Published in final edited form as:

Behav Genet. 2014 March ; 44(2): 165–181. doi:10.1007/s10519-014-9642-1.

Testing for Measured Gene-Environment Interaction: Problems with the use of Cross-Product Terms and a Regression Model Reparameterization Solution

Fazil Aliev,

Department of Psychiatry, Virginia Institute for Psychiatric and Behavioral Genetics, Virginia Commonwealth University, 800 E. Leigh St., PO Box 980126, Richmond, VA 23298-0126, USA;
Department of Statistics, Ankara University, Ankara 06100, Turkey

Shawn J. Latendresse,

Department of Psychiatry, Virginia Institute for Psychiatric and Behavioral Genetics, Virginia Commonwealth University, 800 E. Leigh St., PO Box 980126, Richmond, VA 23298-0126, USA

Silviu-Alin Bacanu,

Department of Psychiatry, Virginia Institute for Psychiatric and Behavioral Genetics, Virginia Commonwealth University, 800 E. Leigh St., PO Box 980126, Richmond, VA 23298-0126, USA

Michael C. Neale, and

Department of Psychiatry, Virginia Institute for Psychiatric and Behavioral Genetics, Virginia Commonwealth University, 800 E. Leigh St., PO Box 980126, Richmond, VA 23298-0126, USA

Danielle M. Dick

Department of Psychiatry, Virginia Institute for Psychiatric and Behavioral Genetics, Virginia Commonwealth University, 800 E. Leigh St., PO Box 980126, Richmond, VA 23298-0126, USA

Abstract

The study of gene-environment interaction ($G \times E$) has garnered widespread attention. The most common way to assess interaction effects is in a regression model with a $G \times E$ interaction term that is a product of the values specified for the genotypic (G) and environmental (E) variables. In this paper we discuss the circumstances under which interaction can be modeled as a product term and cases in which use of a product term is inappropriate and may lead to erroneous conclusions about the presence and nature of interaction effects. In the case of a binary coded genetic variant (as used in dominant and recessive models, or where the minor allele occurs so infrequently that it is not observed in the homozygous state), the regression coefficient corresponding to a significant interaction term reflects a slope difference between the two genotype categories and appropriately characterizes the statistical interaction between the genetic and environmental variables. However, when using a three-category polymorphic genotype, as is commonly done when modeling an additive effect, both false positive and false negative results can occur, and the nature of the interaction can be misrepresented. We present a reparameterized regression equation that

accurately captures interaction effects without the constraints imposed by modeling interactions using a single cross-product term. In addition, we provide a series of recommendations for making conclusions about the presence of meaningful $G \times E$ interactions, which take into account the nature of the observed interactions and whether they map onto sensible genotypic models.

Keywords

Genes; Environment; Modeling; Regression; Cross-product

Introduction

Interest in testing for gene-environment interactions in relation to common, complex disorders is widespread and growing. A PubMed search using the term “gene-environment interaction” produces more than 1,300 papers, more than half of which were published within the last five years. Across multiple disciplines, gene-environment interaction is an active area of research, with large literatures testing candidate gene-environment interactions (e.g., Caspi et al. (2003); Bakermans-Kranenburg and van Ijzendoorn (2006); Petersen et al. (2012); Duncan and Keller (2011), Widaman et al. (2012), Keller (2014)), and growing interest in testing for gene-environment interaction in the context of genome-wide association studies (GWAS) (Khouri and Wacholder (2009); Murcray et al. (2009)). Characterizing genotype by environment interactions presents many challenges, including questions about how best to measure the environment, the relative importance of the timing and duration of the environmental exposure, and the potential for increased problems with multiple testing when large-scale genotypic information is crossed with many potential environmental variables of interest. However, one area that has received little scrutiny is the statistical methods that are routinely applied to test for gene-environment interaction. The most common way to assess interaction effects is by using a cross-product term representing the product of the genetic (G) and environmental (E) variables, within the context of a multiple regression model. A significant nonzero coefficient for the interaction term is interpreted as evidence of statistical interaction. There are two standard ways of coding a genetic polymorphism for this purpose: either as a binary variable (0, 1) or a three category variable, coded as 0, 1, 2. A binary coding implies a particular genetic model, for example, with 0 corresponding to 0 or 1 copies of a selected risk allele and 1 corresponding to two copies of that risk allele (recessive model). Similarly, binary coding can be employed to model genetic dominance, with 0 indicating 0 copies of the risk allele and 1 indicating one or two copies of the risk allele. Often, however, we do not know the true functional model associated with a given marker of interest. Accordingly, a common practice is to conduct initial tests assuming an additive genetic model, in which the gene is coded 0, 1, 2 to indicate the number of copies of a particular reference allele carried by an individual. We demonstrate that modeling gene-environment interaction ($G \times E$) by way of a cross-product term frequently yields misleading results that can generate erroneous conclusions about the presence and magnitude of interaction effects when genotype is coded 0, 1, 2. This result has potentially important implications for the interpretation of gene-environment interaction effects derived using this method, including many of those in previously published articles. In this paper, we illustrate the problem and provide an alternative parameterization which

can be used to estimate regression lines within each genotypic group that more accurately capture the nature of the interaction. Although the same problem exists when the dependent variable is dichotomous, we will explain our methods using continuous linear regression.

Methods and results

Delineation of the problem

The phenotype, or dependent variable, is hereafter denoted by P , with G and E used to represent the genotypic and environmental variables, respectively. The traditional regression equation with an interaction term has the form:

$$P = \beta_0 + \beta_1 G + \beta_2 E + \beta_3 G \times E, \quad (1)$$

which determines four unknown β coefficients (to simplify formulas we omit covariates in the regression model other than the gene and environment variables, and do not use a subscript or error term specific to an individual in the sample). The interaction effect is explained by the slopes resulting from the above equation for different levels of the genotypic variable. The phenotype's regression on the environment at each fixed level of the genotype can be modeled as: $P = \delta_{i0} + \delta_{i1}E$, where i is the genotypic level, δ_{i0} is intercept term for the level, δ_{i1} is the slope coefficient (i takes the value of the genotype, e.g., 0, 1, 2). Figure 1 shows an example of a three level genetic variant and the regression lines for three genotypes in three-dimensional space and corresponding projections in two-dimensional space.

From here forward we will use two-dimensional plots of projections of regression lines onto the plane genotype = 0 instead of three-dimensional plots.

