Psychological Methods

Ubiquitous Bias and False Discovery Due to Model Misspecification in Analysis of Statistical Interactions: The Role of the Outcome's Distribution and Metric Properties

Benjamin W. Domingue, Klint Kanopka, Sam Trejo, Mijke Rhemtulla, and Elliot M. Tucker-Drob

Online First Publication, October 6, 2022. http://dx.doi.org/10.1037/met0000532

AMERICAN PSYCHOLOGICAL ASSOCIATION

Psychological Methods

ISSN: 1082-989X

https://doi.org/10.1037/met0000532

Ubiquitous Bias and False Discovery Due to Model Misspecification in Analysis of Statistical Interactions: The Role of the Outcome's Distribution and Metric Properties

Benjamin W. Domingue[1], Klint Kanopka[2], Sam Trejo[3], Mijke Rhemtulla[4], and Elliot M. Tucker-Drob[5]

[1] Graduate School of Education, Stanford University and Center for Population Health Sciences, Stanford Medicine
[2] Graduate School of Education, Stanford University
[3] Department of Sociology and Office of Population Research, Princeton University
[4] Department of Psychology, University of California, Davis
[5] Department of Psychology and Population Research Center, University of Texas at Austin

Abstract

Studies of interaction effects are of great interest because they identify crucial interplay between predictors in explaining outcomes. Previous work has considered several potential sources of statistical bias and substantive misinterpretation in the study of interactions, but less attention has been devoted to the role of the outcome variable in such research. Here, we consider bias and false discovery associated with estimates of interaction parameters as a function of the distributional and metric properties of the outcome variable. We begin by illustrating that, for a variety of noncontinuously distributed outcomes (i.e., binary and count outcomes), attempts to use the linear model for recovery leads to catastrophic levels of bias and false discovery. Next, focusing on transformations of normally distributed variables (i.e., censoring and noninterval scaling), we show that linear models again produce spurious interaction effects. We provide explanations offering geometric and algebraic intuition as to why interactions are a challenge for these incorrectly specified models. In light of these findings, we make two specific recommendations. First, a careful consideration of the outcome's distributional properties should be a standard component of interaction studies. Second, researchers should approach research focusing on interactions with heightened levels of scrutiny.

Translational Abstract

There is great scientific interest in the degree to which responses to some common stimulus vary across people. Many tests of such variation involve the statistical analysis of interaction terms. We use a variety of evidence (geometric, algebraic, simulation) to argue that incorrect inferences may be made in many cases if details of the outcome variable are not closely monitored. In particular, we show that false positives will result in many cases if a model is not well-suited to the nature of the outcome variable. We offer illustrations from the literature of places where such confusion can occur. We believe that an increased understanding of this problem would lead to improved scientific inquiry and more efficient use of research funds.

Keywords: interactions, misspecification, bias, false, discovery

Supplemental materials: https://doi.org/10.1037/met0000532.supp

Lived experience suggests substantial heterogeneity in how people react to a common intervention, treatment, or exposure.[1] This suggests the hypothesis that some features (of the stimuli, environment, or person) may explain this heterogeneity. For example, which psychological factors account for variation in the success of smoking cessation programs (Halpern et al., 2016)? Does the level of public regard moderate the effect of racial discrimination on depressive symptoms (Seaton & Iida, 2019)? Are the effects of growth mindset interventions modified by environmental features (Claro et al., 2016; Yeager et al., 2019)? These few examples are a small portion of the larger literature probing for such heterogeneities in empirical settings. We believe, along with others (Bryan et al., 2021), that such questions are crucial.

Benjamin W. Domingue https://orcid.org/0000-0002-3894-9049

This research was funded in part by the Jacobs Foundation and National Institutes of Health research Grants R01MH120219 and RF1AG073593. An earlier version of this article was available as a preprint (10.31234/osf.io/932fm).

Correspondence concerning this article should be addressed to Benjamin W. Domingue, Graduate School of Education, Stanford University and Center for Population Health Sciences, Stanford Medicine, 520 Galvez Mall, Room 510, CERAS Building, Stanford, CA 94305, United States. Email: ben.domingue@gmail.com

[1] This is a paraphrase of what we heard Jeremy Freese call the First Law of Sociology—"some do, some don't"—but psychologists also use it (Haaf & Rouder, 2019).

The standard tool for analysis of such heterogeneity is the inclusion of interaction terms in statistical models. However, analysis of such models is fraught. We are not the first to note problems with interaction models; such issues have been remarked upon regularly over the last several decades (e.g., Loftus, 1978; Lubinski & Humphreys, 1990; Rohrer & Arslan, 2021; Tabery, 2008). We briefly discuss previously noted methodological issues below. In this article, we focus specifically on issues in interaction studies that arise from measurement properties of the outcome variables. In particular, we explore problems resulting from a mismatch between the measurement properties of the outcome variable and the distributional assumptions of the statistical model. Statistical models for interactions rely upon strong assumptions regarding the distribution and measurement properties of the outcome variable. When these assumptions are not met, conventional regression models may produce biased interaction effect estimates and high rates of Type I error; that is, such models tend to produce spurious positive findings due to misspecification. Our work here builds specifically on previous studies emphasizing the role of outcome scale in subsequent interpretation (Loftus, 1978), but with a focus on bias and false discovery.

Previously Expressed Concerns Regarding Analysis of Statistical Interactions

Imagine that one is interested in understanding the interaction of two predictors (x and z) in the study of some outcome (y). The outcome is a function of the predictor variables and model parameters. In particular, we assume that $\mathbb{E}(y \mid x, z) = f(\beta_0 + \beta_1 x + \beta_2 z + \beta_3 xz)$ for some function of f. Interest is in an estimate of β_3; therefore, we denote such estimates $\widehat{\beta_3}$. Previous studies have emphasized several potential problems. One problem is that interactions may yield false positives if interactions between additional covariates (w) are not also included in the model (Keller, 2014). If w is a covariate of interest, then analysis of the interaction xz should be based on models that also include xw and zw. A second problem is that if the true data generating model is based on x^2, interaction studies focusing on xz may lead to false positives if x and z are correlated (Lubinski & Humphreys, 1990; MacCallum & Mar, 1995). There are also specialized concerns that might arise due to the specifics of a given context; for example, there are unique concerns associated with analysis of gene-environment interaction (Dudbridge & Fletcher, 2014).

In addition, there are concerns related to interpretation and generalizability of model results. The substantive implications of findings of interactions may be highly contingent on the nature of x and z. In particular, a focus on nonexogenous environments can lead to multiple viable interpretations (Fletcher & Conley, 2013). Under certain configurations, statistical power of interaction studies will be substantially lower than in studies of main effects (see Section 16.4 of Gelman et al., 2020). We highlight these issues for two reasons. First, they help to emphasize that there are numerous challenges to the statistical analysis of interactions. Second, much of the previous scholarship on the estimation of interactions has focused specifically on the variables on the “right-hand side” of the equation; relatively less focus has been paid to outcome variables. We focus explicitly on the “left-hand side” and show that characteristics of the outcome can have major implications for our ability to recover the relevant parameters.

Outcome Types

We consider a range of outcomes that collectively span a broad range of outcomes of interest in psychology. Below we briefly describe the outcomes and the occasions where they may be used in psychological research. We also emphasize analytic approaches used to study them. While most of these outcome types have specialized approaches developed for their analysis, it is also not uncommon to see them analyzed via the linear model (a point we illustrate below). There are a range of reasons why the linear model may be used—including computational feasibility, intuitive interpretation of results, and the lack of complete information about the outcome that would allow for deployment of a superior alternative—but we also emphasize that the linear model can produce high levels of false positives in many cases.

Given that the linear model is used for analysis of the full range of outcomes, we consider analysis of all outcomes via both tailored approaches (when available) and a (misspecified) linear alternative. The linear model is a valuable tool, and its simplicity makes it appealing for many purposes. Nonetheless, we show that the linear model can lead to highly misleading results when used to study interactions as a function of the geometry of the outcome variable.

Binary Outcomes

Binary outcomes are common in analysis of, for example, mental health diagnoses (Ancelin et al., 2017; Culverhouse et al., 2018; Stringa et al., 2020). Such outcomes may be analyzed via logistic or Probit regression, two variants of the generalized linear model (Nelder & Wedderburn, 1972). Here, we focus on the logistic regression approach. The linear model can also be used in the form of what is frequently called the linear probability model (Gomila, 2020). This approach ignores the key feature of the outcome (i.e., it only takes values 0 and 1) for the potential computational gains associated with ordinary least squares estimation of the linear model and for the convenience of being able to work with straightforward coefficient estimates. With the linear probability model, coefficient estimates describe the change in probability associated with a change in the predictor (rather than the more complicated odds ratio interpretation required for working with coefficients from logistic regression).

To motivate evaluation of this type of outcome, we note two recent instances in which statistical interactions were estimated for binary outcomes using a linear probability model. As a first example, consider a recent analysis of behavioral nudges meant to increase COVID vaccination uptake (Dai et al., 2021). The investigators used linear models to examine heterogeneity of nudge efficacy on a binary measure of vaccination status as a function of, for example, flu vaccination status. As a second example, consider the testing of an interaction between genetics and exposure to trauma in the prediction of a binary index of major depressive disorder (Coleman et al., 2020). Importantly, this study considers the logistic and linear approaches in tandem, which provides a useful illustration of the underlying potential for misinterpretation. The two forms of analysis lead to divergent results: Coleman and coauthors find a significant interaction effect when using the linear model but not logistic regression. Their visualization (see their Figure 2) underscores the difference: For the linear model, the regression line representing the association between genetics and probability of major depression is considerably steeper for the trauma exposed group compared to the unexposed group, whereas for the logistic model, the regression lines representing the association between genetics and the log-odds of major depression are parallel for trauma exposed and unexposed groups. We show how such a result can arise as a function of the prevalence of the outcome; in particular, binary outcomes with prevalences far from 50% have high rates of false positive interactions when the linear model is used (this point has been previously raised in the empirical logit analysis setting; Donnelly & Verkuilen, 2017).

Count Outcomes

Count outcomes occur when interest is in the number of times something occurred; they are used to quantify, for example, drug/alcohol use (Angosta et al., 2019; Meyers et al., 2019) or frequency of medical care (Richardson & Ratner, 2005; Minkovitz et al., 2005). Such outcomes may be analyzed via Poisson regression, a specific version of the generalized linear model (Nelder & Wedderburn, 1972). However, analysis via the Poisson model can be challenging due to, for example, the problem of overdispersion (Gardner et al., 1995) and, therefore, analysis via alternatives such as the negative binomial model is common.

