esa ECOSPHERE

SYNTHESIS & INTEGRATION

Quantifying relative importance: computing standardized effects in models with binary outcomes

James B. Grace,[1,†] Darren J. Johnson,[2] Jonathan S. Lefcheck,[3] and Jarrett E. K. Byrnes[4]

[1] *Wetland and Aquatic Research Center, U.S. Geological Survey, Lafayette, Louisiana 70506 USA*
[2] *Cherokee Nations Technical Solutions, Wetland and Aquatic Research Center, Lafayette, Louisiana 70506 USA*
[3] *Bigelow Laboratory for Ocean Science, East Boothbay, Maine 04544 USA*
[4] *Department of Biology, University of Massachusetts, Boston, Massachusetts 02125 USA*

Citation: Grace, J. B., D. J. Johnson, J. S. Lefcheck, and J. E. K. Byrnes. 2018. Quantifying relative importance: computing standardized effects in models with binary outcomes. Ecosphere 9(6):e02283. 10.1002/ecs2.2283

Abstract. Scientists commonly ask questions about the relative importance of processes and then turn to statistical models for answers. Standardized coefficients are typically used in such situations, with the goal being to compare effects on a common scale. Traditional approaches to obtaining standardized coefficients were developed with idealized Gaussian variables in mind. When responses are binary, complications arise that impact standardization methods. In this paper, we review, evaluate, and propose new methods for standardizing coefficients from models that contain binary outcomes. We first consider the interpretability of unstandardized coefficients and then examine two main approaches to standardization. One approach, which we refer to as the latent-theoretical (LT) method, assumes that underlying binary observations, there exists a latent, continuous propensity linearly related to the coefficients. A second approach, which we refer to as the observed-empirical (OE) method, assumes responses are purely discrete and estimates error variance empirically via reference to a classical R^2 estimator. We also evaluate the standard formula for calculating standardized coefficients based on standard deviations. Criticisms of this practice have been persistent, leading us to propose an alternative formula that is based on user-defined "relevant ranges." Finally, we implement all of the above in an open-source package for the statistical software R. Results from simulation studies show that both the LT and OE methods of standardization support a similarly broad range of coefficient comparisons. The LT method estimates effects that reflect underlying latent-linear propensities, while the OE method computes a linear approximation for the effects of predictors on binary responses. The contrast between assumptions for the two methods is reflected in persistently weaker standardized effects associated with OE standardization. Reliance on standard deviations for standardization (the traditional approach) is critically examined and shown to introduce substantial biases when predictors are non-Gaussian. The use of relevant ranges in place of standard deviations has the capacity to place LT and OE standardized coefficients on a more comparable scale. As ecologists address increasingly complex hypotheses, especially those that comparing the influences of different controlling factors (e.g., top-down vs. bottom-up or biotic vs. abiotic controls), comparable coefficients become necessary for evaluations.

Key words: binary response models; coefficients; logistic regression; path models; relevant-range standardization; review; standardization; statistics; structural equation modeling.

Received 7 March 2018; revised 21 April 2018; accepted 2 May 2018. Corresponding Editor: Debra P. C. Peters.

† **E-mail:** gracej@usgs.gov

Introduction

For linear and generalized linear models, coefficients are the currency of the realm in scientific interpretations. Statistical modeling involves specification, estimation, and model selection, all designed to obtain coefficient estimates that can serve as trustworthy measures of the effects of predictors on responses. The coefficient estimates obtained represent not only valuable information about the current relationships among system properties, but how systems got to be in their current state, how they may behave elsewhere, and future possibilities. The importance of coefficient interpretation thus deserves thorough scrutiny because of its central role in the scientific process.

The development of standardized coefficients as an aid to interpretation has a long history in applied statistics (e.g., Gelman and Hill 2007, Fox 2016). Generally speaking, standardization is used by scientists in an attempt to put coefficients on a common scale for the purpose of comparing the relative strengths of processes in models with multiple predictors. It seems that interest in comparing the relative strengths of processes in complex ecological models has increased in recent years. Topics of interest to ecologists commonly focus on broad comparisons, such as top-down vs. bottom-up controls of food webs (e.g., Ecology special feature: Matson and Hunter 1992) and the relative importance of diversity components for ecosystem functioning (Flynn et al. 2011).

While ecological models and scientific aspirations have become increasingly complex, in part due to advances in statistical modeling (Hilborn and Mangel 1997, Shipley 2000, Royle and Dorazio 2008, Grace et al. 2012), advice for computing standardized effects is generally handed down from past traditions. The traditional conceptualization of a standardized coefficient was developed assuming Gaussian variables (Fox 2016:101). The ostensible meaning of a standardized effect is the proportion of standard deviation change predicted for a response variable if the mean value of a predictor was increased by one standard deviation. When standardized coefficients are developed using ideal Gaussian variables they exhibit a number of useful properties. In this paper, we instead focus on the situation where our models fail to meet the idealized assumptions. In particular, we consider ways to overcome the challenges posed by binary response variables. In addition, we also consider persistent criticisms of the traditional formula for coefficient standardization and possible remedies. Generally, we argue that the development of clearly interpretable standardized coefficients is important for certain scientific interpretations and this is true regardless of the distributions of the variables in our models.

In our review of the literature, we have found the greatest discussion of this topic in the fields of sociometrics and econometrics. One goal in this paper is to illustrate these ideas for an ecological audience. We have further observed that the advice given in textbooks for how to interpret coefficients from binary response models is often only appropriate within the narrow context of very simple models. Scientists in many disciplines, including ecology, are increasingly interested in complex integrative models that contain numerous variables and networks of direct and indirect effects. In this paper, we seek methods appropriate for this situation.