Let us first assume that the genotype has two categories, 0 and 1, as follows:

$$\begin{aligned} G = 0 & \text{ has the relation } P|(G=0) = \delta_{00} + \delta_{01}E \\ G = 1 & \text{ has the relation } P|(G=1) = \delta_{10} + \delta_{11}E. \end{aligned}$$

We determine δ_{ij} coefficients using the additive genetic $G \times E$ regression model for all individuals from Eq. (1). When the β coefficients of Eq. (1) are available, for the subset of individuals with $G = 0$, plugging the value 0 for the G variable into Eq. (1) eliminates the G and $G \times E$ terms and we have:

$$P|(G=0) = \beta_0 + \beta_2 E = \delta_{00} + \delta_{01}E.$$

Accordingly, $\delta_{00} = \beta_0$ and $\delta_{01} = \beta_2$. Substituting these known δ coefficients for the β coefficients in Eq. (1), for the case where $G = 1$ we can write:

$$\begin{aligned} P|(G=1) &= \delta_{00} + \beta_1 \cdot 1 + \delta_{01}E + \beta_3 E \\ &= (\delta_{00} + \beta_1) + (\delta_{01} + \beta_3)E = \delta_{10} + \delta_{11}E. \end{aligned}$$

Hence: $\delta_{10} = \delta_{00} + \beta_1$; $\delta_{11} = \delta_{01} + \beta_3$; and $\beta_1 = \delta_{10} - \delta_{00}$, $\beta_3 = \delta_{11} - \delta_{01}$.

Collecting all these formulas we can find the one-to-one relation between δ and β coefficients:

$$\begin{aligned}\beta_0 &= \delta_{00}, & \delta_{00} &= \beta_0, \\ \beta_1 &= \delta_{10} - \delta_{00}, & \delta_{01} &= \beta_2, \\ \beta_2 &= \delta_{01}, & \delta_{10} &= \beta_0 + \beta_1, \\ \beta_3 &= \delta_{11} - \delta_{01}, & \delta_{11} &= \beta_2 + \beta_3.\end{aligned}$$

If the interaction term has a significant nonzero β_3 coefficient, then by substituting $G = 0$ and $G = 1$ in (1) we obtain:

$$\begin{aligned}P|(G=0) &= \beta_0 + \beta_2 E, \\ P|(G=1) &= (\beta_0 + \beta_1) + (\beta_2 + \beta_3) E.\end{aligned}$$

Accordingly, the difference between the two slopes is $(\beta_2 + \beta_3) - \beta_2 = \beta_3$.

As we can see from Eq. (1), β_3 is the coefficient of the interaction term. So the existence of an interaction means that there is a significant slope difference between the two levels of the genotype because $\beta_3 = \delta_{11} - \delta_{01}$, which is also the slope difference. Note that the intercept difference between the two lines is β_1 and it does not affect the interaction and can be expressed using only the G variable, as $\beta_1 = \delta_{10} - \delta_{00}$. Therefore, the slope difference is determined only by slopes δ_{11} and δ_{01} , not by the intercepts δ_{10} and δ_{00} . Thus, in the case of a binary-coded genetic variant (e.g., two genotypic levels, 0 and 1), the cross-product term accurately captures the interaction. Typically, however, genotypes are not binary and recoding them as binary cannot be recommended because doing so discards valuable information.

Use of the simple cross-product term when three genotype levels exist in the data is not recommended either. Only under highly specific circumstances does the cross-product term accurately reflect the interaction. This is because for the three-category genetic variant (yielding three “phenotype by environment” dependencies) there are six unknowns, these being δ_{ij} , where $j = 0, 1$ and $i = 0, 1, 2$. The four β_i coefficients in the interaction regression Eq. (1), cannot in general, determine all six unknown δ_{ij} coefficients. Hence, regression with the traditional cross-product interaction term cannot reliably recover the exact levels for all three genotypic groups. In Appendix A we delineate the specific conditions (which are unlikely to occur in practice) under which the traditional cross-product approach yields equivalent estimates.

The shortage of parameters to estimate all six parameters creates potential problems with modeling $G \times E$ using the cross-product term as delineated in Eq. (1) when there are three G levels, as it imposes a number of constraints, which are generally unacknowledged and may not be accurate for the data. The first is that the lines are always ordered; the second is that the slope differences between adjacent regression lines are always the same; and the third is that the lines always cross at the same point or are parallel. Note that the correspondence between the δ_{ij} and β_i coefficients as derived above means that the equations for $G = 0, 1, 2$ are, respectively:

$$\begin{aligned}
 P| (G=0) &= \beta_0 + \beta_2 E, \\
 P| (G=1) &= (\beta_0 + \beta_1) + (\beta_2 + \beta_3) E, \\
 P| (G=2) &= (\beta_0 + 2\beta_1) + (\beta_2 + 2\beta_3) E.
 \end{aligned}$$

We can see that the differences between the second and the first lines and also the third and the second lines are both $\beta_1 + \beta_3 E$ which means that the differences between the slopes of the lines are always the same and that the slopes are always ordered from the $G = 0$ line to the $G = 1$ line to the $G = 2$ line. However, there is no theoretical reason to assume that the Y axis differences between the lines should always be the same; in principle they could differ and in practice they will always differ due to sampling error and possibly from genuine non-linear $G \times E$ effects. Further, forcing the lines to be ordered (although this may make biological sense because it constrains the model to an additive effect), could lead to inaccurate representation of the data. Essentially it will always predict an ordered effect even when the data do not follow this pattern.

Another problem is that all three lines must cross at the same point or all must be parallel depending on the β_3 coefficient. When $\beta_3 = 0$ all three lines have the same slopes equal to β_2 . This means that all lines are parallel. When $\beta_3 \neq 0$ all three lines cross at the same point. To find the crossing point of two lines we equate the equations. Equating the first and second lines we get

$$\beta_0 + \beta_2 E = (\beta_0 + \beta_1) + (\beta_2 + \beta_3) E.$$

After simplification we find $\beta_1 + \beta_3 E = 0$ or $E = -\beta_1/\beta_3$. Hence the first two lines cross at $E = -\beta_1/\beta_3$. Similarly, by equating the second and third lines we can see that they also cross at the same point. Accordingly, the traditional parameterization of the interaction effect implies that all lines cross at the same point or all must be parallel depending on β_3 coefficient. There is no a priori reason to assume this would be the case or to force this constraint when modeling interaction.