Count outcomes can also be analyzed via the linear model. It has been suggested that the linear model should be relatively robust for analysis of count data given that the linear model is less sensitive to overdispersion (Knief & Forstmeier, 2021). As a specific empirical example, large-scale genetic association studies have used linear approaches to analyze associations between substance use and single nucleotide polymorphisms (Liu et al., 2019). Such analysis is common, and it has been previously noted that "the majority of published research on addictive behaviors continues to report analyses based on ordinary least squares" (Neal & Simons, 2007, p. 441). Or, consider related guidance in a different field: "Surprisingly, despite the ubiquity of count data in linguistics, Poisson regression is used only very little, and most statistics textbooks targeted at linguists do not even mention the approach" (Winter & Bürkner, 2021, p. 2). Here, we illustrate a potentially crippling issue that arises when using the linear model with count data. When data are generated via the Poisson model, spurious interactions can arise as a function of the magnitudes of the main effects of predictors. This issue is resolved when the appropriate model is used but, given the known difficulties associated with application of the Poisson model, caution will be required in interaction analyses of this type.

Censored Outcomes

An observation is censored if it is only partially known; here, we consider censoring of continuous variables. Classic examples of censoring include, for example, time-delimited observations of survival. In the psychological context, censored outcomes can occur when a scale is limited and values above or below a threshold are effectively capped. In many cases, censoring can be readily observed (by, for instance, simply viewing a histogram). Scores constructed via latent variable models can also be censored; for example, indices of academic ability can have ceilings or floors leading to limited information on significant proportions of observations (e.g., Koedel & Betts, 2010). Ceiling effects may be particularly severe in the context of dementia screening instruments used in neuropsychological settings and the context of minimum competency assessments used in statewide educational testing programs, both of which have been specifically designed to be sensitive to borderline and impaired or delayed levels of performance. Similarly, mental health scales designed for use in clinical populations, such as depressive symptoms scales (Djukanovic et al., 2017; Page et al., 2007), may exhibit floor effects, especially when used in samples from the general population. Floor effects are also commonly observed in biomarker research, such as hormone research, in which nontrivial portions of participants exhibit biomarker levels that are below the lower limits of detection (Grotzinger et al., 2018). Censored data may be modeled via the Tobit model (Tobin, 1958). Commonly, however, the censoring is ignored and the linear model is be used.

One challenge to discussion of censoring is that large quantities of research do not report sufficient information (e.g., density plots or histograms) of key outcomes so that an assessment of censoring can be undertaken. Thus, while censoring may exist in a large number of cases, and can often be diagnosed by the researchers who conduct the primary analyses, it is difficult for readers of published research to independently verify. As a consequence, it is hard to know how often the linear model is used with censored variables. We can sometimes infer its existence, however. For example, a gene-by-environment interaction analysis (de Castro-Catala et al., 2017) considered the degree to which depressive symptoms may be moderated by the predictors of interest. Depressive symptoms are based on sum scores of symptoms on a self-report questionnaire; based on the reported descriptive statistics, we estimate that nearly 10% of the observations would have been censored below.[2] This is reason for concern. When censored outcomes are analyzed via the linear model (which ignores censoring), even relatively low levels of censoring can lead to the identification of spurious interaction estimates.

Noninterval Outcomes

Many psychological constructs are measured via aggregating scores across multiple tests or item responses to produce composite indices. For example, the number of correct responses on an achievement test or the number of symptoms indicated on a depression scale can be summed to produce a composite score, or aggregated using more advanced methods (e.g., empirical Bayes) for estimating factor scores in the context of factor analytical or item response theory models (van der Linden, 2016). Previous work has commented on the challenges associated with identification of interactions when working with such latent variables (Embretson, 1996; Kang & Waller, 2005). We focus on the fact that neither the raw sum scores nor the outcomes produced by latent variable models necessarily have interval scales. The interval interpretation allows us to suppose that a one-unit change across the scale consistently has the same meaning. This interpretation is valid for many physical measures (e.g., a 1-m change in length always means a change by a standard amount) but does not necessarily hold for psychological measures (Michell, 2008). Data may have structure necessary for interval interpretations (Domingue, 2014), but in many cases we do

[2] A score of zero would be equivalent to a z-score of -1.3 given the mean and SD for the SCL-90-R (Table 1 in de Castro-Catala et al., 2017). Alternatively, the measure could be heavily skewed but that introduces other problems for interaction studies (Domingue et al., 2022).

not have strong evidence of such structure. A failure to have this equal interval property can have implications for a variety of scale uses (e.g., Ballou, 2009).

In particular, this lack of an interval scale can lead to identification of spurious interactions. For instance, when a test has an unbalanced distribution of easy and difficult items (relative to the ability distribution of the sample), the relationship between the latent ability and the composite index may become nonlinear (see Figure 3 of Tucker-Drob, 2009), and, subsequently, a spurious interaction may result (see Figure 4 of Tucker-Drob, 2019). Here, we consider a concerning example wherein a spurious interaction is introduced due to the noninterval scale. Similar to spurious interactions due to censoring, it is generally challenging, if not impossible, to know the degree to which the scale departs from the equal interval assumption. Given that the degree of departure is typically unknown, appropriate analysis is typically impossible. Thus, in this case, practitioners need to be especially wary about the problem because it is not easy to diagnose and fix via alternative modeling strategies.

Key Contributions

We focus on describing bias and Type I error across the four outcomes, which we divide into two categories. We first examine noncontinuous outcomes (binary and count outcomes). We show that application of linear modeling techniques in such a scenario results in disastrous levels of Type I errors. Second, we consider continuously measured outcome variables that do not meet classical assumptions of the linear model (censored and noninterval outcomes). We show that such outcomes—when viewed as transformed versions of classic linear outcomes—similarly produce undesirable levels of Type I error when analyzed with the linear model. We also provide geometric and algebraic guidance for why Type I error rates are so high. Evidence from both settings suggests a need for heightened scrutiny of the characteristics of the outcome in interaction studies.

Methods

Running Example

We consider a running example so as to describe results in a common way. We suppose that we observe some outcome y and interest is in predictors x and z as well as their interaction. We'll assume that $(x_i, z_i) \sim \mathrm{MVN}(\mu, \Sigma)$ where $\mu = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$ and $\Sigma = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$. We focus on $\rho = 0$ so as to clarify that the issue is about the product term even if x and z are not associated. We define a quantity, Γ, that plays a key role throughout. This quantity is defined as

$$\Gamma_i = \beta_0 + \beta_1 x_i + \beta_2 z_i + \beta_3 x_i z_i. \tag{1}$$

In interaction studies, interest is typically in estimates of β_3. Specific choices for the relevant parameters (i.e., $\beta_0, \beta_1, \beta_2$) and sample size ($N$) are documented throughout but we view them as suggestive; they are chosen to help us illustrate the pronounced problems of bias and false positives across many specific configurations of parameters.

If $\mathbb{E}(y_i \mid x_i, z_i) = \Gamma_i$, then the linear model (LM) is appropriate for estimation of the β parameters. We focus here on cases wherein $\mathbb{E}(y_i \mid x_i, z_i) \neq \Gamma_i$. We first consider transformations of Γ appropriate for discrete outcomes. We do this by supposing that $\mathbb{E}(y_i \mid x_i, z_i) = g(\Gamma_i)$ for different g. Suppose, for example, that we are interested in binary outcomes; in that case, we may assume that $\mathbb{E}(y_i \mid x_i, z_i) = \sigma(\Gamma_i)$ where σ is the logistic sigmoid, $\sigma(x) = 1/(1 + e^{-x})$. We consider outcomes belonging to the family of generalized linear models (GLM; Nelder & Wedderburn, 1972) with nonidentity link (binary and count outcomes). We then suppose that $\mathbb{E}(y_i^\star \mid x, z) = \Gamma_i$ but instead of observing $y^\star$ we observe some transformed version y (i.e., $g(y_i^\star) = y_i$ for some transformation g). We focus on transformations g that induce metric limitations in y_i.

Four Approaches for Illustrating Bias and False Discovery

We take four approaches to illustrating bias and false discovery.

1. The first two approaches are based on visualizations for select values of the relevant parameters. We first consider figures that plot y or $\mathbb{E}(y \mid x, z)$ as a function of xz. Such figures capture the breakdowns in symmetry that induce false positives.

2. We then consider plots showing y or $\mathbb{E}(y \mid x, z)$ for select values of z over the range of x. These plots illustrate the fact that the best fit conditional regression lines for different values of z are not parallel (thus, indicating an interaction effect) as we vary the key parameters.

3. We next consider Taylor series expansions of the true model. Taylor series expansions allow for representation of a function as a function of an infinite sum of the function's derivatives. Use of this technique requires certain smoothness conditions on the function (and thus won't be applicable in all cases). However, where applicable, this technique allows us to offer additional insight into the algebra leading to Type I errors. Derivations of key expansions are shown in Appendix A. In addition, these expansions reveal how product terms end up in the linear approximation even when they are not in the true generating model, which is useful in understanding the extent which the issues raised in this article are specifically about challenges to the identification of interactive effects (see Appendix B).

4. Finally, to demonstrate the ubiquity of the problem, we calculate bias and Type I error rates via simulation studies as a function of the relevant model parameters and sample size.

This sequence allows us to visualize the geometry of correlations between xz and the outcome, illustrate the differences in linear fits that suggest the existence of an interaction, offer an algebraic rationale for why this comes about, and provide guidance on how large a problem this is under different conditions.

Simulation Details

In all cases, interest will focus on estimates of β_3. We denote such estimates as $\widehat{\beta_3}$ when the correct model is used, and $\widehat{\beta_3^{LM}}$ when the LM is used but emphasize that in fact $\beta_3 = 0$ (i.e., we always study false positives). We consider sample sizes of 250 to 1,000.[3] Note that the problems identified for samples of 1,000 would only be magnified in larger samples. For any given configuration of conditions, we consider results from analysis of 10,000 simulated data sets. All analyses were conducted in R. Estimation of linear, logistic, and Poisson models was done via base R functions; estimation of Tobit models via a tailored package (Henningsen, 2020). Code to replicate all analyses is available on GitHub.[4]

Results

Transformations of Γ

Binary Outcomes

We begin with a discussion of binary outcomes. We embed some additional didactic components into this discussion given that the remaining outcome types will follow a similar format. Suppose y_i is a binary outcome; we suppose that y_i is a Bernoulli random variable with $\Pr(y_i = 1) = (1 + \exp(-\Gamma))^{-1}$. That is, we assume that data are generated via the logistic model but the key observation regarding false discovery and the LM is not sensitive to this fact (e.g., data could be generated from the Probit model). Such data may be analyzed either logistic regression (i.e., the GLM with the logistic link) or via the LM. The latter approach, commonly known as the "linear probability model," is based on ignoring the fact that y_i is binary. Use of the LM in this case makes the simplifying assumption that $y_i \sim \text{Normal}(\Gamma_i, \sigma_y^2)$. Why make this assumption? There are several rationales. Application of the LM may be useful given that parameter estimates are straightforward to interpret and there may also be computational reasons for its use (e.g., estimation in the context of relatively large data sets with hierarchical structure can become computationally expensive and the LM may help to offset this fact).