The Challenges Posed by Binary Responses

Overview of binary response models

Many ecologically interesting phenomena are expressed as binary (0/1) observations. Examples include species presence, infection status, sexual maturity, reproductive status, and survival. A collection of methods, generally referred to as binary response models (BRMs), are commonly used when response sets are {present/absent}, {yes/no}, or {1,0} (Maddala 1983, McCullagh and Nelder 1989, Long 1997, Menard 2010, Agresti 2013, Fox 2016). Texts presenting statistical methods specifically for ecologists and environmental scientists most commonly illustrate BRMs from the perspective of logistic regression (e.g., Floyd 2001, Quinn and Keough 2002, Gotelli and Ellison 2004, Zuur et al. 2007, Bolker 2008, Legendre and Legendre 2012, Buckley 2015). It has come to be nearly universal that ecologists approach binary response modeling from within a generalized linear modeling framework using the logit link, which can be implemented using either maximum likelihood or Bayesian techniques.

A generalized linear model (or GLM; Nelder and Wedderburn 1972) consists of three parts: a random component, a linear predictor, and a link function. The GLM allows us to model certain inherently nonlinear relationships between predictors and responses using a linear model in combination with a link function that transforms the expected values to be linearly related to the predictor. This model type can be applied using a number of link functions corresponding to different response distributions, giving the GLM great flexibility. The random component of the GLM specifies the response variable (y) and its probability distribution. The individual observations, y_i, constitute a vector of values $\mathbf{y}$ that are assumed to be independent and follow a probability density distribution from within the exponential family (or extensions thereof).

The linear predictor component of the GLM consists of a vector of predicted scores $\boldsymbol{\eta}$, with one score for each y_i observation. Scores are generated from a matrix of explanatory variable observations, $\mathbf{x} = x_{ij}$ (where i indexes observations and j indexes the p predictor variables), and a vector of estimated coefficients $\boldsymbol{\beta} = \beta_j$ as represented in Eq. 1.

$$\boldsymbol{\eta} = \mathbf{x}\boldsymbol{\beta} \tag{1}$$

Link functions $\mathbf{g}(\cdot)$ transform the expected values for the response variable $E(\mathbf{y})$ so as to have a linear relation to the predictor. These link functions are generally required to have the property of being invertible. This leads to the following set of relations, where $\boldsymbol{\mu}$ is the expected value of y, that is, $P(y_i = 1)$.

$$E(\mathbf{y}) \equiv \boldsymbol{\mu} \tag{2}$$

$$g(\boldsymbol{\mu}) = \boldsymbol{\eta} \tag{3}$$

$$\boldsymbol{\mu} = g^{-1}(\boldsymbol{\eta}) \tag{4}$$

We can begin to discern some of the interpretive choices that will have to be made by recognizing the chain of computational relationships connecting predictors to responses shown in the symbolic Eq. 5.

$$\mathbf{x} \rightarrow \boldsymbol{\eta} \rightarrow \boldsymbol{\mu} \rightarrow \mathbf{y} \tag{5}$$

While there are a number of link functions that can be applied to binary data, there are two of particular interest in this discussion because of their wide-spread usage, the logit and (to a lesser degree currently) the probit. For the probabilities of the individual observations, $P(y_i = 1) = \mu_i$, vs. $P(y_i = 0) = 1 - \mu_i$, the logit model can be expressed in the following form:

$$\begin{aligned}\operatorname{logit}(\mu_i) &= \log\left(\frac{\mu_i}{1-\mu_i}\right) = \log\left(\frac{P(y_i=1)}{P(y_i=0)}\right)\\ &= \sum_{j=1}^{p} \beta_j x_{ij}\end{aligned} \tag{6}$$

The probit model takes the inverse of the standard normal cumulative density function Φ as the link function (Eq. 7). Throughout the rest of the paper, our primary emphasis will be on the logit model because it is so commonly used in ecology, though the treatment extends naturally to the probit model.

$$\operatorname{probit}(\mu_i) = \Phi^{-1}(\mu_i) = \sum_{j=1}^{p} \beta_j x_{ij} \tag{7}$$

Interpretations of raw coefficients from the logit model

The chances that an event will occur are perhaps most naturally thought of in terms of probabilities, which range between a minimum of 0 and a maximum of 1. The coefficient estimates returned from logistic regression do not, however, directly represent predicted effects of changes in x on $P(y = 1)$. Rather, the relationship between predictors (e.g., months of drought) and the probability of observing an event (e.g., tree mortality) is nonlinear, as illustrated by the solid line in Fig. 1. What methods like GLM using a logit link do is to find coefficient values (represented by the dashed line in Fig. 1) that relate the predictors to the logits or log odds ratios.

Returning to the standard advice given to ecologists in biostatistics textbooks, we find that the interpretation of coefficients most commonly focuses on the log odds ratios. For example, Zuur et al. (2007) illustrate logistic regression using data from a study of fish distributions along salinity gradients in estuaries. They interpreted a parameter estimate of −0.12 obtained from a GLM with logit link as the expected change in the log odds ratio for observed outcomes in response to a unit change in the predictor. Going further, they point out that the log odds value of −0.12

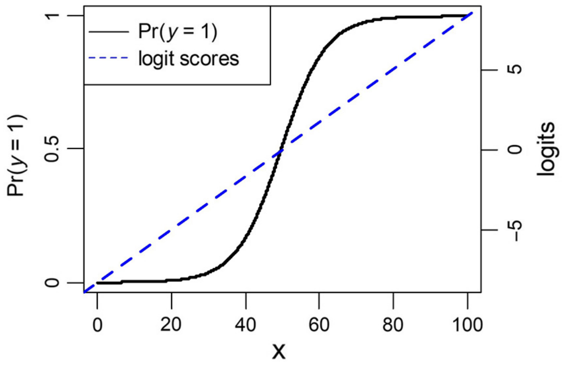

Fig. 1. Illustration of the theoretical relationship between predictors (x), probabilities of observing a binary response (solid line and left axis), and logit scores (dashed line and right axis).

means a 1-unit change in salinity translates into a $e^{-0.12} = 0.88$ reduction in the odds of observing a fish. Rather than emphasizing log odds or odds ratios, some authors discuss interpretations in terms of probabilities (e.g., Gotelli and Ellison 2004, Bolker 2008, Legendre and Legendre 2012). The merits of alternative approaches to interpretation of raw coefficients from logistic regression are discussed at some length by Allison (2012).