In summary, modeling gene-environment interactions in the common way using the cross-product term, with three categories of the interacting variable (i.e., three levels of G) yields three critical problems. Because the regression lines for a three category genotypic variable have more unknowns than are in the traditional regression Eq. (1) with a cross-product term, only under very specific circumstances (when the coefficients satisfy Eqs. (A6) and (A7) in Appendix A) will the regression lines adequately capture the nature of the interaction in the data. Accordingly, modeling gene-environment interaction with an interaction regression line with four parameters produces lines with several constraints: first, the lines will always be ordered (e.g., 0, 1, 2 but never 0, 2, 1); secondly, differences between the lines must be the same (e.g., the difference between $G = 0$ and $G = 1$ will always be the same as the slope difference between $G = 1$ and $G = 2$); thirdly, the lines will always cross at a single point (or always will be parallel in the event of no interaction). None of these are necessarily true for any given $G \times E$, making the cross-product parameterization of $G \times E$ problematic.

A reparameterization of the equation to address the problem

In the second half of this paper we focus on constructing a model that appropriately captures gene-environment interactions with a three level genotype, i.e., one that accurately produces Eqs. (A1)–(A3) in Appendix A.

One approach is to recode the genotype data into three orthogonal components, taking the values:

$$\begin{aligned} H_0 &= (G - 1)(G - 2)/2 = \{1, 0, 0\}, \\ H_1 &= -G(G - 2)/2 = \{0, 1, 0\}, \\ H_2 &= G(G - 1)/2 = \{0, 0, 1\}. \end{aligned}$$

$$P = H_0(\delta_{00} + \delta_{01}E) + H_1(\delta_{10} + \delta_{11}E) + H_2(\delta_{20} + \delta_{21}E). \quad (2)$$

The first coefficient in Eq. (2), $(G - 1)(G - 2)/2$ is zero if $G = 1$ or $G = 2$ and is 1 if $G = 0$. Accordingly, the second $-G(G - 2)/2$ and third $-G(G - 1)/2$ coefficients also are 0 or 1 depending on the corresponding value of G . So, Eq. (2) is equivalent to Eqs. (A1), (A2) and (A3) in Appendix A when $G = 0, 1$ and 2 , respectively. Thus, the regression equation in (2) yields the correct equations for (A1)–(A3).

Substituting terms for H_0 , H_1 , and H_2 into Eq. (2), it is easy to see that the model involves G^2 and G^2E terms and corresponds to the full model with trichotomous G and continuous E variables with interaction (see, Kleinbaum et al. 2002, Chapters 6 & 7), which can be rewritten as:

$$P = \gamma_0 + \gamma_1 G + \gamma_2 E + \gamma_3 G \times E + \gamma_4 G^2 + \gamma_5 G^2 \times E. \quad (3)$$

The relationship between Eq. (3) and the coefficients from Eqs. (A1)–(A3) are as follows:

$$\begin{aligned} \gamma_0 &= \delta_{00}, \quad \delta_{00} = \gamma_0, \\ \gamma_1 &= -1.5\delta_{00} + 2\delta_{10} - 0.5\delta_{20}, \quad \delta_{01} = \gamma_2, \\ \gamma_2 &= \delta_{01}, \quad \delta_{10} = \gamma_0 + \gamma_1 + \gamma_4, \\ \gamma_3 &= -1.5\delta_{01} + 2\delta_{11} + 0.5\delta_{21}, \quad \delta_{11} = \gamma_2 + \gamma_3 + \gamma_5, \\ \gamma_4 &= 0.5\delta_{00} - \delta_{10} + 0.5\delta_{20}, \quad \delta_{20} = \gamma_0 + 2\gamma_1 + 4\gamma_4, \\ \gamma_5 &= 0.5\delta_{01} - \delta_{11} + 0.5\delta_{21}, \quad \delta_{21} = \gamma_2 + 2\gamma_3 + 4\gamma_5. \end{aligned}$$

In order to illustrate why this method is more appropriate for modeling $G \times E$ with three category genotypic variables we used R software to plot graphs showing properties of regression lines with an interaction term modeled using Eq. (1) [the four parameter model] and Eq. (3) [the six parameter model]. The R script is presented in Appendix B. For $n = 10,000$ individuals we assigned random values for a prototypical environmental variable (E) taking values between 0 and 10 with the same probability, and G variables having three levels: 0, 1 and 2. The G variable satisfies Hardy–Weinberg equilibrium conditions with minor allele frequency $p = 0.5$, so the numbers of individuals for $G = 0, 1$ and 2 levels are

$np(1-p) = 2,500$, $2np(1-p) = 5,000$ and $n(1-p)(1-p) = 2,500$, respectively. Then for each G level we assumed that the Phenotypes were linearly dependent on the environmental variable with the error having a normal distribution with mean = 0 and $SD = 0.1$. Figure 2 presents simulated data and observed regression lines from the four parameter and six parameter models for three scenarios in which a different one of the three constraints is violated in each one. In Figs. 2a, b and c the linear dependence of P on E for the G levels 0, 1 and 2 has the forms:

$$\begin{aligned} P(G=0) &= \text{Intercept}_0 + \text{slope}_0 \cdot E \\ P(G=1) &= \text{Intercept}_1 + \text{slope}_1 \cdot E \\ P(G=2) &= \text{Intercept}_2 + \text{slope}_2 \cdot E. \end{aligned}$$

Intercept and Slope coefficients are given in Table 1.

In Fig. 2a, the simulated data has equivalent slope differences between adjacent lines and the lines all cross at the same point; however, the data does not show ordered genotypic groups. As illustrated in Fig. 2a, the four-parameter model does not fit the data points and yields slopes in increasing order for G levels 0, 1 and 2, despite the fact that this is not the nature of the simulated lines (which are ordered 0, 2, 1). The interaction term in the four-parameter model is highly significant with $p < 10^{-16}$, which misleadingly indicates the existence of an additive interaction. This result is spuriously generated by the gene dose effects, which are not additive. The six parameter model accurately represents the data and captures the true nature of the regression lines.

Figure 2b illustrates the case when differences between slopes are different in the simulated data, but the other two constraints met. Because the four parameter model requires the slope differences to be the same between genotypic groups, the implied regression lines from the model will always be a poor fit to the actual shape of the data, as illustrated in the figure.

In Fig. 2c the simulated data for each genotypic group do not cross at the same point (note that this also means that the slopes are not ordered in an additive fashion). Again, although all three real lines do not cross at one point, the four parameter model shows all three lines as crossing. It yields a highly significant interaction term with $p < 10^{-16}$, but the predicted regression lines are a very poor representation of the data. The six parameter model accurately reflects the interaction.