While there may be merit in application of the LM with binary outcomes in some cases, we emphasize one key point. Interaction analysis with the LM may be a flawed approach depending on the prevalence of the outcome. We have discussed this issue in examples related to gene–environment interactions elsewhere (Domingue et al., 2020; Trejo et al., 2018) but here show it in its general form. We focus specifically on the problem induced by changes in the intercept (β_0), which is related to the prevalence of the outcome (i.e., given that x and z have zero mean, $\mathbb{E}(y) = \frac{1}{1+e^{-\beta_0}}$). Figure 1 considers a scatterplot of $\mathbb{E}(y \mid x, z)$ and xz. When $\beta_0 = 0$, there is a clear structure to the attached plot but we emphasize that there is no linear association between $\mathbb{E}(y \mid x, z)$ and xz (note the small correlation). However, as β_0 increases, the symmetry in the scatterplot breaks down; the resulting asymmetry leads to a large correlation (upper left in each panel) and thus false positives (i.e., spurious interaction estimates).

Why does this occur? To illustrate the underlying geometry driving this correlation, Figure 2 shows $\mathbb{E}(y \mid x, z)$ as a function of x along with the implied logistic (solid lines) and linear (dashed lines) fits for two different values of z ($z = \pm 1$). When $\beta_0 = 0$, the geometry of the $\mathbb{E}(y \mid x, z)$ values is such that both the logistic and linear fits for $z = \pm 1$ are clearly parallel (thus, indicating $\beta_3 = 0$). However, as β_0 increases, both the slope and intercept of the linear fit exhibit a dependence on z, and the geometry of the $\mathbb{E}(y \mid x, z)$ values is such that there is an implied interaction. This is due to the fact that the $z = 1$ values in blue are observed at a location where the true underlying sigmoid is quite flat; thus, leading to an inference that $\beta_3 < 0$. Note that when $\beta_0 = 4$, the fit line for $z = 1$ has become effectively horizontal; as β_0 continues to increase, the red line will begin to take the same shape and the spurious interaction under the linear model will diminish (we will observe this similar pattern in Figure 3).

To complement the geometric presentation in Figure 2, we now consider an algebraic illustration of the problem. Suppose that $\mathbb{E}(y \mid x, z) = \sigma(\beta_0 + \beta_1 x + \beta_2 z)$, where $\sigma(\cdot)$ is the standard logistic sigmoid. Using a Taylor series expansion, we can rewrite the right-hand side:

$$\mathbb{E}(y \mid x, z) = \frac{1}{2} + \frac{1}{4}(\beta_0 + \beta_1 x + \beta_2 z) - \frac{1}{48}(\beta_0 + \beta_1 x + \beta_2 z)^3 + \cdots. \tag{2}$$

After expanding the cubic term and multiplying by the leading coefficient, there is an interaction term including x and z: $-\frac{1}{8}\beta_0\beta_1\beta_2 xz$. By modeling this expectation as $\mathbb{E}(y \mid x, z) = b_0 + b_1 x + b_2 z + b_3 xz$, we should expect false discovery of an interaction effect due to this term. That is, we might anticipate $b_3 = -\frac{1}{8}\beta_0\beta_1\beta_2 xz \neq 0$ if we omit higher-order terms. Moreover, if $\beta_1\beta_2 > 0$, we would anticipate $\widehat{\beta_3^{LM}} < 0$ which is indeed what we observe in Figure 3. Balanced classes (or $\beta_0 = 0$) improve the situation but do not fully solve this problem, as the cubic term in the Taylor series still expands to contain the higher order interaction terms $\frac{1}{16}\beta_1^2\beta_2 x^2 z$ and $\frac{1}{16}\beta_1\beta_2^2 xz^2$ that the LM may attribute to an interaction between x and z.

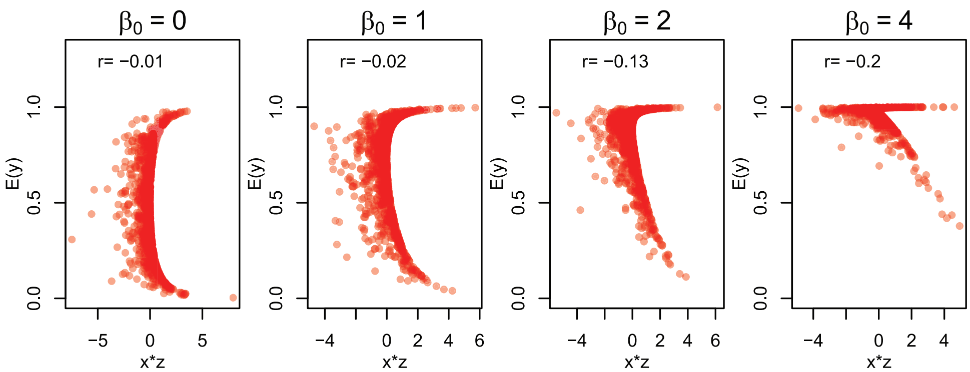
With both this geometric and algebraic intuition at hand, we focus on the associated bias in estimates from the LM. We simulate data via $\Pr(y_i = 1) = (1 + \exp(-\Gamma))^{-1}$ and based on additional notes discussed in Simulation Details. Crucially, we set $\beta_3 = 0$ in all simulations such that there is never a true interaction. We denote the LM-based estimates of the interaction coefficient as $\widehat{\beta_3^{LM}}$; interest is in this estimate as a function of variation in β_0. Given that x and z are centered at 0, expected prevalence of the outcome is $\frac{1}{1+e^{-\beta_0}}$ (e.g., expected prevalence is .88 for $\beta_0 = 2$). In

[3] To offer context for this choice of sample size, we surveyed 150 studies published between January 1, 2016 and October 20, 2020 focusing on depression (based on the Pubmed query: `(depression[Title]) AND (heterogeneity[Title] OR moderation[Title] OR interaction[Title]))`. Of the ascertained studies, 30% were empirical studies along the lines of what we consider here. The sample sizes ranged widely (min = 22, median = 576, IQR = 292–2,279, max = 134,357); our choice of 250–1,000 is meant to reflect the center of this distribution.

[4] See https://github.com/ben-domingue/interaction-problems. Note also that an early version of this manuscript was available as a preprint.

Figure 1

Scatterplot of $\mathbb{E}(y|x,z)$ and xz for Different Values of β_0 ($\beta_1 = \beta_2 = 1, \beta_3 = 0, N = 1{,}000$) When y is a Binary Outcome



Note. See the online article for the color version of this figure.

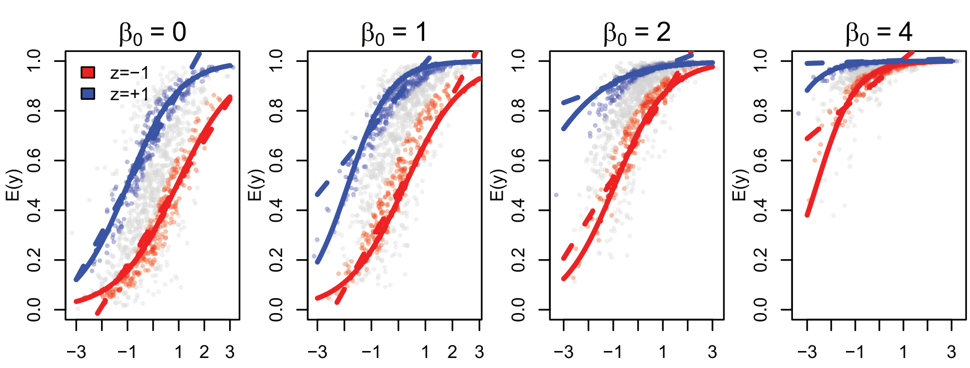
the left panel of Figure 3, we can see that LM estimates of β_3 are unbiased when $\beta_0 = 0$ but heavily biased when $\beta_0 \approx 2$; in our simulated data sets, the 10th through 90th percentiles of estimates are far from zero. As suggested above in Figure 2, this bias decreases for large values of β_0 (but note that the outcome will become extremely rare in such scenarios). Note that the degree of bias is independent of sample size.

The right panel of Figure 3 shows the implications in terms of Type I error for two selected values of β_0. When $\beta_0 = 0$, levels of false discovery are appropriate (as expected given the right panel). However, for $\beta_0 = 2$, false discovery is uniformly problematic and increasing sample size only increases the level of false discoveries. Here, we also consider GLM estimates; note that they have uniformly correct Type I error (given $\alpha = .05$). Selection of the correct link function here (rather than relying on the linear model) resolves the problem.

We end this section by emphasizing that false discovery when using the linear model to detect interactions effects on unbalanced binary outcomes is a fundamental problem. The parameter bias introduced by modeling interactions for binary outcomes using a linear model is not resolved by increasing sample size (to the contrary, increasing sample size will typically increase the Type I error rate). Further, the degree and direction of bias will depend upon specifics regarding the main effects β_1 and β_2 (if $\beta_1\beta_2 < 0$ then we will observe $\widehat{\beta_3^{LM}} > 0$ in contrast with what we observed in Figure 3). We begin to unpack these dependencies in the left panel of Figure 4. When $\beta_1 = \beta_2 = 1$ and prevalence is above .75, bias leads to elevated Type I error rates across all sample sizes considered. When samples are relatively large, bias is a problem even for prevalence approaching .5. The fact that larger samples lead to higher levels of Type I error is clearly apparent in the fact that the red regions extend lower for a given vertical strip when sample size is larger. To show that there is also sensitivity to other parameters, we consider $\beta_1 = .25$ and $\beta_2 = .75$ in the right figure panel. Here, bias leads to less problematic Type I error for smaller samples no matter the underlying prevalence. Thus, rather than

Figure 2

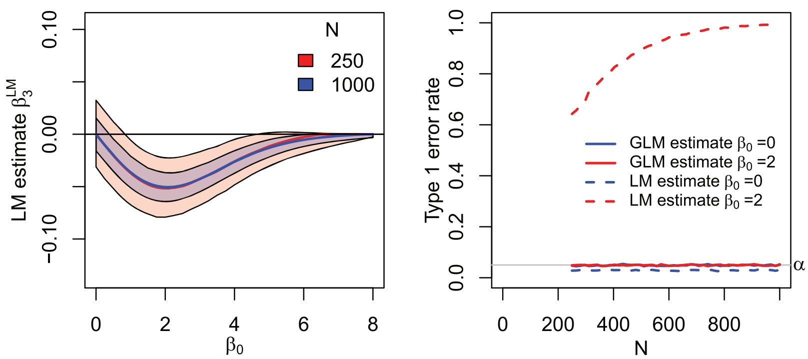
Illustration of the Geometry Driving False Discovery Due to Variation in β_0 When the Linear Model is Used ($\beta_1 = \beta_2 = 1$, $\beta_3 = 0$, $N = 1{,}000$) for Analysis of Binary Outcomes



Note. Blue and red dots represent those data points within half a unit of their respective z values (i.e., z such that $|z-1| < 0.5$ are in blue and z such that $|z+1| < 0.5$ are in red. Fitted lines for ± 1 are similarly shaded). Solid lines are fits from logistic regression model while dashed lines are fits from linear model. See the online article for the color version of this figure.