There are additional problems that arise for interpreting log odds ratios for certain types of comparisons. One issue has to do with the property of collapsibility. Greenland et al. (1999) state that a measure of association between two variables is “strictly collapsible” with regard to a third variable if the measure of association is constant regardless of the value of the third variable. This property holds for linear Gaussian response models without interactions (see demonstrations in Appendix S1). Greenland et al. (1999) show for binary responses that differences between probabilities (risk differences) are collapsible, but risk ratios (ratios of probabilities) and odds ratios are non-collapsible. This means log odds ratios are not comparable across models (reason being that error variance is fixed, regardless of predictors, and thus implied total variance also varies across models). This illustrates a common situation where raw coefficients from BRMs may not provide a satisfactory capacity for interpreting and comparing coefficients, for example, when comparing models with different sets of predictors or when comparing coefficients within path models containing more than one binary response. Further, non-collapsibility for the raw coefficients also means they cannot be compared across samples; this is true despite the fact that the usual advice given is to use the raw coefficients for comparisons across samples. Problems such as these add motivation to the search for alternative approaches to interpretation (e.g., Allison 1999, Mood 2010).

Proposed Remedies

Linear approximations

Despite there being a nonlinear relationship between predictors and probabilities, scientists in some disciplines, particularly econometrics, nonetheless prefer to use linear models for binary responses (Hellevik 2009). The arguments made in support of this practice are that impacts on tests of significance are minor and that raw coefficients can be directly interpreted as differences in probabilities. Others have argued directly against this practice because there are many situations where the approximation will fail (Long 1997, Allison 2012). The apparently wide-spread use of a linear model in certain fields (https://statisticalhorizons.com/when-can-you-fit) is a testament to the problems associated with reliance on log odds ratios for interpretation as described in the previous section.

Latent-linear conceptualization

An alternative approach to adopting a linear interpretation of the effects of predictors on binary responses is known as the latent propensity model (McKelvey and Zavoina 1975). To see how this approach works, let us imagine that underlying our binary observations of fish (yes or no) at sample points along continuous gradations in salinity, there is a continuously varying propensity or suitability for observing a fish or not. If these fish prefer moderate levels of salinity, we can imagine that as salinity increases, conditions become increasingly unfavorable and this continues well past the point where the probability of observing a live fish approaches zero. Thus, in this example, it is fairly natural to think about the landscape of habitat suitability for the fish in terms of a latent (unmeasured) propensity.

If we assume that a continuous but latent y variable (designated y^*) is linearly related to a vector of observed xs via a vector of βs, we can represent the relationship through the following equation:

$$y_i^* = \mathbf{x_i}\boldsymbol{\beta} + \varepsilon_i \quad (8)$$

A key to understanding the latent propensity approach is the assumption made about the data-generating process. Individual values of y_i^* are presumed to be related to observed binary values through a pair of inequalities that compare y^* to a threshold cutpoint τ, which in turn determines whether we observe a 1 or 0 (Eq. 9).

$$\begin{aligned} y_i &= 1 \text{ if } y_i^* > \tau \\ y_i &= 0 \text{ if } y_i^* \leq \tau \end{aligned} \quad (9)$$

These relations are visualized in Fig. 2.

Adopting a continuous latent propensity approach hypothetically permits many of the interpretations usually reserved for models with Gaussian responses and linear relations between predictors and responses (Greene 2012). At the same time, it is possible to connect the latent-linear conceptualization to logit or probit models (Long 1997) by recognizing that there is a distribution of probabilities for observing $y_i = 1$ or $y_i = 0$ that varies depending upon where the $E(y^* \mid x)$ line in Fig. 2 falls relative to the cutpoint (τ). The biggest difference between the probit and logit models is that the variance for the probit distribution equals 1, while for the logistic distribution, the variance equals $\pi^2/3$. Both forms have a long history of use with binary data.

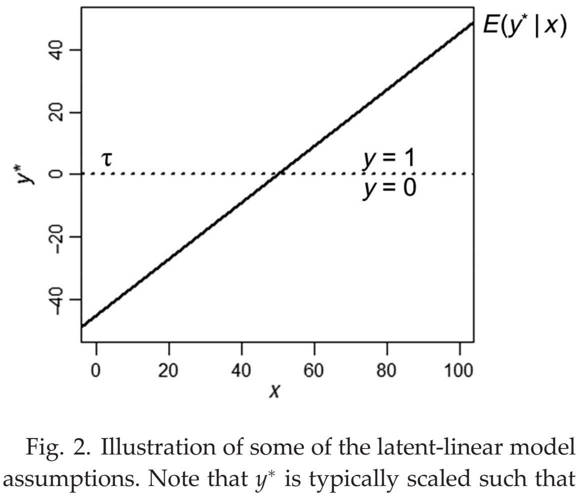

Fig. 2. Illustration of some of the latent-linear model assumptions. Note that y^* is typically scaled such that τ = 0.

While conceptually convenient, a problem with introducing latent variables into models is that key properties are unknown, such as their scale (mean) and variance. This requires that we make arbitrary assumptions about scale and variance to permit maximum-likelihood (ML) estimation via partial identification. We demonstrate explicitly the implications of this later in the paper.