In the above examples, a significant interaction would be detected by the four parameter model, but the implied nature of the interaction is erroneously indicated by the regression lines of the four parameter model. We consider a final example (indicated in the last columns of Table 1 and corresponding to Appendix C), illustrating a case where the four parameter model is not significant; however, an interaction exists in the data, and is accurately captured by the six parameter model. In this case the six parameter model accurately shows interaction with $p < 4.9 \times 10^{-12}$ but the four parameter model shows no interaction ($p = 0.239$) and corresponds to Type II error.

Testing for significance of gene-environment interaction effects using the reparameterized equation

In the previous section we demonstrated that the six parameter model more accurately captures gene-environment interaction effects with the addition of the G^2 terms as written in Eq. (3). However, the introduction of additional terms complicates how to test for the significance of an interaction effect. Accordingly, we provide guidelines for evaluating interaction effects here.

First we rewrite Eq. (3) for each of the three genotype groups:

$$\begin{aligned} \text{when } G=0, P|(G=0) &= \gamma_0 + \gamma_2 E, \\ \text{when } G=1, P|(G=1) &= (\gamma_0 + \gamma_1 + \gamma_4) + (\gamma_2 + \gamma_3 + \gamma_5) E, \\ \text{when } G=2, P|(G=2) &= (\gamma_0 + 2\gamma_1 + 4\gamma_4) + (\gamma_2 + 2\gamma_3 + 4\gamma_5) E. \end{aligned}$$

Recall that the significance of the interaction is determined by the slope difference between the lines for different groups.

The slope difference A between $G = 1$ and $G = 0$ is

$$A \equiv (\gamma_2 + \gamma_3 + \gamma_5) - \gamma_2 = \gamma_3 + \gamma_5.$$

Similarly, slope differences between $G = 2$ and $G = 1$, and between $G = 2$ and $G = 0$ are, respectively,

$$B \equiv (\gamma_2 + 2\gamma_3 + 4\gamma_5) - (\gamma_2 + \gamma_3 + \gamma_5) = \gamma_3 + 3\gamma_5,$$

and

$$C \equiv (\gamma_2 + 2\gamma_3 + 4\gamma_5) - \gamma_2 = 2\gamma_3 + 4\gamma_5.$$

Note also that $C = A + B$. Accordingly, we have three hypotheses and corresponding p -values

$$\begin{aligned} \text{Test } A: H_0: A &= 0, H_1: A \neq 0, p\text{-value } P_A, \\ \text{Test } B: H_0: B &= 0, H_1: B \neq 0, p\text{-value } P_B, \\ \text{Test } C: H_0: C &= 0, H_1: C \neq 0, p\text{-value } P_C. \end{aligned} \quad (4)$$

Generally, to test for any interaction effect means that one of the coefficients A , B or C is different from 0 (i.e., the slope difference between two of the genotypic groups is significant). In this case we can use the smallest of P_A , P_B and P_C as the general interaction p value.

However, in the context of studying gene-environment interaction, the assumed genetic model should be taken into account when determining whether the data support a meaningful interaction. That is to say, a three-category genotype cannot be treated

agnostically as any three level variable because each category corresponds to a particular genotype. Accordingly, whether an observed interaction is meaningful in a biological sense should be taken into account when evaluating the data. Figure 3 shows the six possible outcomes for the ordering of the genotypic categories and the signs associated with the corresponding slope differences. The signs and significance of A , B and C provide the necessary information to draw conclusions about whether the interaction is likely to be biologically meaningful. We suggest that this information should be taken into account when evaluating statistically significant interaction coefficients, as failure to conform to a likely biological model suggests an increased probability that the interaction is a false positive finding. We provide the framework for this evaluation process in Fig. 3. Note that in the situations represented in Fig. 3c to 3f, whether the slope difference (A or B) that has the opposite sign is significant, determines whether the implied genetic model is likely to be sensible. Only when the slope difference with the sign opposite the other two slope differences is nonsignificant does the interaction represent a recessive or dominant genetic model. Otherwise, the interaction reflects overdominance, a situation whereby the mean of the heterozygous group is not between the means of the two homozygous groups. The existence of overdominance is thought to be rare and remains somewhat controversial (Lynch and Walsh (1998)). We propose that, based on the current state of knowledge and the high propensity for post hoc explanations in the $G \times E$ literature, interactions that do not correspond to a classic genetic model (additive, dominant, or recessive) should be viewed skeptically without a clear reason why a more complex genetic model is appropriate for that study. We have listed the situations in which one would conclude that a significant $G \times E$ is supported by the data and the likely genetic model associated with the interaction (Fig. 3).

Hypothesis testing theory and procedures to test Hypotheses A , B and C above are explained in Appendix C. For the test we can either use the standardized t -distribution or an asymptotic of the t -distribution to create the Wald statistic. We provide R scripts and examples that work with general linear model outputs.

Extensions of the equation—We have delineated models to test for interaction using a three category (G) variable, as that is the most common situation in which the problems delineated here would arise in the context of testing for gene-environment interaction. However, we note that the problems described here associated with using cross-product terms to test for interaction are not specific to tests of gene-environment interaction. They will exist when testing for interaction with any variable that has three or more categories. For a four or more category interaction variable it is possible to use powers of that variable (for n category variable regression one would include powers of that variable up to $n - 1$ and also interactions/multiplications of all powers with the other interacting variable). For example, the extension of the model to a four category variable (using G and E as the interacting variables for ease of comparison to the previous equations) would be

$$P \sim G + G^2 + G^3 + E + G \times E + G^2 \times E + G^3 \times E.$$

A second extension of the model would be to allow for nonlinear interaction effects. These can also readily be incorporated into extensions of the model. For polynomial regression

(i.e. polynomial on environment variable) with a three category variable the regression representing all subclasses of the gene variable has the form

$$\begin{aligned}
 P \sim & G + G^2 + E + G \times E + G^2 \times E \text{ (linear)} \\
 & + E^2 + G \times E^2 + G^2 \times E^2 \text{ (quadratic)} \\
 & + E^3 + G \times E^3 + G^3 \times E^3 \text{ (cubic)} \\
 & \dots, \\
 & + E^k + G \times E^k + G^2 \times E^k \text{ (k=polynomial)}.
 \end{aligned}$$

Although it is not advisable to test for more complex interactions without an explicit a priori rationale, due to the loss of power when estimating an increasing number of parameters and the possibility of collinearity problems, we note that it is possible to test for them when circumstances warrant.