Figure 3

Variation in Bias (Left) and Type I Error Rate (Right) Associated With Estimates of β_3 When $\beta_3 = 0$ for Binary Outcomes (for $\beta_0 = 0$, $\beta_1 = \beta_2 = 1$)



Note. Left: Estimates $\widehat{\beta_3^{LM}}$ based on the LM for two sample sizes as a function of β_0. Shaded regions capture span of estimates between 10th and 90th percentile while the solid line shows median estimates. Right: Levels of Type I error as a function of sample size for two values of β_0. See the online article for the color version of this figure.

providing guidelines or rules of thumb regarding when a LM may be appropriate, we generally encourage the implementation of more appropriate link functions (e.g., logistic or probit).

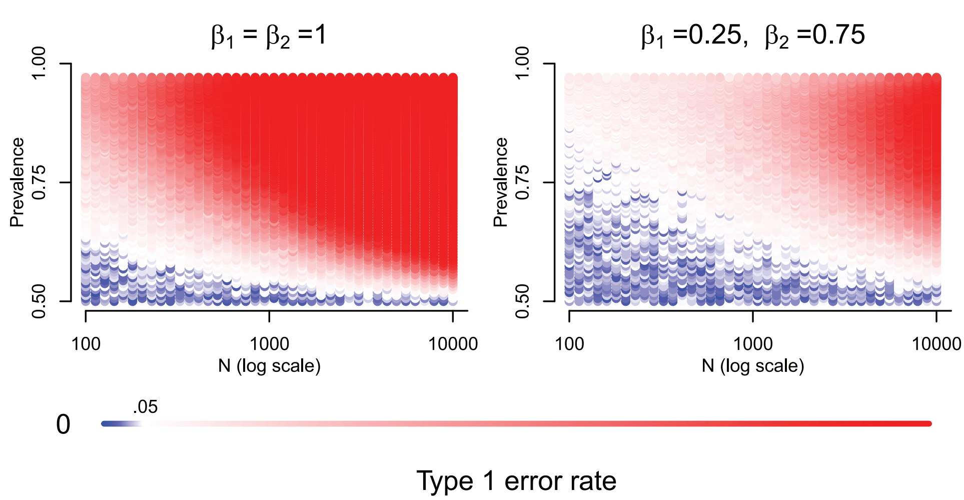
Count Outcomes

We now pivot to another outcome in the GLM framework and focus on the scenario where y_i is a count outcome. We suppose that data are generated via the Poisson model; thus, $\mathbb{E}(y_i \mid x, z) = \exp(\Gamma_i)$. False discovery is again a catastrophic problem if the linear model is used in this setting. We begin to illustrate this fact via Figure 5. We first show a scatterplot of $\mathbb{E}(y \mid x, z)$ and xz; however, this time we introduce variation in β_2 to emphasize the fact that spurious interactions can arise as a function of the main effect coefficients in this case. Note that there is no linear correlation apparent in the scatterplot when β_2 is relatively small. However, increase in β_2 induces an association despite the fact that $\beta_3 = 0$. We illustrate the geometry leading to this false discovery in Figure 6. We continue to vary β_2 and hold $\beta_1 = .2$. As β_2 increases, seeming variation in $\mathbb{E}(y \mid x, z)$ as a function of z is introduced. This would imply $\beta_3 \neq 0$ in the incorrectly specified linear approach.

As an algebraic illustration, we can again consider a Taylor series expansion. Suppose that $\mathbb{E}(y \mid x, z) = \exp(\beta_1 x + \beta_2 z) = \exp(\beta_1 x)\exp(\beta_2 z)$. We can use Taylor series expansion to write this as

Figure 4

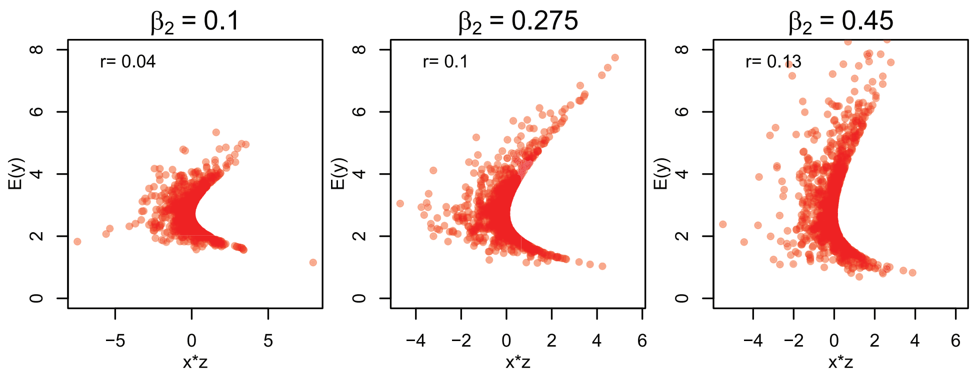
Type I Error Rate as a Function of Sample Size and Prevalence When Using LM to Analyze Binary Outcomes



Note. In all cases, $\beta_3 = 0$ while β_0 varies (such that stated prevalence is equivalent to $\frac{e^{\beta_0}}{1+e^{\beta_0}}$). The β_1 and β_2 coefficients are as shown for each panel. Blue regions indicate areas where Type I error is appropriate whereas red indicates regions wherein bias is leading to elevated levels of Type I error. See the online article for the color version of this figure.

Figure 5

Scatterplot of $\mathbb{E}(y\,|\,x,z)$ and xz for Different Values of β_2 ($\beta_0 = 1, \beta_1 = 0.2, N = 1{,}000$) When y_i is a Count Outcome



Note. See the online article for the color version of this figure.

$$\mathbb{E}(y|x,z) = (1+\beta_1 x+(\beta_1 x)^2/2+\cdots)(1+\beta_2 z+(\beta_2 z)^2/2+\ldots)$$
$$= 1+\beta_1 x+\beta_2 z+\beta_1\beta_2 xz+\ldots. \quad (3)$$

If we mistakenly assume that $\mathbb{E}(y\,|\,x,z) = b_1x + b_2z + b_3xz$, we would anticipate $b_3 = \beta_1\beta_2$ if we omit higher order terms. Thus, if either β_1 or β_2 does not equal zero, false discovery will result when the LM is deployed for analysis.

Implications of this problem for bias and false discovery are shown in Figure 7. At left, bias in $\widehat{\beta_3^{LM}}$ increases as a function of β_2; the performance of the LM for analysis of count outcomes will be sensitive to the magnitude of the main effects. At right, we observe that the bias introduced by β_2 induces Type I errors. Analysis of larger samples will be plagued by false positives in such an instance. However, analysis of the appropriately specified GLM (i.e., Poisson regression) yields the correct Type I error rate. Note that we see higher than expected levels of false discovery by the LM even when $\beta_2 = 0$, but the false discovery rate is unaffected by sample size. While the Taylor series expansion suggests we should not see false discovery when $\beta_2 = 0$, the transformation of Γ_i induces heteroskedasticity in the outcome. This leads the standard LM to underestimate standard errors, which manifests as false discovery.

Transformations of $y^{\star}$

Censored Outcomes

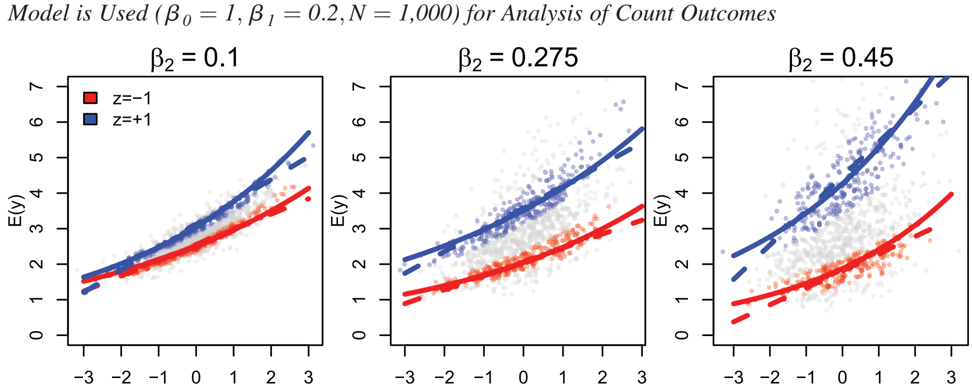
We first consider transformations g of $y^{\star}$ (where $\mathbb{E}(y_i^{\star}\,|\,x,z) = \Gamma_i$) that lead to scales with a ceiling or floor (Garin, 2014). We focus on a floor but the same concerns would apply to ceilings. To simulate data, we utilize the Tobit model (Tobin, 1958). We suppose that

$$y_i^{\star} \sim \mathrm{N}(\Gamma_i, \sigma^2_{y^{\star}}). \quad (4)$$

However, we do not observe $y_i^{\star}$; rather, we observe

Figure 6

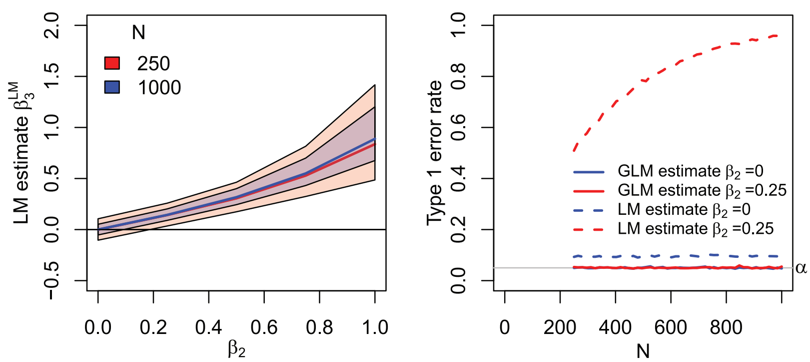
Illustration of the Geometry Driving False Discovery Due to Variation in β_2 When the Linear Model is Used ($\beta_0 = 1, \beta_1 = 0.2, N = 1{,}000$) for Analysis of Count Outcomes



Note. Blue and red dots represent those data points within half a unit of their respective z values (i.e., z such that $|z-1| < 0.5$ are in blue and z such that $|z+1| < 0.5$ are in red. Fitted lines for ± 1 are similarly shaded). Solid lines are fits from Poisson regression model while dashed lines are fits from linear model. See the online article for the color version of this figure.