Standardization of coefficients from binary response models

There has been considerable discussion of options for standardizing the coefficients from BRMs (summarized in Menard 2010, 2011). Reference to Eq. 5 suggests some of the possibilities. The coefficients, **β**s, which we wish to standardize, are chosen through ML estimation to compose a predictor **η** whose transformation into probabilities, $E(\mathbf{y})$, best matches the observed zeros and ones. In this somewhat complex process, the dependent variable that relates directly to the predictors is not y, but rather logit(y). Problematic for standardization is that the individual logit(y_i) values transform to $-\infty$ for the case logit (0/1) and $+\infty$ for logit(1/0) (refer to Eq. 6). The logit(y) itself represents a transformation of the ratio of the probabilities of observing ones or zeros and thus is not analogous to a traditional dependent variable made up of a systematic component plus a random (unexplained) component. Instead, the error of prediction in a GLM is represented by the residual deviance for the likelihood function rather than the unexplained portion of the total variance in logit(y). The use of deviance criteria means that it is not possible to directly compute standard deviations for the dependent variable in the usual fashion because we lack a direct estimate of error variance.

Several approaches to solving the just-described problem have been proposed. These can be divided into two types. The first we can refer to as a latent-theoretical approach while the second we refer to as a observed-empirical approach.

Latent-theoretical approach to standardization.— As early as 1975, McKelvey and Zavoina (1975) proposed the latent propensity model described previously (Eqs. 8 and 9). To solve the problem that the error was unidentified, they set a fixed value related to the assumed model form:

Var(ε) = 1 for probit regression, which was later extended by Winship and Mare (1984) to include the logit case, where it is traditional to set Var(ε) = $\pi^2/3$. These values come from the theoretically defined properties of the associated distributions (see Long 1997). McKelvey and Zavoina then proposed estimating the variance of y^* as the sum of the predictor variance plus the assumed error variance.

$$\sigma^2_{y*} = \sigma^2_{x\beta} + \sigma^2_{\epsilon} \qquad (10)$$

This, incidentally, provides a means of computing a latent-linear model R^2 using an equation of the form typically used for Gaussian response models (Eq. 11).

$$R^2 = \frac{\sigma^2_{\eta}}{\sigma^2_{y^*}} \qquad (11)$$

It also opens the door to standardizing coefficients using the ratio of the standard deviations (σ) of the predictor (x) and y^*, as given in Eq. 12 for $j = p$ predictors (coefficient superscript "s" indicates coefficient is standardized).

$$b^s_{yx_j} = b_{yx_j} \times \frac{\sigma_x}{\sigma_{y^*}} \qquad (12)$$

Observed-empirical approach to standardization.— Menard (1995, 2010) has proposed an alternative method that we refer to here as an "observed-empirical" approach to standardization. A first major difference from the "latent-theoretical" approach is that Menard's method does not assume binary variables are underlain by a continuous latent propensity and does not assume data generation involves a cutpoint, but rather, an error distribution. The second major difference is that Menard's method does not use a theoretical approach to estimate the variance of error, but instead relies on an estimate derived from a calculation of R^2 for the model. While there are several different formulae for R^2s and pseudo-R^2s of BRMs that have been proposed (Menard 2010: Chapter 3), for this purpose, he relies on a traditional ordinary least squares (OLS) definition that is the ratio of the predicted ($\hat{y}$) vs. observed values of y (Eq. 13). This choice is motivated by the fact that only the OLS R^2 involves actual variances and thus permits direct computation of a standard deviation (Menard, *personal communication*).

$$R^2 = \frac{\sigma^2_{\hat{y}}}{\sigma^2_y} \qquad (13)$$

Rearranging and taking the square root of both sides allows us to develop an estimator for the standard deviation of the response variable (i.e., σ_y) for use in standardization (Eq. 14).

$$\sigma^2_y = \frac{\sigma^2_{\hat{y}}}{R^2} \text{ and } \sigma_y = \frac{\sigma_{\hat{y}}}{R} \qquad (14)$$

Recognizing that b_{yx} is still in logits, we nonetheless have the materials for a standardization procedure (Eq. 15).

$$b^{s_M}_{yx} = b_{yx} \times \sigma_x/\sigma_y \qquad (15)$$

Using this approach, we generate a standardized coefficient that is roughly analogous to standardized coefficients from Gaussian models but do so without resorting to a latent propensity framework. The performance of this method is influenced by the fact that it is based on a linear approximation to a nonlinear relationship between $\hat{y}$ and y.

While it is beyond our purpose to cover in detail in this paper, we note that it is possible to extend the LT and OE standardization methods to models where categorical responses include more than two outcomes. The two most common model types for this situation are (1) ordered categorical models and (2) multinomial models (note that we follow Fox 2016 in using the strict definition of multinomial, i.e., comprising multiple, nominal variables). One difference between these two model types is the first makes a strict assumption of a constant effect of predictors on the conditional log odds of transitioning from one category to the next. Thus, models for ordered categorical responses, such as the proportional-odds model (Fox 2016: 400–407), return only a single slope coefficient, which can be standardized using any of the methods we present. In contrast, models for multiple nominal responses return multiple slope coefficients. In such models, one outcome is treated as the baseline and the log odds for each alternative outcome are benchmarked to the baseline. Multiple slope coefficients are returned for models of this case, so separate standardizations are required. Nonetheless, the methods presented in this paper are appropriate for this case as well.

Addressing the criticisms of standardization based on standard deviations—relevant ranges.—There has been persistent criticism through the years of the use of standard deviations as the basis for standardizing coefficients (Tukey 1954, Turner and Stevens 1959, Greenland et al. 1991, Pedhazur 1997, Fox 2016). Despite this, the computation of standardized coefficients using standard deviations is routinely conducted without critical evaluation.