Discussion

There has been a proliferation of studies on gene-environment interaction in recent years, in both candidate gene and GWAS. Most studies in this large and growing literature use a cross-product term to model interaction between a genotype and environmental variable. In the case of binary variables, such as genotype of males on the X chromosome, the use of the cross-product term is appropriate. For loci with very low minor allele frequencies, such that no homozygotes for the minor allele exist in the population, this simple model also yields unbiased results. In all other cases, i.e., where three genotype categories exist, the cross-product approach should be avoided. This is because when three genotype categories exist (0, 1, 2), the use of the only one cross-product term only reproduces interaction effects accurately under certain very unusual circumstances. When the minor allele frequency is low, say 10 % then about 1 % of data is expected as rare homozygous. In this case the majority of data will lay on other two strata and the model with only one cross product term can reflect the model enough well as in the case of having only two strata. Also note that for very rare alleles when there is no third strata G^2 and G are linearly dependent (if G gets values 0 or 1 and there is no $G = 2$, then $G^2 = G$, if G gets the values 1 or 2 and there is no $G = 0$, then $G^2 = 3G - 2$) the model should not have G^2 terms, otherwise the model will have collinearity issues. Essentially, the use of only one cross-product term with three genotype categories imposes constraints on the interaction: it forces the slope (and also intercept) difference to be the same between adjacent genotypic groups (e.g., the difference in slope between the groups $G = 0$ and $G = 1$ is constrained to be the same as the slope difference between the groups where $G = 1$ and $G = 2$) and it forces the lines for the three genotypic groups to all cross at the same point when an interaction exists, which also means that slopes and intercepts of all three lines are ordered. Accordingly, an interaction will only be accurately represented by the cross-product term when these conditions are met. Particularly concerning in the context of interpreting gene-environment interactions is that use of a cross-product term to model interactions will produce regression lines that always follow an ordered interaction pattern that maps onto an additive effect even when the true shape of the interaction in the data does not follow that pattern. It is common practice to plot the estimated regression lines to illustrate interactions; however, in many cases doing so will be misleading because it will not accurately represent the shape of the interaction in the data. If

the researcher did not plot the actual data points and instead relied on a significant p-value associated with the cross-product term as evidence of interaction, this inaccuracy would go unnoticed. Biased estimates and increased Type I and Type II error rates would result.

We provide a reparameterization of the traditional regression model that incorporates a G^2 term and we demonstrate that the use of this expanded equation solves the problems that currently exist when a cross-product term is used to model a three-category genotype variable. This reparameterization allows for the accurate reproduction of interactions that may exist in the data, and removes the constraint that the slope differences must be the same between groups and always follow the order 0, 1, 2 and the requirement that per strata regression lines must cross at the same point. Accordingly, as in the case of a binary variable interaction (e.g., an X-linked locus in males), the significance of the interaction term is once again based entirely on the slope differences between genotypic groups without confounding the test with information about the intercepts or mean differences between genotypic groups. This revised model allows for accurate characterization and visualization of any gene-environment interaction effects that may exist in the data. We note that alternatively it is possible to use dummy-coded variables; for a three-category genotype two dummy variables and their cross-products with the environmental variable would be necessary. The use of dummy-coded genotypic variables for modeling interaction is mathematically equivalent to including the G^2 term, as well as allele-based coding, within a multiple regression framework. Its omission is parallel to the situation in latent variable interaction models for twin studies (Purcell 2002), where non-additive effects of the genotype can mimic or obscure $G \times E$ interaction (Van Hulle et al. 2013).

Accurately modeling the nature of the interaction effect using the G^2 term also allows the researcher to use information about the nature of the observed interaction to evaluate the likelihood that the finding represents a real effect. The use of p-value cut-offs is essentially a statistical guide to reduce the likelihood that a finding is a false positive; because genotypes have biological meaning, we suggest that this information should also be used to evaluate the likelihood that a statistically significant effect actually represents a potentially biologically meaningful phenomenon. We suggest that interactions that do not conform to commonly accepted genetic models (additive, dominant, recessive) be viewed with extreme skepticism and only be accepted as evidence of a meaningful gene-environment interaction effect when there is a clear a priori reason to expect that a more complex genetic model is applicable.

In summary, we demonstrate that modeling gene-environment interaction using a cross-product term when the genotype is coded using three categories is problematic and should be abandoned. We present a reparameterization of the regression equation that accurately captures the shape of potential interactions that may exist in the data. We suggest that researchers interested in the study of gene-environment interaction effects (as well as any researchers testing for interaction with polytomous variables) use a six-parameter model equation (or more, depending upon the number of levels within the categorical variable) in future tests of interaction effects. Moreover, existing reports of interaction effects in the literature that applied the cross-product regression approach (the four-parameter model) to three-category genotype variables must be re-evaluated to assess their accuracy.

Acknowledgments

This work grows out of a program of research supported by K02AA018755 (to DMD). MCN was supported by R37DA18673.

Appendix A

Here we delineate the specific conditions under which the traditional cross-product approach yields accurate estimates for a G variable with three levels.

Suppose that we have groups of individuals with three distinct genotypes ($G = 0, 1$ and 2 meaning each group has 0, 1 or 2 copies of the reference allele, respectively). Assume that there is a linear regression in each group between the phenotype and the environment.

$$G=0 \text{ group has the relation } P|(G=0) = \delta_{00} + \delta_{01}E \quad (A1)$$

$$G=1 \text{ group has the relation } P|(G=1) = \delta_{10} + \delta_{11}E \quad (A2)$$

$$G=2 \text{ group has the relation } P|(G=2) = \delta_{20} + \delta_{21}E. \quad (A3)$$

The aim is to determine the δ_{ij} coefficients using the $G \times E$ additive regression model for all groups of individuals/genotypes from Eq. (1). When β_i coefficients of (1) are available for all individuals, for the subset of individuals with $G = 0$, equating $G = 0$ in (1) we have:

$$P|(G=1) = \beta_0 + \beta_2E.$$

Combining this with Eq. (A1), which also represents the $G = 0$ group, we get:

$$P|(G=1) = \beta_0 + \beta_2E = \delta_{00} + \delta_{01}E.$$

Accordingly, $\delta_{00} = \beta_0$ and $\delta_{01} = \beta_2$. Note that β_0 and β_2 coefficients are already defined by δ_{00} and δ_{01} as $\beta_0 = \delta_{00}$ and $\beta_2 = \delta_{01}$. Using these values in Eq. (1) changes it to the equivalent equation:

$$P = \delta_{00} + \beta_1G + \delta_{01}E + \beta_3G \times E. \quad (A4)$$

Similarly, for the case where $G = 1$, using Eq. (A4) now instead of Eq. (1) and combining with Eq. (A2), the regression line specific to the $G = 1$ group yields:

$$\begin{aligned} P|(G=1) &= \delta_{00} + \beta_1 \cdot 1 + \delta_{01}E + \beta_3E \\ &= (\delta_{00} + \beta_1) + (\delta_{01} + \beta_3)E = \delta_{10} + \delta_{11}E. \end{aligned}$$

Hence $\delta_{10} = \delta_{00} + \beta_1$ and $\delta_{11} = \delta_{01} + \beta_3$ or $\beta_1 = \delta_{10} - \delta_{00}$ and $\beta_3 = \delta_{11} - \delta_{01}$.