Figure 7

Variation in Bias (Left) and Type I Error Rate (Right) Associated With $\widehat{\beta_3^{LM}}$ When $\beta_3 = 0$ for Count Outcomes (for $\beta_0 = 0, \beta_1 = 0.5$)



Note. Left: Estimates of β_3 based on the LM for two sample sizes as a function of β_2. Shaded regions capture span of estimates between 10th and 90th percentile while the solid line shows median estimates. Right: Levels of Type I error as a function of sample size for two values of β_2. See the online article for the color version of this figure.

$$y_i = g(y_i^\star) = \begin{cases} y_i^\star & y_i^\star > c \\ c & y_i^\star \leqslant c \end{cases} \tag{5}$$

for some constant c. Note that the censoring violates the smoothness conditions required to conduct the Taylor series expansion. Thus, we do not consider such an analysis here.

We begin to illustrate the reason for Type I errors when the LM is deployed due to the existence of floors in Figure 8. We show y as a function of xz in gray versus $y^\star$ in red. When the floor is sufficiently low, few points are affected and recovery with the censored data yields the correct inference. This is due to the fact that we observe the full shape of (xz, y). However, when the floor begins to censor a significant number of cases, we no longer observe the bottom half of the shape. When these points are raised toward the y mean, they are down-weighted in the sum leading to the correlation coefficient and thus lead to the resulting (spurious) positive correlation between xz and y.

The implications for fitted trajectories for different values of z are shown in Figure 9. Trajectories for relatively large values of z are unaffected (i.e., the solid and dashed blue lines are similar in all cases). However, as c increases, there is an increasing effect on the red line; when c is at its largest, the trajectory based on observed values is flatter than the trajectory based on uncensored values thus suggesting an interaction under the naive LM.

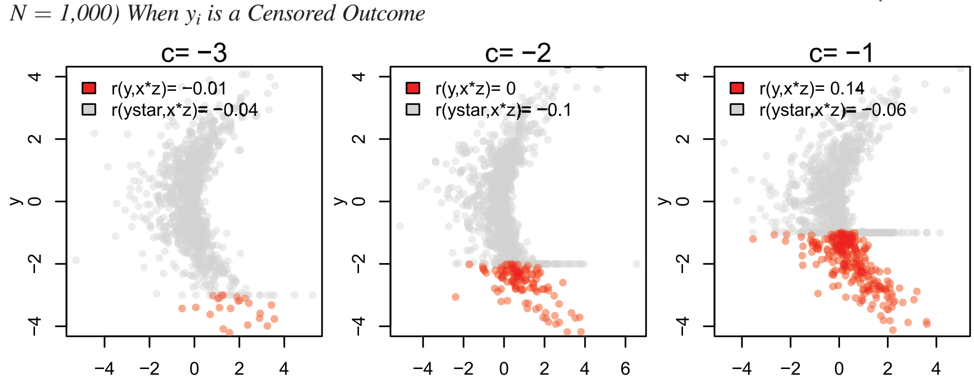
Figure 10 extends this geometric logic to an analysis of bias and Type I error as a function of c. On the left, we first consider LM-based estimates $\widehat{\beta_3^{LM}}$. As c increases, so does bias in $\widehat{\beta_3^{LM}}$. This translates into an increase in Type I error which we consider on the right. Bias leads to highly elevated levels of false discovery even for modestly sized samples and the relatively low values of c that would not result in substantial amounts of censoring. The bias leads to catastrophic Type I error as the censoring affects a larger proportion of outcomes (i.e., $c = -1$); in such a case, increasing sample size serves only to increase the salience of the problem. In contrast, Tobit estimates produce Type I errors at the expected rate for both values of c irrespective of sample size.

Noninterval Outcomes

As a final example, we consider the impact of distortions that lead to noninterval scales on our ability to make accurate

Figure 8

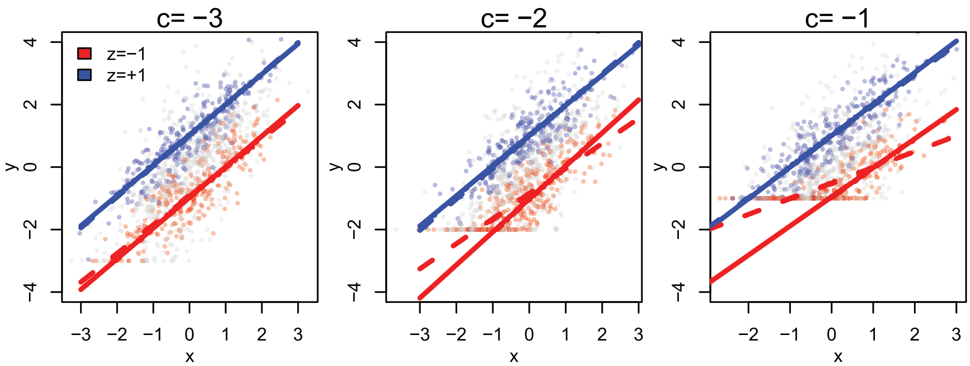
Scatterplot of y or $y^\star$ and xz for Different Values of c ($\beta_0 = 0, \beta_1 = \beta_2 = 1, \sigma_y^2 = 0.25$, $N = 1{,}000$) When y_i is a Censored Outcome



Note. See the online article for the color version of this figure.

Figure 9

Illustration of the Geometry Driving False Discovery Due to Variation in c When the Linear Model is Used ($\beta_0 = 0, \beta_1 = \beta_2 = 1, \sigma_y^2 = 0.25, N = 1{,}000$) for Analysis of Censored Outcomes



Note. Blue and red dots represent those data points within half a unit of their respective z values (i.e., z such that $|z-1|<0.5$ are in blue and z such that $|z+1|<0.5$ are in red. Fitted lines for ±1 are similarly shaded). Solid lines are fits from Tobit regression model while dashed lines are fits from linear model. See the online article for the color version of this figure.

inferences. To do this, we consider a monotonic but nonlinear transformation of $y_i^\star$. Such a transformation is motivated by, for example, the work of Michell (1986) which surfaces the issue of measurement scales in the context of psychological constructs. Scales are frequently assumed to be interval but need not be. In particular, we consider “Lord’s transformation” (Briggs & Betebenner, 2009) which stretches the scale for larger values of $y_i^\star$. That is, we suppose that $\mathbb{E}(y_i^\star \mid x, z) = \Gamma_i$ but that we observe

$$y_i = g(y_i^\star) = 1.05^{(y_i^\star - \alpha)/\lambda}. \tag{6}$$

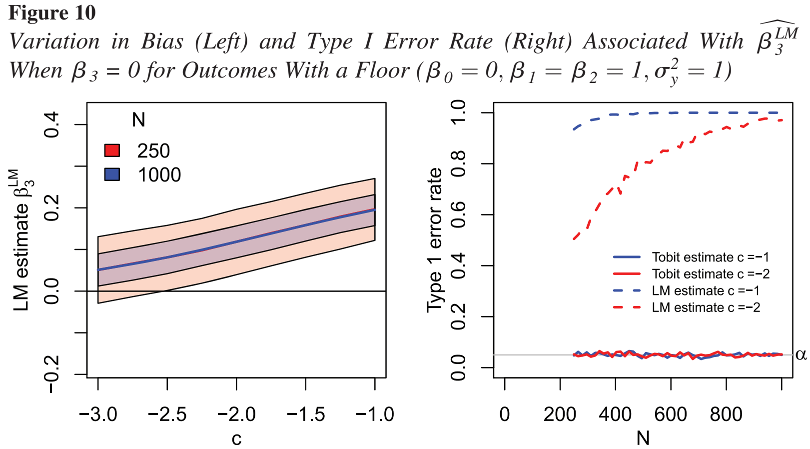
Here, we consider α = 0 and allow λ to vary. This transformation has the effect of making a one-unit difference in y have differential meaning in the $y^\star$ metric (see illustration in Figure 11). When λ is small, distances between smaller values of $y^\star$ are compressed while distances between larger values are inflated. Differences between the distances quickly dissipate as λ increases. The transformation’s effect has similarities to the effect of a floor in that, in both cases, differences between small values of $y^\star$ are being minimized (i.e., false positives are generated here for the same reason as in Figure 8). Given these similarities, we provide additional figures in the online supplemental materials but do not discuss them further.

First, we look to a Taylor series expansion of Equation 6 about zero. Supposing that α = 0, we have $\mathbb{E}(y \mid x, z) = 1.05^{\frac{1}{\lambda}\Gamma}$. Looking at the first three terms, we see:

$$\begin{aligned}\mathbb{E}(y \mid x, z) &= 1.05^{\frac{1}{\lambda}\Gamma} \\ &= 1 + \left(\ln 1.05 \frac{1}{\lambda}\right)\Gamma + \frac{1}{2}\left(\ln 1.05 \frac{1}{\lambda}\right)^2 \Gamma^2 + \cdots.\end{aligned} \tag{7}$$

Figure 10

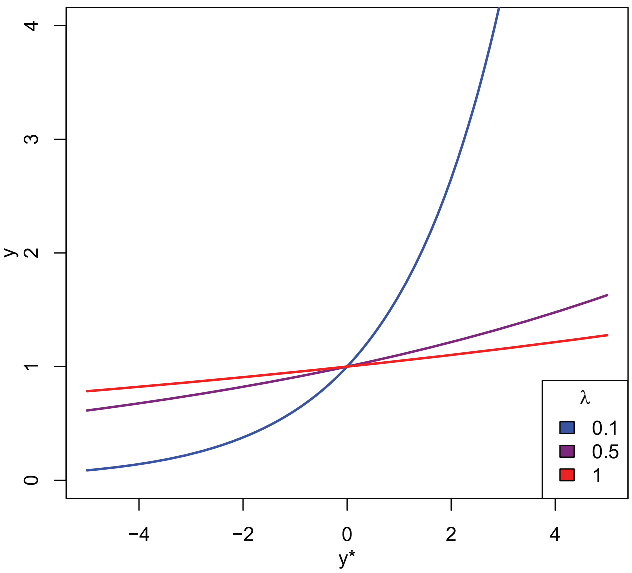
Variation in Bias (Left) and Type I Error Rate (Right) Associated With $\widehat{\beta_3^{LM}}$ When $\beta_3 = 0$ for Outcomes With a Floor ($\beta_0 = 0, \beta_1 = \beta_2 = 1, \sigma_y^2 = 1$)



Note. Left: Estimates of β_3 based on the LM for two sample sizes as a function of c. Shaded regions capture span of estimates between 10th and 90th percentile while the solid line shows median estimates. Right: Levels of Type I error as a function of sample size for two values of c. See the online article for the color version of this figure.

Figure 11

Effect of Transformation When α = 0 for Various Values of λ



Note. See the online article for the color version of this figure.