The usual definition of a standardized coefficient is "the expected change in y in standard deviation units if we were to increase x by one standard deviation." Traditionally, completely standardized coefficients have been obtained either by standardizing (z-transforming) the input variables or by multiplying the raw coefficients by the ratios of the standard deviations of the predictors to the response variables. In this treatment, we treat the latter (computational) method as the more general case because it can return both raw and standardized coefficients and in order to be consistent with the general references we draw upon in this review (e.g., Long 1997, Menard 2010, Fox 2016).

It is understood that a standard deviation represents a significant portion of the range of a variable, and thus, standardization based on standard deviations is a proxy for standardization by the ranges (Fox 2016). The motivation behind using standard deviations as proxies of ranges comes from the fact that simple ranges are estimated from only two of the available data points, while standard deviations are based on all the observations, and thus less subject to random errors. This surrogacy works fine as long as all variables are consistently Gaussian (both predictors and responses). Sometimes, however, the minimum and maximum for a variable can be interpreted as definitional quantities (e.g., variables bounded between 0.0 and 1.0). When that is the case, ranges may be more meaningful than standard deviations. For this reason, Grace and Bollen (2005) suggested that by replacing standard deviations with ranges, standardized coefficients can be produced that are interpreted as "If we were to vary x across its range, y would change by a certain proportion of its range." This approach expands coefficient comparability to include coefficients involving binary predictors and those that otherwise deviate from a strict Gaussian distribution.

Consideration of the main criticisms made against standardization informs our evaluation. The first problem with the use of standard deviations for standardization often raised (Fox 2016) is the inconsistent relationships between standard deviations and ranges in data. For example, the standard deviation for a binary variable is 50% of the range (for a binary variable with an even proportion of zeros and ones). For Gaussian variables, in contrast, one standard deviation is ~15% of its range when $n = 1000$ and ~20% when $n = 100$. Treating standard deviations as if they are equivalent across variables impairs comparability of coefficients to a degree that is often not small in magnitude.

A second complaint long launched against the use of standard deviations is that estimating the variances of variables (from which the standard deviations are derived) is difficult and very sensitive to sampling methods. As Pedhazur (1997: 319) put it, "The size of a [standardized coefficient] reflects not only the presumed effect of the variable with which it is associated but also the variances. . . of the variables in the model (including the dependent variable), as well as the variances of the variables not in the model and subsumed under the error term. In contrast, [the unstandardized coefficient] remains fairly stable despite differences in the variances and the covariances of the variables in different settings or populations." It may be difficult to eliminate all facets of this complaint. However, range standardization provides an opportunity for addressing at least some of these concerns.

There are situations where standardizing based on ranges seems quite natural. For some variables, such as the percentage of flowers that set seed, we know that the range of possible values spans from 0% to 100%. In such a case, using standard deviations for standardization may seem less appropriate than using the range, which in this case has a clear and transportable meaning. Many variables have such clearly defined minimum and maximum values. An obvious example is a binary indicator used to represent treated vs. untreated individuals. For these and many other cases, ranges may provide a good scale for comparisons.

Estimating the ranges of variables can bring with it a host of problems as well if one is not careful. For variables unbounded in their values, arriving at a suitable estimate of range may be challenging. Further, the observed range is strongly dependent on sample size, with rare values increasingly likely at higher sample sizes. For Gaussian variables, estimates of sample standard deviations stabilize to become independent of sample size (not true for all variable types).

The concept of a relevant range emphasizes that judgment is called for when defining a set of ranges that can establish some basis for coefficient comparability. Rather than automatically choosing the minimum and maximum empirical values to define the range, the investigator may need to use their scientific judgment. The benefit of doing so is that the interpretation of the resulting coefficient is consciously defined. Choosing a relevant range can be approached in several ways:

First, raw coefficients are estimated based on the data in hand. It is important to consider whether the estimated slope seems applicable to some range of predictor values less than or greater than the range in the sample. Making this determination will be greatly aided through the use of appropriate plots of data relationships and a priori knowledge of the system under investigation.

Second, as mentioned previously, many variables have naturally defined minima and maxima. A minimum value of zero applies to many of the quantities measured by scientists. Natural maxima, such as 100%, also occur. An important caveat is that even if a naturally delimited range exists for a variable, the data in hand may not cover enough of that range to automatically justify using the natural range. This is another case where scientific judgment is needed.

Third, there will be times when the range for a relationship is best estimated indirectly, most obviously from the estimated standard deviations. A commonly used rule for relating ranges to standard deviations is that six standard deviations will include 99.973% of the values for a Gaussian distribution. If a simple rule is to be applied in order to avoid relying on the empirical minimum and maximum values, one could use six times the standard deviation as a default. It is possible to refine the conversion of standard deviations to relevant range using two pieces of information, the sample size and the approximate shape of the distribution. With that information, it is possible to simulate data and then compute the average ratio of standard deviations to range, creating a conversion factor specific to the case if desired.

Fourth, there are cases where hypothetical intervention scenarios might serve as the basis for defining relevant ranges. Perhaps restoration dollars are estimated to allow one to either plant 10,000 tree seedlings or cull deer populations by 10%. If it seemed of value to construct coefficients that explicitly represented the relative consequences of these two interventions, it would be possible to do so.

The reward for going through the extra steps to select relevant ranges is the creation of a thoughtfully constructed basis for interpreting coefficients within a meaningful scientific context. Reporting the ranges used when applying range standardization (along with the raw coefficients) should be considered mandatory, as subsequent analyses may choose to use a different set of ranges for standardization. As an example, if an initial study chooses to use a range of 10 for a variable, re-analyses that wish to combine data or to compare across datasets may choose to expand the range to encompass the new samples. This example highlights one of the more important potential applications of range standardization, which is the facilitation of across-sample comparisons. We provide an explicit example of this later in the paper.