This means Eq. (A4) changes to

$$P = \delta_{00} + (\delta_{10} - \delta_{00})G + \delta_{01}E + (\delta_{11} - \delta_{01})G \times E. \quad (A5)$$

Accordingly, for the case where $G = 2$, applying Eqs. (A5) and (A3) yields

$$\begin{aligned} P| (G=2) \\ &= \delta_{00} + (\delta_{10} - \delta_{00}) \cdot 2 + \delta_{01}E + (\delta_{11} - \delta_{01}) \cdot 2 \cdot E \\ &= \delta_{20} + \delta_{21}E. \end{aligned}$$

Accordingly,

$$\delta_{20} = 2\delta_{10} - \delta_{00}, \quad (A6)$$

$$\delta_{21} = 2\delta_{11} - \delta_{01}. \quad (A7)$$

This means that the six δ_{ij} coefficients are not free because there is a linear dependency among them. Only when Eqs. (A6) and (A7) are satisfied are Eqs. (A1)–(A3) consistent. This means that only under very specific conditions will the use of the traditional cross-product interaction term yield the appropriate regression lines to model those interactions observed in the data. These conditions, dictated by Eqs. (A6)–(A7), are equivalent to

$$\delta_{20} - \delta_{00} = 2(\delta_{10} - \delta_{00}), \delta_{21} - \delta_{01} = 2(\delta_{11} - \delta_{01}). \quad (A8)$$

Appendix B

In this section we generate gene (G) variable under Hardy–Weinberg condition and phenotype (P) variable using random environment (E) variable and then run tests for interaction with the two parameterizations explained in the paper (Eqs. 1 and 3).

```

# We generate gene variable with H-W condition using p as allele frequency
# regression for G = 0, 1, 2 are P = g0_a + g0_b*E, P = g1_a + g1_b*E,
# P = g2_a + g2_b*E

n=10000          # number of individuals
p=0.5            # allele frequency
g0_intercept=0.4  # intercept for G=0
g0_slope= -0.4    # slope for Gene=0
g1_intercept=-0.4 # intercept for G=1
g1_slope= 0.4     # slope for Gene=1
g2_intercept=0.0  # intercept for G=2
g2_slope= 0.0     # slope for Gene=2

# H-W satisfied numbers of individuals
n0=round(n*(1-p)*(1-p)) # number of (allele1,allele1) individuals
n1=round(2*n*p*(1-p))   # number of (allele1,allele2) individuals
n2=n-n0-n1              # number of (allele2,allele2) individuals

# generating residuals
P=rnorm(mean=0,sd=0.1,n)

# generating env values
G=rep(0,n);          # initial Gene values
E=sample(0:10, n, replace=T) # Env variable

# calculating P values for G=0,1,2 levels
for (i in 1:n) {
  if (i<=n0) { G[i]=0; P[i]=P[i]+g0_intercept+g0_slope*E[i] }
  else if (i<=n0+n1) { G[i]=1; P[i]=P[i]+g1_intercept+g1_slope*E[i] }
  else { G[i]=2; P[i]=P[i]+g2_intercept+g2_slope*E[i] }
}

G_G <- G^2          # calculating Gene^2
G_E <- G*E          # calculating Gene*Env
G2_E <- G^2*E       # calculating Gene^2*Env

# running six parameters regression model
tt=lm(P ~ G + E + G_E + G_G + G2_E)$coefficients

# running four parameters regression model
ff=lm(P ~ G + E + G_E)$coefficients

# defining lines for graph
xline=0:10          # x line for Env variable

# gene layers from 6 variable regression
t0=tt[1]+tt[3]*xline

```

```

t1=(tt[1]+tt[2]+tt[5])+(tt[3]+tt[4]+tt[6])*xline
t2=(tt[1]+2*tt[2]+4*tt[5])+(tt[3]+2*tt[4]+4*tt[6])*xline
# gene layers from 4 variable regression
f0=ff[1]+ff[3]*xline
f1=(ff[1]+ff[2])+(ff[3]+ff[4])*xline
f2=(ff[1]+2*ff[2])+(ff[3]+2*ff[4])*xline
# calculating a range for plots
yrange<-range(c(t0,t1,t2,f0,f1,f2))
par(mfrow=c(1,2))
# value +0.03 and -0.03 below are added to shift lines
# to make all overlapped colors visible
plot(E[1:n0]+0.03, P[1:n0],type='p',xlab='Environment',
      main="Four Parameter Model",
      ylab='Phenotype',xlim=c(0,10),ylim=yrange,col="green")
par(new=T);
plot(E[(n0+1):(n0+n1)]-0.03,P[(n0+1):(n0+n1)],type='p',xlab='',
      main="",ylab='', xlim=c(0,10),ylim=yrange,col="red")
par(new=T);
plot(E[(n0+n1+1):n], P[(n0+n1+1):n],type='p',xlab='',main="",ylab='',
      xlim=c(0,10),ylim=yrange,col="blue")
par(new=T);

plot(xline,f0,type='l',xlab='',ylab='',xlim=c(0,10),ylim=yrange,col="green")
par(new=T);
plot(xline,f1,type='l',xlab='',ylab='',xlim=c(0,10),ylim=yrange,col="red")
par(new=T);
plot(xline,f2,type='l',xlab='',ylab='',xlim=c(0,10),ylim=yrange,col="blue")
legend(0,yrange[2], c("Gene=0", "Gene=1", "Gene=2"),
      col = c("green","red","blue"),lty=c(1,1,1))
plot(E[1:n0]+0.03, P[1:n0],type='p',xlab='Environment',main="Six Parameter
Model",
      ylab='Phenotype',xlim=c(0,10),ylim=yrange,col="green")
par(new=T);
plot(E[(n0+1):(n0+n1)]-0.03,
P[(n0+1):(n0+n1)],type='p',xlab='',main="",ylab='',
      xlim=c(0,10),ylim=yrange,col="red")
par(new=T);
plot(E[(n0+n1+1):n], P[(n0+n1+1):n],type='p',xlab='',main="",ylab='',
      xlim=c(0,10),ylim=yrange,col="blue")
par(new=T);
plot(xline,t0,type='l',xlab='',ylab='',xlim=c(0,10),ylim=yrange,col="green")
par(new=T);
plot(xline,t1,type='l',xlab='',ylab='',xlim=c(0,10),ylim=yrange,col="red")
par(new=T);
plot(xline,t2,type='l',xlab='',ylab='',xlim=c(0,10),ylim=yrange,col="blue")
legend(0,yrange[2], c("Gene=0", "Gene=1", "Gene=2"),
      col = c("green","red","blue"),lty=c(1,1,1))