Expanding the third term, we produce a term of $\frac{\ln^2 1.05}{\lambda^2}\beta_1\beta_2 xz$. If $\beta_1\beta_2 \neq 0$; this term will thus lead to false discovery of an interaction coefficient when the LM is used. Given that λ appears in the denominator, this problem is exacerbated at smaller values of λ.

Figure 12 focuses on the bias and Type I error induced by this transformation. We do not consider estimates that are adjusted for the outcome's specific distributional properties because the specific function governing how the observed data map to an interval scale is typically not something that is directly known (Domingue, 2014). This increases the salience of the following observation about the LM estimates because they would likely be used in place of a hypothetical alternative that adjusts for the role of Lord's transformation. We can see a clear sensitivity in estimates of β_3 to the value of λ. For small values of λ, there are larger deviations from the true interval scale which leads to an increase in bias. This then translates into a higher level of Type I error rate when λ is relatively small. Implementing an ordinal link function (Bürkner & Vuorre, 2019)—for example, an ordered logit—may be worth considering in this scenario.

Discussion

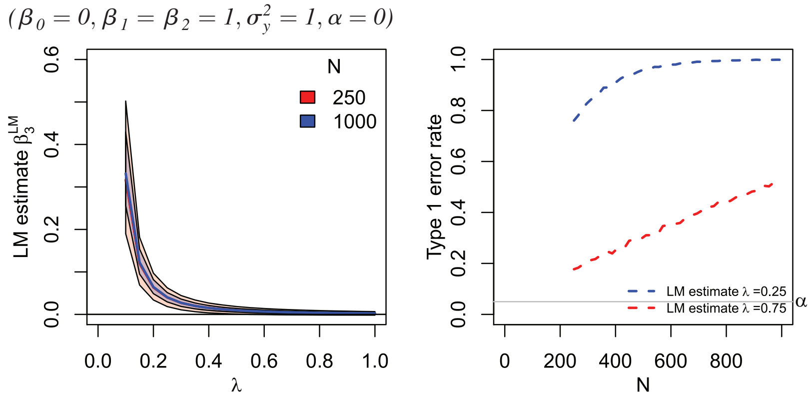
The study of interactions is of clear interest in a wide range of settings. However, the specific properties of the outcome variable in question can make accurate inference challenging. To illustrate this point, we considered two broad types of outcomes: outcomes based on a transformation of Γ and outcomes based on a transformation of some intermediate variable $y^\star$ where $\mathbb{E}(y_i^\star \mid x, z) = \Gamma_i$. In both cases, there are a variety of problems that may threaten inference; some of these issues are relatively easy to diagnose and resolve, others less so. Here, we separately discuss implications for interaction analysis of both types of outcomes before concluding with general thoughts on the study of interactions.

For outcomes based on transformations of Γ, there are clear gains to using the correctly specified alternative. In all cases, the correctly specified approach led to correct levels of Type I error whereas the LM-based approach led to highly elevated levels of Type I error. There are considerations—including computational feasibility, interpretability, and other concerns (Battey et al., 2019; Gomila, 2020)—that may lead one away from the GLM approaches. In certain scenarios (e.g., binary outcomes with prevalences near 50%), the benefits of the linear model may be worth considering (Knief & Forstmeier, 2021). However, the potential for spurious interaction results from the linear model should be seriously considered by researchers making model specification decisions. We would urge heightened scrutiny when attempting to model interaction effects using alternatives that may be incorrectly specified.

For outcomes based on transformations of $y^\star$, our results suggest that common problems will serve to increase Type I error. Some

Figure 12

Variation in Bias (Left) and Type I Error Rate (Right) Associated With $\widehat{\beta_3^{LM}}$ When $\beta_3 = 0$ for Outcomes Observed Following Lord's Transformation ($\beta_0 = 0, \beta_1 = \beta_2 = 1, \sigma_y^2 = 1, \alpha = 0$)



Note. Left: Estimates of β_3 based on the LM for two sample sizes as a function of λ. Shaded regions capture span of estimates between 10th and 90th percentile while the solid line shows median estimates. Right: Levels of Type I error as a function of sample size for two values of λ.

of these problems are perhaps fairly easy to identify (i.e., the existence of a floor) and can frequently be addressed using standard solutions (i.e., models for censored data). Other problems (i.e., noninterval scales) may be hard to identify and nevertheless have deleterious effects on statistical inference. There is no straightforward general mechanism for identifying and correcting this issue. We expect that in many instances, the primary analyst can employ their knowledge of the measurement protocol combined with content knowledge about the constructs of interest and data visualization (e.g., histograms, LOESS plots) to make an informed appraisal of quality of the measurement, and possible implications for interaction testing. Another possibility would be to consider analyses that only require the outcome be ordinal (Bürkner & Vuorre, 2019). The lack of a straightforward resolution is, in our view, a rationale for statistical humility when studying outcomes with challenging measurement properties of the kind that abound in social science. In particular, a single study should rarely be viewed as dispositive evidence regarding heterogeneity absent of very large samples and very high-quality measurement.

Much of what we discuss is not novel in the sense that these findings are anticipated by other studies focusing on power and false discovery in the context of misspecified models or other issues. That said, we view the issues of interactions as warranting special attention. Consider, for example, power studies that may be performed in proposals of future research. Our findings suggest the need for carefully constructed power analyses if the goal is to derive appropriate guidance about sample sizes, for example. Such studies may be constructed using recently developed techniques (Jaccard & Brinberg, 2021). However, also note that many of the problems identified here produce bias such that they will only become worse with larger samples. In other words, more data will not resolve these problems; informed decisions about appropriate measurement and modeling will require conceptual understanding. One potential response to these findings is that the problem of bias is easily avoided if correctly specified models are used. On the one hand, we agree and think this an underappreciated issue in the context of interaction studies. On the other hand, especially when considering latent variables as outcomes (a common occurrence in psychology), it may be nontrivial to identify the correct model.

Turning to the issue of false discovery, we make two key points. First, the issues of false discovery raised here cannot be dealt with by utilizing robust standard errors. The problems illustrated in Figures 2, 6, and 8 show that it is the parameter estimates themselves—not merely the standard errors—which are flawed (for a related argument about the limitations of robust standard errors, see King & Roberts, 2015). We have focused on Type I errors because they directly call attention to the high rates of false positives in interaction studies using outcomes that present measurement challenges. The second (closely related) point is that biased parameter estimates, like those observed here, may lead to a large quantity of spurious findings in research focusing on interactions. This is a concern and should lead to increased attention to the nature of the outcome variable in studies of statistical interactions.

We acknowledge limitations. We do not consider analysis using techniques such as structural equation modeling or mixed models. There are also outcome types that we do not consider (e.g., skewed outcomes; Domingue et al., 2022). In addition, we do not consider spline-based approaches that may allow for nonparametric analysis of outcome response surfaces as a (potentially nonlinear) function of predictors. Despite the fact that our simulations only speak conclusively to some cases, they demonstrate the need for increased caution in studies of interactions, especially when focal interest is on tests of statistical significance. Finally, some of the issues that we discuss may be more general problems that afflict main effect estimates in certain settings. Our Taylor series expansions help clarify how xz product terms are produced in linear approximations of the true nonlinear response surfaces (see Appendix B). Future research may use this technique to further probe the scenarios in which similar concerns arise when considering main effects.

What should researchers do about the problems highlighted in this article? First, it is important to reemphasize that the false discovery of interactions highlighted here are the results from parameter bias arising from model misspecification; these problems will not be resolved by increasing sample sizes (in fact, Figure 4 makes clear that increasing sample size will typically exacerbate Type I error rates). Such bias arises from a confluence of several data characteristics that are not easily reducible to consultation of standard rubrics or guidelines about when a linear model may be acceptable. Rather, we encourage researchers to implement link functions that may be better suited to the characteristics of their data (e.g., logistic or probit for binary and ordered categorical outcomes; Poisson for count outcomes; Tobit for censored outcomes that may otherwise be normally distributed). We stress that, while in some cases the appropriate link function may be challenging to identify, in many cases there will be available options for a link function that will resolve the concerns discussed here. Such link functions are universally available across statistical software. There are also several widely used software options (e.g., MPlus TYPE = COMPLEX and TYPE = TWOLEVEL, SAS PROC GLIMMIX and PROC NLMIXED, R lme4) for implementation of link functions in conjunction with other advanced modeling approaches when these issues co-occur with other data complexities (e.g., nesting of observations within clusters; the need to implement sample weights). In some cases, there nevertheless may be complexities of study design that require modeling adjustments that lead to interest in linear approximations. In such cases, we would urge both caution and a reliance on simulation studies to guide subsequent inferences. Such simulation studies would need to consider not just the sample size in conjunction with the metric properties of their outcomes, but also the magnitudes of main effects. Continuously distributed noninterval outcomes may be especially concerning as they are difficult to identify. When they are a potential concern, we suggest considering ordinal (Bürkner & Vuorre, 2019) approaches.

If the "some do, some do not" formulation holds true—and, we believe that it often does—interaction studies will clearly be of interest. Yet, poorly designed interaction studies can lead fields into dead-ends: Spurious interactions will be detected and true interactions will be obscured, reversed, or overestimated. Consider, for example, the era of candidate gene-by-environment studies (Duncan & Keller, 2011) which is now viewed as a vast literature consisting almost entirely of false positives. A failure to steer clear of dead-ends can lead to wasting large quantities of resources—both in terms of finite research dollars and even scarcer researcher time. Thus, we encourage researchers pursuing questions focused on studies of heterogeneity using statistical interactions to take a more realistic perspective on the quality of inference likely to result from their particular data context.