Finally, one additional situation where range standardization may provide more interpretable results is when one is using a linear approximation for a nonlinear true relationship. Raw coefficients are rarely useful as summaries for nonlinear relationships since the slope of a nonlinear line is not constant with regard to x. Yet, linear approximations are frequently used for what are truly nonlinear functions. A range-standardized coefficient represents the predicted net change in y over the full range of variation in x. Thus, it seems a more appropriate approximation of the relationship than the expected change over a standard deviation of change, which would vary across x. This attribute of range standardization can be specifically useful in the context of this paper since binary response models are inherently nonlinear and some standardization approaches do represent approximations of net change.

We feel there is much to recommend the use of relevant-range standardization. That determination, however, depends on how far off are the usual assumptions associated with the use of standard deviations. We thus defer further discussion of the merits of various approaches until the *Discussion*, after we have considered the performance of standardized coefficients in a variety of ecological settings.

Evaluations

Approach

We use simulation studies in this paper to make the problems being confronted clear and to evaluate proposed remedies. We consider two basic models, a multiple regression and structural path model, so that our evaluation applies to a range of circumstances. We initially describe these models in conventional linear form for comparison with the standard Gaussian results. For more discussion of the contrast between regression and causal models, refer to Lindley (2002) or Pearl (2009).

A Regression model with two predictors.—Consider a case where a focal response of primary interest, y, is hypothesized to be associated with two predictors, x and z (Fig. 3A). This simple two-predictor regression, omitting the intercept, can be represented by the following formula (Eq. 16).

$$y = \beta_{yx\cdot z}x + \beta_{yz\cdot x}z + \varepsilon_y \tag{16}$$

There are two possible interpretations for coefficients, depending upon whether we are claiming that descriptive or causal (structural) relationships are encoded in the model. In descriptive regression models, we would interpret $\beta_{yx\cdot z}$ as the predicted change in y we would expect to observe with a single unit increase in x, controlling for the value of z. The intended point being made by the notation here is that the relationship is associational, one expects certain values of y when they see certain values of x and z. In contrast, for structural models, we would interpret $\beta_{yx\cdot z}$ as the predicted change in y we would observe whether we were to physically change the value of x by one unit, while holding z constant (i.e., as a causal effect). Pearl (2009) refers to this as $p(y|do(x,z))$. Interpretations for ε_y also differ between descriptive and causal models. For descriptive models, we interpret the error term in Eq. 16 as representing the residual deviations between predicted and observed values. For causal models, the error term represents the combined effects of unspecified causal factors. Causal models are accompanied by a significant number of assumptions that are required to be met if the coefficients are to be interpreted as unbiased causal effect estimates (i.e., absent confounding due to back-door relationships). In this paper, we assume that the regression models examined are descriptive equations, while the structural equation models (described below) are intended to represent a causal hypothesis. However, for simplicity, we will use the term "effect" to cover both cases: For the descriptive case, we mean the effect on y that we expect to see when we observe a different value of x, while for the causal case, we mean the anticipated effect of physically changing the value of the predictor on the value of the response.

The coefficients returned from an OLS solution of the model represented in Eq. 16 possess a number of familiar and useful characteristics. The coefficients, along with the predictor variables, are estimated so as to form a vector of predicted y scores, $\hat{y}$, that permit the minimization of $\text{Var}(\varepsilon_y)$. In addition to assuming that errors are

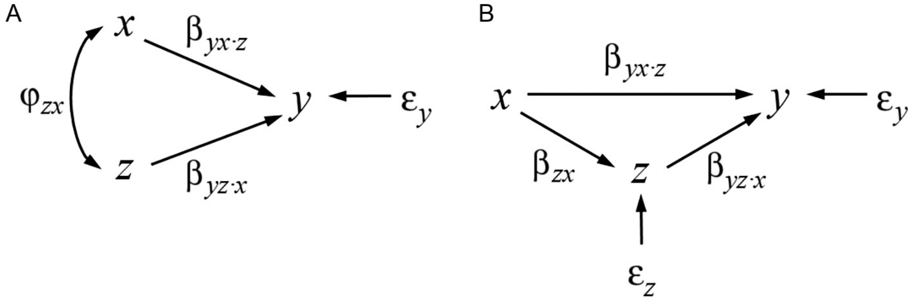

Fig. 3. Generic representations of (A) regression and (B) structural equation path models.

independent from each other, we also assume that Cov($\hat{y}$, ε_y) = 0 (i.e., we expect errors to be uncorrelated with predictor scores, which requires predictors to be measured without error and unspecified causes to be orthogonal to the included causes).

Above and beyond the basic assumptions described above, one expectation we have when analyzing data is that an adequate model should be able to recover the true parameter values associated with the data-generating process (which are known in simulation studies). Following Boos and Stefanski (2013: Section 6.3.1) and Faraway (2014: Section 2.9), we say that a model is identified or fully identified if we can recover the true parameter values using an appropriate analysis method. The easiest way to verify such a claim is by demonstrating that the parameter value used to simulate data given some data-generating assumptions can be recovered by an analysis model (e.g., Kéry and Schaub 2012). The term identified is also sometimes used in specific contexts (especially SEM circles) to refer to cases of partial identification where parameter values can be obtained, but their relationship to the true parameters of the data-generating process is unknown (e.g., there are unknown scaling factor).

To illustrate the expected behavior of models at large sample sizes, we used a single simulation of 20,000 (MacKinnon et al. 2007). Details of the simulation studies can be found in Appendix S1. For simulation study #1, we generated independent random normal vectors for x and z that were shown to be uncorrelated. We then simulated values of y using the following settings: intercept = 0, $\beta_{yx\cdot z}$ = 0.60, and $\beta_{yz\cdot x}$ = 0.80 (these are listed as input parameter values in Table 1). To the resulting expected values, we added random normal errors with σ_ε = 5.0. We then performed OLS linear regression on the simulated data using the lm function in the R base package (R Core Team 2017). For the full model, the statement used was lm(y ~ x + z). The recovered estimates are as shown in Table 1 under the heading, full model. It can be seen that, as expected, estimates recovered were very close to the true (input) values.

