Sample size					
γ_{11}	25	50	100	250	500	1000
-10	.139	.157	.214	.375	.605	.863
-5	.105	.125	.155	.281	.476	.741
-2.5	.074	.083	.091	.153	.240	.429
-1	.054	.058	.058	.067	.089	.122
-0.5	.050	.051	.050	.055	.061	.073
-0.25	.051	.047	.052	.053	.050	.058
0	.048	.046	.051	.051	.051	.049
0.25	.046	.050	.050	.053	.057	.055
0.5	.053	.052	.054	.054	.059	.074
1	.051	.053	.060	.071	.084	.130
2.5	.073	.079	.092	.146	.244	.430
5	.107	.124	.163	.281	.465	.726
10	.143	.152	.203	.365	.612	.869

Note. Based on *Nsim* = 10,000 samples.

Table 7

Percent of significant treatment effects at $\alpha = .05$ for GLM with unequal group sizes in the population

	Sample size					
γ_{11}	25	50	100	250	500	1000
-10	.399	.429	.436	.434	.445	.438
-5	.173	.193	.199	.205	.204	.207
-2.5	.082	.087	.089	.090	.089	.089
-1	.054	.057	.055	.058	.052	.055
-0.5	.051	.050	.049	.053	.053	.054
-0.25	.049	.049	.051	.051	.051	.052
0	.049	.047	.050	.051	.051	.048
0.25	.047	.051	.050	.053	.053	.051
0.5	.052	.053	.054	.048	.052	.058
1	.050	.052	.056	.055	.056	.055
2.5	.086	.085	.090	.094	.097	.093
5	.176	.198	.200	.208	.211	.204
10	.399	.431	.432	.447	.434	.427

Note. Based on *Nsim* = 10,000 samples.

Table 8

Percent of significant treatment effects $\alpha = .05$ *for Wald with unequal group sizes in the population*

	Sample size					
γ_{11}	25	50	100	250	500	1000
-10	.040	.048	.049	.050	.052	.048
-5	.044	.052	.051	.052	.051	.050
-2.5	.051	.053	.052	.048	.050	.048
-1	.054	.054	.052	.051	.045	.049
-0.5	.055	.051	.049	.052	.051	.052
-0.25	.053	.051	.052	.052	.051	.053
0	.052	.052	.053	.053	.052	.049
0.25	.049	.053	.051	.053	.053	.051
0.5	.054	.055	.055	.048	.051	.057
1	.050	.051	.053	.050	.051	.051
2.5	.054	.052	.052	.051	.055	.051
5	.045	.053	.047	.052	.052	.050
10	.041	.045	.048	.049	.047	.051

Note. Based on *Nsim* = 10,000 samples.

Table 9

Percent of significant treatment effects at $\alpha = .05$ *for Wald with equal group sizein the population and a quadratic effect function*

	Sample size					
γ_{12}	25	50	100	250	500	1000
-10	.080	.073	.062	.054	.055	.051
-5	.061	.060	.058	.055	.051	.053
-2.5	.050	.056	.054	.047	.050	.053
-1	.047	.054	.053	.054	.050	.053
-0.5	.044	.054	.053	.051	.051	.045
-0.25	.044	.057	.051	.052	.053	.050
0	.042	.051	.060	.055	.052	.045
0.25	.043	.054	.057	.049	.052	.048
0.5	.049	.052	.052	.051	.053	.049
1	.044	.053	.054	.050	.047	.050
2.5	.048	.053	.053	.050	.049	.054
5	.060	.065	.054	.052	.052	.054
10	.080	.069	.060	.053	.050	.046

Note. Based on *Nsim* = 10,000 samples.

Table 10

Percent of significant treatment effect for the Wald test with unequal group size in the population and a quadratic effect function

	Sample size					
γ_{12}	25	50	100	250	500	1000
-10	.074	.070	.059	.052	.052	.052
-5	.060	.064	.058	.054	.055	.053
-2.5	.043	.051	.053	.050	.048	.048
-1	.044	.053	.051	.048	.048	.049
-0.5	.040	.053	.051	.054	.049	.054
-0.25	.041	.056	.053	.053	.050	.048
0	.042	.052	.054	.051	.051	.052
0.25	.044	.049	.053	.054	.051	.050
0.5	.046	.052	.054	.051	.050	.049
1	.045	.053	.054	.051	.050	.050
2.5	.045	.055	.054	.053	.053	.054
5	.061	.056	.054	.055	.050	.048
10	.074	.067	.064	.059	.051	.051

Note. Based on Nsim = 10,000 samples.

Appendix A: Proof

$$
\begin{aligned}
E(Y|X,Z) &= E[E(Y|X,U,Z)|X,Z] \quad \text{(always true)} \\
&= E[\, f_0(U) + f_1(U)\cdot X \,|\, X,Z] \quad \text{[see Eqs. (8), (11)]} \\
&= E[\, f_0(U)|X,Z] + E[f_1(U)|X,Z]\cdot X \\
&= E[\, f_0(U)|Z] + E[f_1(U)|\,Z]\cdot X \quad \text{[see Eqs. (9), (10)]} \\
&= g_0(Z) + g_1(Z)\cdot X,
\end{aligned}
$$

defining $g_0(Z) := E[f_0(U)|Z]$ and $g_1(Z) := E[f_1(U)|Z]$. Finally, the fact that $E\,[g_1(Z)]$ is the average causal effect follows from $E[g_1(Z)] = E(E[f_1(U)|Z]) = E[f_1(U)]$.

Appendix B: Lisrel Input Files

Linear model: estimate, standard error and *t*-value for the

average effect found under PAR(1)

```
DA NI=4
RA FI=TRY.DAT
LA
Y  X  Z  ZX
MO NY=1 NX=3 NE=1 NK=3 GA=FU,FR LY=ID LX=ID TD=FI TE=FI c
   TY=FI TX=FI KA=FR AL=FR AP=1  FIXED-X
CO PAR(1) = KA(2)*GA(1,3) + GA(1,1)
OU WP ND=4
```

Quadratic model: estimate, standard error and *t*-value for the

average effect found under PAR(1)

```
DA NI=6
RA FI=TRY.DAT
LA
Y X Z ZZ ZX ZZX
MO NY=1 NX=5 NE=1 NK=5 GA=FU,FR TY=FI TX=FI LY=ID  c
   LX=ID TD=FI TE=FI KA=FR AL=FR AP=1  FIXED-X
CO PAR(1) = GA(1,1) + KA(2)*GA(1,4) + KA(3)*GA(1,5)
OU WP ND=4
```