```

Appendix C

Consider the linear regression model $y = \beta_0 + \beta_1 x_1 + \dots + \beta_p x_p + \varepsilon$ with dependent variable y and independent variables $x_1, \dots, x_p$ and assume that error term ε is normally distributed. Denote $\boldsymbol{\beta} = (\beta_0, \beta_1, \dots, \beta_p)'$, where $'$ represents the transpose sign. Having a sample of size n with values of $y_i, x_{1i}, \dots, x_{pi}$ ($i = 1, \dots, n$) the least squares estimate for $\boldsymbol{\beta}$ can be expressed

as $\boldsymbol{\beta}' = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$, where $\mathbf{y} = (y_1, \dots, y_n)'$ and $\mathbf{X}$ is the matrix containing n rows $(1, x_{1i}, \dots, x_{pi}), i = 1, \dots, n$.

To test the hypothesis $H_0: \mathbf{b}'\boldsymbol{\beta} = 0$ vs $H_1: \mathbf{b}'\boldsymbol{\beta} \neq 0$ with $\mathbf{b} = (b_0, b_1, \dots, b_p)'$ being a contrast vector, note that $\mathbf{b}'\boldsymbol{\beta} = \mathbf{b}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ has a t -distribution with standard error

$s \sqrt{\mathbf{b}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{b}}$ under the null hypothesis. s^2 is the estimated residual variance. So, for the test we can either use the standardized t -distribution or use the fact that the corresponding Wald statistic $\mathbf{b}'\boldsymbol{\beta}(\mathbf{b}'s^2\mathbf{b}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{b})^{-1}\mathbf{b}'\boldsymbol{\beta}$ has asymptotic χ^2 distribution.

The `linearHypothesis` function from the R `car` package tests contrasts i.e. linear combinations of model coefficients. The `linearHypothesis` function tests linear hypotheses and methods for linear models, generalized linear models, multivariate linear models, linear and generalized linear mixed-effects models. `linearHypothesis` computes either a finite-sample F statistic or asymptotic Chi squared statistic for carrying out a Wald-test-based comparison between a model and a linearly restricted model. For mixed-effects models, the tests are Wald Chi square tests for the fixed effects.

Assuming that variables are named as G-Gene, E-Environment, $G_E = G * E$ and $G2_E = G^2 * E$ and linear model results are named as `model_result` to test Hypotheses $A = 0$ (which is same as $\gamma_3 + \gamma_5 = 0$), $B = 0$ and $C = 0$ explained with Eqs. (4) we can use

```
linearHypothesis(model_result, "1 * G_E + 1 * G2_E = 0"),
linearHypothesis(model_result, "1 * G_E + 3 * G2_E = 0"),
linearHypothesis(model_result, "2 * G_E + 4 * G2_E = 0"),
```

respectively. Note that “car” package needs to be installed in order to use `linearHypothesis` function.

Alternatively, we provide an R function which is based on the t -distribution for direct calculation of the contrast test statistic and p-value.


```

# Explanations for the contrast_test function:
# Usage
# contrast_test(model_result, vector_b, G_E="G_E", G2_E="G2_E")
# Arguments
# model_result  result object from linear regression or mixed models
# vector_b      2d vector of contrast coefficients (for GxE and G*2xE)
# G_E          variable name of GxE variable in the linear model
# G2_E         variable name of G*2xE variable in the linear model
# This function calculates estimation, st. error, t-value and p-value
# for the contrast with given coefficients vector_b
contrast_test <- function(model_result, vector_b, G_E="G_E", G2_E="G2_E")
{ summaryl <- summary.lm(model_result)
  # check if model is consistent, i.e. all coefficients are non NA
  if (length(rownames(summaryl$coef))!=6)
  { # finding row numbers with G_E and G2_E and calculating contrast
    coefficients
    linear_vec <- vector_b[1]*(rownames(summaryl$coef)==G_E)+
      vector_b[2]*(rownames(summaryl$coef)==G2_E)
    mod_sigma <- summaryl$sigma
    df.error <- summaryl$df[2]
    # calculating estimated contrast value
    estimation <- c(t(linear_vec) %*% coef(model_result))
    stderr <- sqrt(c(t(linear_vec) %*%
      (mod_sigma^2*summaryl$cov.unscaled) %*% linear_vec))
    t <- estimation/stderr
    # calculating p-value from t-distribution
    p <- 2*(1-pt(abs(t), summaryl$df[2]))
    return(list(est=estimation, stderr=stderr, t=t, p=p))
  } else { return(list(est="NA", stderr="NA", t="NA", p="NA")) }
}

```

Now assuming that variables are named as $G_E = \text{gene} \times \text{env}$ and $G2_E = \text{gene2} \times \text{env}$ and linear model results are named as `model_result` to test Hypotheses A, B and C from Eqs. (4) we can use

```

contrast_test(model_result, c(1,1), G_E, G2_E),
contrast_test(model_result, c(1,3), G_E, G2_E),
contrast_test(model_result, c(2,4), G_E, G2_E),

```

respectively.