Table 1. Inputs and results from simulation studies #1 and #2.

Model	Input value (β or ε)	Recovered value (*b* or *e*)
Gaussian regression example: simulation study #1		
Full model		
$\beta_{yx\cdot z}$ or $b_{yx\cdot z}$	0.60	0.613
$\beta_{yz\cdot x}$ or $b_{yz\cdot x}$	0.80	0.805
$\sigma_{\varepsilon y}$ or σ_{ey}	5.00	5.015
Single-predictor models		
β_{yx} or b_{yx}	0.60	0.616
$\sigma_{\varepsilon y}$ or σ_{ey}	5.00	6.442
β_{yz} or b_{yz}	0.80	0.806
$\sigma_{\varepsilon y}$ or σ_{ey}	5.00	5.343
Gaussian path model example: simulation study #2		
Full model		
β_{zx} or b_{zx}	0.08	0.080
$\beta_{yx\cdot z}$ or $b_{yx\cdot z}$	0.20	0.198
$\beta_{yz\cdot x}$ or $b_{yz\cdot x}$	2.00	2.017
$\sigma_{\varepsilon z}$ or σ_{ez}	0.25	0.247
$\sigma_{\varepsilon y}$ or σ_{ey}	0.75	0.751
Path relations		
Direct effect: $\beta_{yx\cdot z}$ or $b_{yx\cdot z}$	0.20	0.198
Indirect effect: $\beta_{zx} \times \beta_{yz\cdot x}$ or $b_{zx} \times b_{yz\cdot x}$	0.08 × 2.0 = 0.16	0.080 × 2.017 = 0.160
Total effect = direct + indirect	0.36	0.358
Reduced-form model		
β_{yx} or b_{yx}	0.36	0.358

Notes: Full models and parameter labels are shown in Fig. 3. For these simulations and analyses, response variables were created and modeled as Gaussian.

There are other expectations we would have for a linear Gaussian model with multiple predictors, as in simulation #1. For example, if x and z are completely orthogonal (uncorrelated), their simple coefficients β_{yx} and β_{yz} (marginal relations) should make stable, independent contributions to the model predicted scores; that is, we would anticipate $\beta_{yx \cdot z} = \beta_{yx}$ and $\beta_{yz \cdot x} = \beta_{yz}$. This coefficient property is one manifestation of the desired property of collapsibility. To demonstrate coefficient collapsibility, we analyzed the data from simulation #1 using single-predictor models lm($y \sim x$) and lm($y \sim z$). Comparisons of the coefficients recovered from these models with those from the full model (Table 1) show that coefficients were stable, that is, collapsible, as the coefficients from each single model were the same as those from the multiple regression. The recovered error estimates, however, were larger than those used to simulate the data because omitting predictors increased prediction error.

A simple structural equation model.—In order to progress from descriptive regression models to models more appropriate for causal inference, it is necessary to permit a full causal ordering of variables (Grace et al. 2014). To do this, we must shift from the classical statistical model $\mathbf{Y} = f(\mathbf{X})$, where $\mathbf{Y}$ is a vector of one or more response variables and $\mathbf{X}$ is a vector of potentially intercorrelated predictor variables, to the structural equation model $\mathbf{Y} = f(\mathbf{X}, \mathbf{Y})$. Once we are able to represent network hypotheses by allowing y (response) variables to depend on other y variables, a new set of questions and coefficient comparisons become possible (Grace et al. 2012). We consider some of these questions and comparisons in a second simulation study (simulation study #2) where z is now a mediator of part of the effect of x on y (Fig. 3B). In particular, we wish to compare the effect of x on y through z (an indirect effect in the model) to the remaining (direct) effect that operates independent of z.

The two predictors of y in this model, the x and z variables, are no longer orthogonal in this situation, in contrast to the data generated in simulation #1. Another difference from simulation #1 is that there are two equations required to represent the relationships among variables in this model.

$$z = \beta_{zx}x + \varepsilon_z \qquad (17)$$

and

$$y = \beta_{yx \cdot z}x + \beta_{yz \cdot x}z + \varepsilon_y \qquad (18)$$

When responses are linear and Gaussian, the coefficients recovered can be interpreted in the usual fashion, as expected change in the response variable if the predictor were increased by 1 unit while other factors are fixed. We also expect that the strength of an indirect effect can be computed as the product of the coefficients along that path. Simulation study #2 provides us with an opportunity to illustrate these points. As presented in Table 1 under the section *Path Model Example*, we are again able to recover the input parameter values. This was true for both coefficients and for error standard deviations. Coefficients recovered were used to compute direct, indirect, and total effects and compared to expectations based on input parameters. To do this, the indirect effect of x on y mediated by z was computed by calculating the product of the coefficients along the indirect path; then, to that we added the direct effect, yielding an estimate of the total effect of x on y.

There exists a so-called reduced form of the path model that absorbs the effects of intervening variables. This can be thought of as the net effect of x on y. In this example, the reduced-form model can be represented as in Eq. 19.

$$y = \beta_{yx}x + \varepsilon_{y.\mathrm{net}} \qquad (19)$$

In order to distinguish the errors in Eq. 19 from those in Eq. 18, we use the notation $\varepsilon_{y.\mathrm{net}}$. The reduced form of the model returns a coefficient value (0.357), which is very close to the total effect reported for the mediation model in the linear Gaussian case (0.358). This demonstration illustrates the property of decomposability of net/total effects into direct and indirect components (an effect is the sum or its parts). In summary, regression and path models that contain Gaussian responses produce coefficients that demonstrate a number of consistent properties (collapsibility and decomposability). In the next section, we show precisely how these properties change when we are modeling binary responses.