References

Ancelin, M.-L., Scali, J., Norton, J., Ritchie, K., Dupuy, A.-M., Chaudieu, I., & Ryan, J. (2017, March). Heterogeneity in HPA axis dysregulation and serotonergic vulnerability to depression. *Psychoneuroendocrinology*, *77*, 90–94. https://doi.org/10.1016/j.psyneuen.2016.11.016

Angosta, J., Steers, M.-L. N., Steers, K., Riggs, J. L., & Neighbors, C. (2019, November). Who cares if college and drinking are synonymous? Identification with typical students moderates the relationship between college life alcohol salience and drinking outcomes. *Addictive Behaviors*, *98*, Article 106046. https://doi.org/10.1016/j.addbeh.2019.106046

Ballou, D. (2009). Test scaling and value-added measurement. *Education Finance and Policy*, *4*(4), 351–383. https://doi.org/10.1162/edfp.2009.4.4.351

Battey, H., Cox, D., & Jackson, M. (2019). On the linear in probability model for binary data. *Royal Society Open Science*, *6*(5), Article 190067. https://doi.org/10.1098/rsos.190067

Briggs, D., & Betebenner, D. (2009). *Is growth in student achievement scale dependent*. Unpublished manuscript. https://www.cde.state.co.us/sites/default/files/documents/research/download/pdf/growthscaledependent.pdf

Bryan, C. J., Tipton, E., & Yeager, D. S. (2021). Behavioural science is unlikely to change the world without a heterogeneity revolution. *Nature Human Behaviour*, *5*(8), 980–989. https://doi.org/10.1038/s41562-021-01143-3

Bürkner, P.-C., & Vuorre, M. (2019). Ordinal regression models in psychology: A tutorial. *Advances in Methods and Practices in Psychological Science*, *2*(1), 77–101. https://doi.org/10.1177/2515245918823199

Claro, S., Paunesku, D., & Dweck, C. S. (2016). Growth mindset tempers the effects of poverty on academic achievement. *Proceedings of the National Academy of Sciences of the United States of America*, *113*(31), 8664–8668. https://doi.org/10.1073/pnas.1608207113

Coleman, J. R. I., Peyrot, W. J., Purves, K. L., Davis, K. A. S., Rayner, C., Choi, S. W., Hübel, C., Gaspar, H. A., Kan, C., Van der Auwera, S., Adams, M. J., Lyall, D. M., Choi, K. W., Dunn, E. C., Vassos, E., Danese, A., Maughan, B., Grabe, H. J., Lewis, C. M., . . . Breen, G. (2020). Genome-wide gene-environment analyses of major depressive disorder and reported lifetime traumatic experiences in U.K. Biobank. *Molecular Psychiatry*, *25*(7), 1430–1446. https://doi.org/10.1038/s41380-019-0546-6

Culverhouse, R. C., Saccone, N. L., Horton, A. C., Ma, Y., Anstey, K. J., Banaschewski, T., Burmeister, M., Cohen-Woods, S., Etain, B., Fisher, H. L., Goldman, N., Guillaume, S., Horwood, J., Juhasz, G., Lester, K. J., Mandelli, L., Middeldorp, C. M., Olié, E., Villafuerte, S., . . . Bierut, L. J. (2018). Collaborative meta-analysis finds no evidence of a strong interaction between stress and 5-HTTLPR genotype contributing to the development of depression. *Molecular Psychiatry*, *23*(1), 133–142. https://doi.org/10.1038/mp.2017.44

Dai, H., Saccardo, S., Han, M. A., Roh, L., Raja, N., Vangala, S., Modi, H., Pandya, S., Sloyan, M., & Croymans, D. M. (2021). Behavioural nudges increase covid-19 vaccinations. *Nature*, *597*(7876), 404–409. https://doi.org/10.1038/s41586-021-03843-2

de Castro-Catala, M., Peña, E., Kwapil, T. R., Papiol, S., Sheinbaum, T., Cristóbal-Narváez, P., Ballespí, S., Barrantes-Vidal, N., & Rosa, A. (2017, November). Interaction between FKBP5 gene and childhood trauma on psychosis, depression and anxiety symptoms in a non-clinical sample. *Psychoneuroendocrinology*, *85*, 200–209. https://doi.org/10.1016/j.psyneuen.2017.08.024

Djukanovic, I., Carlsson, J., & Årestedt, K. (2017). Is the hospital anxiety and depression scale (HADS) a valid measure in a general population 65–80 years old? A psychometric evaluation study. *Health and Quality of Life Outcomes*, *15*(1), 1–10. https://doi.org/10.1186/s12955-017-0759-9

Domingue, B. (2014). Evaluating the equal-interval hypothesis with test score scales. *Psychometrika*, *79*(1), 1–19. https://doi.org/10.1007/s11336-013-9342-4

Domingue, B. W., Kanopka, K., Mallard, T. T., Trejo, S., & Tucker-Drob, E. M. (2022). Modeling interaction and dispersion effects in the analysis of gene-by-environment interaction. *Behavior Genetics*, *52*(1), 56–64. https://doi.org/10.1007/s10519-021-10090-8

Domingue, B., Trejo, S., Armstrong-Carter, E., & Tucker-Drob, E. (2020, September). Interactions between polygenic scores and environments: Methodological and conceptual challenges. *Sociological Science*, *7*, 365–486. https://doi.org/10.15195/v7.a19

Donnelly, S., & Verkuilen, J. (2017, June). Empirical logit analysis is not logistic regression. *Journal of Memory and Language*, *94*, 28–42. https://doi.org/10.1016/j.jml.2016.10.005

Dudbridge, F., & Fletcher, O. (2014). Gene-environment dependence creates spurious gene-environment interaction. *The American Journal of Human Genetics*, *95*(3), 301–307. https://doi.org/10.1016/j.ajhg.2014.07.014

Duncan, L. E., & Keller, M. C. (2011). A critical review of the first 10 years of candidate gene-by-environment interaction research in psychiatry. *American Journal of Psychiatry*, *168*(10), 1041–1049. https://doi.org/10.1176/appi.ajp.2011.11020191

Embretson, S. E. (1996). Item response theory models and spurious interaction effects in factorial ANOVA designs. *Applied Psychological Measurement*, *20*(3), 201–212. https://doi.org/10.1177/014662169602000302

Fletcher, J. M., & Conley, D. (2013). The challenge of causal inference in gene–environment interaction research: Leveraging research designs from the social sciences. *American Journal of Public Health*, *103*(S1), S42–S45. https://doi.org/10.2105/AJPH.2013.301290

Gardner, W., Mulvey, E. P., & Shaw, E. C. (1995). Regression analyses of counts and rates: Poisson, overdispersed Poisson, and negative binomial models. *Psychological Bulletin*, *118*(3), 392–404. https://doi.org/10.1037/0033-2909.118.3.392

Garin, O. (2014). Ceiling effect. In A. C. Michalos (Ed.), *Encyclopedia of quality of life and well-being research* (pp. 631–633). Springer. https://doi.org/10.1007/978-94-007-0753-5_296

Gelman, A., Hill, J., & Vehtari, A. (2020). *Regression and other stories*. Cambridge University Press.

Gomila, R. (2020). Logistic or linear? Estimating causal effects of experimental treatments on binary outcomes using regression analysis. *Journal of Experimental Psychology: General*, *150*(4), 700–709. https://doi.org/10.1037/xge0000920

Grotzinger, A. D., Briley, D. A., Engelhardt, L. E., Mann, F. D., Patterson, M. W., Tackett, J. L., Tucker-Drob, E. M., & Harden, K. P. (2018, April). Genetic and environmental influences on pubertal hormones in human hair across development. *Psychoneuroendocrinology*, *90*, 76–84. https://doi.org/10.1016/j.psyneuen.2018.02.005

Haaf, J. M., & Rouder, J. N. (2019). Some do and some don't? Accounting for variability of individual difference structures. *Psychonomic Bulletin & Review*, *26*(3), 772–789. https://doi.org/10.3758/s13423-018-1522-x

Halpern, S. D., French, B., Small, D. S., Saulsgiver, K., Harhay, M. O., Audrain-McGovern, J., Loewenstein, G., Asch, D. A., & Volpp, K. G. (2016). Heterogeneity in the effects of reward-and deposit-based financial incentives on smoking cessation. *American Journal of Respiratory and Critical Care Medicine*, *194*(8), 981–988. https://doi.org/10.1164/rccm.201601-0108OC

Henningsen, A. (2020). censreg: Censored regression (Tobit) models [Computer software manual]. https://CRAN.R-project.org/package=censReg (R package version 0.5-32)

Jaccard, J., & Brinberg, M. (2021). Monte Carlo simulations using extant data to mimic populations: Applications to the modified linear probability model and logistic regression. *Psychological Methods*, *26*(4), 450–465. https://doi.org/10.1037/met0000383

Kang, S.-M., & Waller, N. G. (2005). Moderated multiple regression, spurious interaction effects, and IRT. *Applied Psychological Measurement*, *29*(2), 87–105. https://doi.org/10.1177/0146621604272737

Keller, M. C. (2014). Gene× environment interaction studies have not properly controlled for potential confounders: The problem and the (simple) solution. *Biological Psychiatry*, *75*(1), 18–24. https://doi.org/10.1016/j.biopsych.2013.09.006

King, G., & Roberts, M. E. (2015). How robust standard errors expose methodological problems they do not fix, and what to do about it. *Political Analysis*, *23*(2), 159–179. https://doi.org/10.1093/pan/mpu015

Knief, U., & Forstmeier, W. (2021). Violating the normality assumption may be the lesser of two evils. *Behavior Research Methods*, *53*(6), 2576–2590. https://doi.org/10.3758/s13428-021-01587-5

Koedel, C., & Betts, J. (2010). Value added to what? How a ceiling in the testing instrument influences value-added estimation. *Education Finance and Policy*, *5*(1), 54–81. https://doi.org/10.1162/edfp.2009.5.1.5104

Liu, M., Jiang, Y., Wedow, R., Li, Y., Brazel, D. M., Chen, F., Datta, G., Davila-Velderrain, J., McGuire, D., Tian, C., Zhan, X., Choquet, H., Docherty, A. R., Faul, J. D., Foerster, J. R., Fritsche, L. G., Gabrielsen, M. E., Gordon, S. D., Haessler, J., . . . Vrieze, S. (2019). Association studies of up to 1.2 million individuals yield new insights into the genetic etiology of tobacco and alcohol use. *Nature Genetics*, *51*(2), 237–244. https://doi.org/10.1038/s41588-018-0307-5

Loftus, G. R. (1978). On interpretation of interactions. *Memory & Cognition*, *6*(3), 312–319. https://doi.org/10.3758/BF03197461

Lubinski, D., & Humphreys, L. G. (1990). Assessing spurious "moderator effects": Illustrated substantively with the hypothesized ("synergistic") relation between spatial and mathematical ability. *Psychological Bulletin*, *107*(3), 385–393. https://doi.org/10.1037/0033-2909.107.3.385

MacCallum, R. C., & Mar, C. M. (1995). Distinguishing between moderator and quadratic effects in multiple regression. *Psychological Bulletin*, *118*(3), 405–421. https://doi.org/10.1037/0033-2909.118.3.405

Meyers, J. L., Salvatore, J. E., Aliev, F., Johnson, E. C., McCutcheon, V. V., Su, J., Kuo, S. I.-C., Lai, D., Wetherill, L., Wang, J. C., Chan, G., Hesselbrock, V., Foroud, T., Bucholz, K. K., Edenberg, H. J., Dick, D. M., Porjesz, B., & Agrawal, A. (2019). Psychosocial moderation of polygenic risk for cannabis involvement: The role of trauma exposure and frequency of religious service attendance. *Translational Psychiatry*, *9*(1), 1–12. https://doi.org/10.1038/s41398-019-0598-z

Michell, J. (1986). Measurement scales and statistics: A clash of paradigms. *Psychological Bulletin*, *100*(3), 398–407. https://doi.org/10.1037/0033-2909.100.3.398

Michell, J. (2008). Is psychometrics pathological science? *Measurement*, *6*(1–2), 7–24. https://doi.org/10.1080/15366360802035489