The next function tests all three tests for the gene x environment effect and returns an overall p-value.

```

# Explanations for the combined_test function:
# Usage
# combined_test(model_result, vector_b, correct=T)
# Arguments
# model_result  result object from linear regression or mixed models
# correct       when TRUE corrects for multiple testing, for additive
#               GxE interaction
# G_E           variable name of GxE variable in the linear model
# G2_E          variable name of G^2xE variable in the linear model
# This function calculates overall p-value based on three p-values for
# the slope differences
combined_test <- function(model_result, correct=TRUE, G_E="G_E", G2_E="G2_E")
{ # p-values for A, B and C
  p_a <- contrast_test(model_result, c(1,1), G_E, G2_E)$p
  p_b <- contrast_test(model_result, c(1,3), G_E, G2_E)$p
  p_c <- contrast_test(model_result, c(2,4), G_E, G2_E)$p
  a <- contrast_test(model_result, c(1,1), G_E, G2_E)$est
  b <- contrast_test(model_result, c(1,3), G_E, G2_E)$est
  return_pval <- min(p_a,p_b,p_c)
  # check if p-values are determined
  if (return_pval != "NA")
  { if (correct) {
      return_pval <- (min(1, 2 * return_pval)) # corrected p-value
    } else {
      if (sign(a)*sign(b)<0) { return_pval <- 1 } # no GxE interaction
    }
  } else { return_pval = "NA" }
  return(return_pval)
}

```

R example with simulation:

```

# We simulate data for G and E variables and create Phenotype
# using linear equations and adding random normal component
set.seed(346917)

n <- 1000 # sample size
G <- sample (0:2, n, replace=T) # creating gene variable
E <- sample (0:10, n, replace=T) # creating env variable
G_G <- G^2 # calculating gene^2
G_E <- G * E # calculating gene*env
G2_E <- G^2 * E # calculating gene^2*env

# assigning delta coefficients for three lines
d00=0; d01=0.005; d10=0; d11=0.0; d20=2; d21=-0.005

# calculating gamma coefficients of the full model
g0=d00 #Intercept
g1=-1.5*d00+2*d10-0.5*d20 #Gene (G)
g2=d01 #Environment (E)
g3=-1.5*d01+2*d11-0.5*d21 #GxE
g4= 0.5*d00- d10+0.5*d20 #G^2
g5= 0.5*d01- d11+0.5*d21 #G^2xE

# creating phenotype(P) with additional normal noise
P <- g0+g1*G+g2*E+g3*G_E+g4*G_G+g5*G2_E+0.058*rnorm(n)

# running combined_test with created 6 parameter model
model_result=lm(P ~ G + E + G_G + G_E + G2_E) # model
final_pvalue <- combined_test(model_result, correct=T, G_E="G_E",
G2_E="G2_E" )
final_pvalue

# running the test with four parameter model
summary(lm(P ~ G + E + G_E))

```

References

- Bakermans-Kranenburg MJ, van Ijzendoorn MH. Gene-environment interaction of the dopamine D4 receptor (DRD4) and observed maternal insensitivity predicting externalizing behavior in preschoolers. *Dev Psychobiol.* 2006; 48(5):406–409. [PubMed: 16770765]
- Caspi A, Sugden K, Moffitt TE, Taylor A, Craig IW, Harrington H, McClay J, Mill J, Martin J, Braithwaite A, Poulton R. Influence of life stress on depression: moderation by a polymorphism in the 5-HTT gene. *Science.* 2003; 301(5631):386–389. [PubMed: 12869766]
- Duncan L, Keller MC. A critical review of the first 10 years of measured gene-by-environment interaction research in psychiatry. *Am J Psychiatry.* 2011; 168(10):1041–1049. [PubMed: 21890791]
- Keller M. Gene 9 Environment interaction studies have not properly controlled for potential confounders: the problem and the (simple) solution. *Biol Psychiatry.* 2014; 75(1):18–24. [PubMed: 24135711]
- Khoury MJ, Wacholder S. Invited commentary: from genomewide association studies to gene-environment-wide interaction studies—challenges and opportunities. *Am J Epidemiol.* 2009; 169(2):227–230. [PubMed: 19022826]
- Kleinbaum, DG.; Klein, M.; Pryor, ER. Logistic regression: a self-learning text. 2nd edn.. New York: 2002. ISBN 0387953973
- Lynch, M.; Walsh, B. Genetics and analysis of quantitative traits. Sinauer Associates Inc; Sunderland: 1998.
- Murcray CE, Lewinger JP, Gauderman WJ. Genome-environment interaction in genome-wide association studies. *Am J Epidemiol.* 2009; 169(2):219–226. [PubMed: 19022827]
- Petersen IT, Bates JE, Goodnight JA, Dodge KA, Lansford JE, Pettit GS, Latendresse SJ, Dick DM. Interaction between serotonin transporter polymorphism (5-HTTLPR) and stressful life events in

adolescents' trajectories of anxious/depressed symptoms. *Dev Psychol.* 2012; 48(5):1463–1475. [PubMed: 22390669]

Purcell S. Variance components models for gene-environment interaction in twin analysis. *Twin Res.* 2002; 5(6):554–571. [PubMed: 12573187]

Van Hulle CA, Lahey BB, Rathouz PJ. Operating characteristics of alternative statistical methods for detecting gene-by-measured environment interaction in the presence of gene-environment correlation in twin and sibling studies. *Behav Genet.* 2013; 43(1):71–84. [PubMed: 23090766]

Widaman KF, Helm JL, Castro-Schilo L, Pluess M, Stallings MC, Belsky J. Distinguishing ordinal and interactions. *Psycholog Methods.* 2012; 17(4):615–622.

Fig. 1.
Regression lines as illustrated for a three category genotype

Fig. 2.

a Simulated data violates the constraint of ordered genotypes. Parameters are the same as in Appendix B and Table 1. **b** Simulated data violates the constraint of equivalent slope differences between adjacent genotypic groups. Parameters are the same as in Table 1. **c** Simulated data violates the constraint of all lines crossing at a single point. Parameters are the same as in Table 1

Fig. 3.

Possible outcomes for slope differences and corresponding conclusions about interactions

Table 1

Coefficients for scenarios for Fig. 2a, b and c and for the example

	<u>Fig. 2a</u>		<u>Fig. 2b</u>		<u>Fig. 2c</u>		<u>Example ($n = 1,000$)</u>	
	Intercept	Slope	Intercept	Slope	Intercept	Slope	Intercept	Slope
G = 0	5.0	-0.8	0.4	-0.4	0.1	-0.1	0.0	0.005
G = 1	0.0	0.0	-0.4	0.4	0.0	0.0	0.0	0.000
G = 2	-0.2	0.8	0.0	0.0	-1.0	0.9	2.0	-0.005