Minkovitz, C. S., Strobino, D., Scharfstein, D., Hou, W., Miller, T., Mistry, K. B., & Swartz, K. (2005). Maternal depressive symptoms and children's receipt of health care in the first 3 years of life. *Pediatrics*, *115*(2), 306–314. https://doi.org/10.1542/peds.2004-0341

Neal, D. J., & Simons, J. S. (2007). Inference in regression models of heavily skewed alcohol use data: A comparison of ordinary least squares, generalized linear models, and bootstrap resampling. *Psychology of Addictive Behaviors*, *21*(4), 441–452. https://doi.org/10.1037/0893-164X.21.4.441

Nelder, J. A., & Wedderburn, R. W. (1972). Generalized linear models. *Journal of the Royal Statistical Society: Series A (General)*, *135*(3), 370–384. https://doi.org/10.2307/2344614

Page, A. C., Hooke, G. R., & Morrison, D. L. (2007). Psychometric properties of the depression anxiety stress scales (DASS) in depressed clinical samples. *British Journal of Clinical Psychology*, *46*(Pt 3), 283–297. https://doi.org/10.1348/014466506X158996

Richardson, C. G., & Ratner, P. A. (2005). Sense of coherence as a moderator of the effects of stressful life events on health. *Journal of Epidemiology & Community Health*, *59*(11), 979–984. https://doi.org/10.1136/jech.2005.036756

Rohrer, J. M., & Arslan, R. C. (2021). Precise answers to vague questions: Issues with interactions. *Advances in Methods and Practices in Psychological Science*, *4*(2). https://doi.org/10.1177/25152459211007368

Seaton, E. K., & Iida, M. (2019). Racial discrimination and racial identity: Daily moderation among black youth. *American Psychologist*, *74*(1), 117–127. https://doi.org/10.1037/amp0000367

Stringa, N., Milaneschi, Y., van Schoor, N. M., Suanet, B., van der Lee, S., Holstege, H., Reinders, M. J. T., Beekman, A. T. F., & Huisman, M. (2020). Genetic liability for depression, social factors and their interaction effect in depressive symptoms and depression over time in older adults. *The American Journal of Geriatric Psychiatry*, *28*(8), 844–855. https://doi.org/10.1016/j.jagp.2020.02.011

Tabery, J. (2008). RA Fisher, Lancelot Hogben, and the origin (s) of genotype–environment interaction. *Journal of the History of Biology*, *41*(4), 717–761. https://doi.org/10.1007/s10739-008-9155-y

Tobin, J. (1958). Estimation of relationships for limited dependent variables. *Econometrica: Journal of the Econometric Society*, *26*(1), 24–36. https://doi.org/10.2307/1907382

Trejo, S., Belsky, D. W., Boardman, J. D., Freese, J., Harris, K. M., Herd, P., Sicinski, K., & Domingue, B. W. (2018, August). Schools as moderators of genetic associations with life course attainments: Evidence from the WLS and Add Heath. *Sociological Science*, *5*, 513–540. https://doi.org/10.15195/v5.a22

Tucker-Drob, E. M. (2009). Differentiation of cognitive abilities across the life span. *Developmental Psychology*, *45*(4), 1097–1118. https://doi.org/10.1037/a0015864

Tucker-Drob, E. M. (2019, December). Cognitive aging and dementia: A life-span perspective. *Annual Review of Developmental Psychology*, *1*, 177–196. https://doi.org/10.1146/annurev-devpsych-121318-085204

van der Linden, W. J. (2016). *Handbook of item response theory: Volume 1: Models*. CRC Press.

Winter, B., & Bürkner, P.-C. (2021). Poisson regression for linguists: A tutorial introduction to modelling count data with brms. *Language and Linguistics Compass*, *15*(11), Article e12439. https://doi.org/10.1111/lnc3.12439

Yeager, D. S., Hanselman, P., Walton, G. M., Murray, J. S., Crosnoe, R., Muller, C., Tipton, E., Schneider, B., Hulleman, C. S., Hinojosa, C. P., Paunesku, D., Romero, C., Flint, K., Roberts, A., Trott, J., Iachan, R., Buontempo, J., Yang, S. M., Carvalho, C. M., . . . Dweck, C. S. (2019). A national experiment reveals where a growth mindset improves achievement. *Nature*, *573*(7774), 364–369. https://doi.org/10.1038/s41586-019-1466-y

(*Appendices follow*)

Appendix A

Taylor Series Expansions

We use Taylor series expansions to approximate a function $f(x)$ as an infinite sum of polynomials $\sum_{n=0}^{\infty} a_n x^n$. Below we derive values for the constants a_n for different choices of f that are relevant here. In the main text, we consider the expansion where we replace the argument x with Γ. In particular, we identify the implied terms in the infinite series that involve a linear product of the predictors (i.e., the xz term where x now refers to the predictor from Equation 1, not the argument in the above discussion of f).

Expansion for Binary Outcomes

Suppose that f is the logistic function such that $f(x) = (1 + e^{-x})$. We then have:

$$f'(x) = (1 + e^{-x})^{-2}(e^{-x}) \quad (8)$$

$$f''(x) = 2(1 + e^{-x})^{-3}(e^{-2x}) - (1 + e^{-x})^{-2}(e^{-x}) \quad (9)$$

$$f'''(x) = 6(1 + e^{-x})^{-4}(e^{-3x}) - 4(1 + e^{-x})^{-3}(e^{-2x}) - f''(x). \quad (10)$$

Using these, we first compute derivatives at $x = 0$.

$$f'(0) = \frac{1}{4} \quad (11)$$

$$f''(0) = 0 \quad (12)$$

$$f'''(0) = -\frac{1}{8} \quad (13)$$

Using these (and the fact that $f(0) = \frac{1}{2}$), we can construct the Taylor series expansion around $x = 0$:

$$f(x) = f(0) + \frac{f'(0)}{1!}x + \frac{f''(0)}{2!}x^2 + \frac{f'''(0)}{3!}x^3 + \cdots \quad (14)$$

$$f(x) = \frac{1}{2} + \frac{1}{4}x + 0x^2 - \frac{1}{48}x^3 + \cdots. \quad (15)$$

We use this expansion in Equation 2.

Expansion for Count Outcomes

Here, we use the well-known expansion of $f(x) = e^x$ around $x = 0$:

$$e^x = \sum_{n=0}^{\infty} \frac{x^n}{n!}. \quad (16)$$

This expansion is used in Equation 3.

Expansion for Noninterval Outcomes

Suppose that $f(x) = a^{bx}$. We again take derivatives:

$$f'(x) = (b\ln a)a^{bx} \quad (17)$$

$$f''(x) = (b\ln a)^2 a^{bx} \quad (18)$$

and then compute values

$$f(0) = 1 \quad (19)$$

$$f'(0) = b\ln a \quad (20)$$

$$f''(0) = (b\ln a)^2. \quad (21)$$

Using these facts, we can write

$$f(x) = 1 + (b\ln a)x + \frac{1}{2}(b\ln a)^2 x^2 + \cdots. \quad (22)$$

which we use in Equation 7.

(*Appendices continue*)

Appendix B

Distinguishing Interactions From Main Effects Via the Taylor Series

We use the generic notion of Taylor series expansions to illustrate some key elements related to the unique problem of interactions when a linear model is assumed. We will assume $\mathbb{E}(y \mid x, z) = f(\Gamma)$ and that we approximate f via a Taylor series, $f(\Gamma) = a_0 + a_1\Gamma + a_2\Gamma^2 + \cdots$. Suppose first that $\Gamma = \beta_0 + \beta_1 x$; i.e., $\beta_2 = \beta_3 = 0$ and z has no effect on y. We would then have:

$$\mathbb{E}(y \mid x, z) = a_0 + a_1(\beta_0 + \beta_1 x + \beta_2 z + \beta_3 xz) + a_2(\beta_0 + \beta_1 x + \beta_2 z + \beta_3 xz)^2 + \cdots \tag{23}$$

$$= a_0 + a_1(\beta_0 + \beta_1 x) + a_2(\beta_0 + \beta_1 x)^2 + \cdots \tag{24}$$

$$= a_0 + a_1\beta_0 + a_1\beta_1 x + a_2\beta_0^2 + 2a_2\beta_0 x + a_2\beta_1^2 x^2 + \cdots. \tag{25}$$

Note first that z appears nowhere; given that z always appears in a $\beta_2 z$ product in the Taylor series expansion, if $\beta_2 = 0$ we will not detect spurious main effects of z.

Suppose now that $\beta_2 \neq 0$ but $\beta_3 = 0$. In this case, the higher-order Γ terms will inevitably produce product terms. For example,

$$\Gamma^2 = \beta_0^2 + 2\beta_0\beta_1 x + 2\beta_0\beta_2 z + \beta_1^2 x^2 + 2\beta_1\beta_2 xz + \beta_2^2 z^2. \tag{26}$$

We emphasize the $2\beta_1\beta_2 xz$ term. As long as $\beta_1\beta_2 \neq 0$, if a misspecified linear regression model $\mathbb{E}(y \mid x, z) = b_0 + b_1 x + b_2 z + b_3 xz$ is then fit to resulting data, then these interaction terms from the expansion will lead to nonzero estimates of b_3 even though $\beta_3 = 0$. That is, well-powered studies will necessarily result in nonzero b_3 estimates (the problem of spurious interactions we focus on here). Matters are somewhat more complicated in the case of binary outcomes generated via the logistic regression model; in that case, $a_2 = 0$ but higher-order terms will still include xz terms that may induce spurious interactions (under certain assumptions about β_0, see discussion in main text).

Reverting back to the assumption that $\beta_2 = \beta_3 = 0$, we make some final remarks related to x given that $\mathbb{E}(y \mid x, z) = (a_0 + a_1\beta_0) + (a_1\beta_1 + 2a_2\beta_0)x + \cdots$. Suppose a misspecified regression model of $\mathbb{E}(y \mid x, z) = b_0 + b_1 x$ is deployed. Estimates of b_1 will retain sensitivity to, for example, β_0 given the $2a_2\beta_0$ term in the expansion. In general, the degree to which the misspecified regression produces reasonable results seems dependent on the degree to which a linear approximation is appropriate (i.e., is $\mathbb{E}(y \mid x, z) - (a_0 + a_1\beta_0) - (a_1\beta_1 + 2a_2\beta_0)x$ small over the range of x?). This assumption regarding linearity (or, perhaps even more importantly, monotonicity) is an important consideration that is beyond the scope of our article. For our purposes, we think the potential shortcomings of estimates of b_1, in this case where a linear model is misspecified, are distinctive from the interaction case wherein the existence of main effects will induce spurious interaction coefficients (i.e., the analysis of Γ^2 above).

Received November 12, 2021
Revision received July 7, 2022
Accepted August 10, 2022 ■