Evaluation of results under the latent propensity conceptualization

In this section, we take a similar approach to that used above for regression and path models having Gaussian responses. For this illustration, we work with binary response models having the same underlying relationships as used in simulations 1 and 2, but with binary (0/1) observations determined by the application of a threshold cutpoint, as is classically assumed for this model (Fig. 2). Our goal is to determine the capacity of BRMs to allow for interpretations similar to those from a Gaussian response model as some have proposed. The goal of this section is to make clear what it means to say that under the latent propensity model, the true parameter values are unidentified (or, more properly, not fully identified). For reference, we show results in Appendix S3, based on simulations #3 and #4 (Appendix S1), that for data generated under traditional assumptions for logit model, input parameter values are recovered by logistic regression.

To represent the latent propensity model, we may create a model for binary responses using a variant of the general linear model where y is replaced by y^*, a continuous latent variable representing the propensity for an occurrence. For this situation, we convert the model used in simulation #1 (Eq. 16) to a latent version of that model (Eq. 20). Again, a key difference from traditional logit response models is the assumption that binary observations emerge from application of a threshold (Long 1997) rather than because observations are drawn from a binomial distribution.

$$y^* = \beta_{y^*x\cdot z}x + \beta_{y^*z\cdot x}z + \varepsilon_{y^*} \qquad (20)$$

The linkage between y^* and y is further understood by recognizing that the expected values of y in Eq. 16 (the solid line in Fig. 1) are a range of values from a nonlinear probability distribution spanning from 0 to 1.

$$E(y=1) = \Pr(y=1) \qquad (21)$$

For the latent-linear model (Eq. 20), it is important to recognize that neither y^* nor ε_{y^*} are directly estimable. The reason for this is that there exists an infinity of combinations of y^* and ε_{y^*} values that are consistent with the data. This condition of non-identification is partially solved by making assumptions about error distributions and then fixing the error variance to a value corresponding to the distribution form. Making these assumptions allows us to recover estimates (i.e., achieves partial identification of the model). However, the values recovered are not estimates of the true underlying parameters (as we show below). Despite that, the adoption of arbitrary assumptions about error variance, while they do not identify individual estimates for y^* and ε_{y^*}, nor estimates of true effects, they do provide us with values for variance components for all the terms in the model (Eq. 22).

$$\mathrm{Var}(y^*) = \mathrm{Var}(\beta_{y^*x\cdot z}x + \beta_{y^*z\cdot x}z) + \mathrm{Var}(\varepsilon_{y^*}) \qquad (22)$$

If we assume a logit error distribution, the standard assumption is that $\mathrm{Var}(\varepsilon_{y^*}) = \pi^2/3$. For the probit distribution, the standard assumption is that $\mathrm{Var}(\varepsilon_{y^*}) = 1$ (Long 1997).

To restate, assuming an error variance value allows for partial identification, which is sufficient to estimate variance components, but not causal parameter estimates for the y^* model. We must now recognize that because of the causal non-identification, the coefficients recovered by the software are not the true parameter values; instead, the whole model has been rescaled by a scaling factor (σ_u) that converts the true error standard deviation to the fixed value (Breen et al. 2013). Unfortunately, the value of the scaling factor is unknown. If we assume a logit distribution, what we now have is as shown in Eq. 23. Here, we use the notation of $b_{yx\cdot z}$ and $b_{yz\cdot x}$ to refer to the *observed* coefficients, while $\beta_{y^*x\cdot z}$ and $\beta_{y^*z\cdot x}$ refer to the *true* latent coefficients.

$$\begin{aligned}\mathrm{logit}[\Pr(y^* > 0)] &= b_{yx\cdot z}x + b_{yz\cdot x}z \\ &= \frac{\beta_{y^*x\cdot z}}{\sigma_u}x + \frac{\beta_{y^*z\cdot x}}{\sigma_u}z \qquad (23)\end{aligned}$$

As Eq. 23 makes explicit, attempts to use binary regression models such as the logistic and probit to return coefficient estimates consistent with a latent propensity conceptualization will fail because the assumed parameters are scaled by some unknown quantity (σ_u). This causes no problems for the identification of probabilities or log odds ratios associated with traditional binomial models because the scaling cancels out. To make the problem more explicit, we illustrate the issues with simulation studies #5 and #6.

*Latent regression.—*In simulation study #5, we return to the data obtained from simulation #1. We first recast the continuous version of y as y^* and treat it as a latent quantity. We then create observed binary outcomes using a cutpoint threshold and center the y^* values. We now refer to the observed 0s and 1s as *y*. This means we use the data-generating assumptions associated with the latent propensity model in this case. In all other respects, the simulation input parameters are the same as used in simulation #1. Details of the simulation are in Appendix S1.

As can be seen in the results reported in Table 2, the true parameters for simulation study #5 (those used to simulate the y^* values) are *not* recovered from the GLM analysis of the data (e.g., the first input value of 0.60 leads to an estimated value of 0.205). The value for the error standard deviation is the same for all models (1.814) because it is fixed by assumption. Because the error variance is a fixed quantity, the variance of y^* varies among models as the predictor variance changes. This results in inconsistent estimates of coefficients across models even though the predictors (*x* and *z*) in the regression case are orthogonal (i.e., coefficient estimates recovered exhibit non-collapsibility).

*Latent path model.—*As the results from simulation #5 made explicit, binary regression models based on latent propensity assumptions and using cutpoints do not return the true parameter values. This is again observed when a path model including mediation (simulation study #6) has a binary response (compare input and recovered values for parameters in Table 2).

Given that we cannot recover the true underlying coefficients in a latent y^* model, there are still opportunities to develop approaches to interpreting the coefficients that are recovered. Before we examine those, we here illustrate one more manifestation of non-identification—the recovery of estimated coefficients that are consistent with more than one underlying model. This is easily understood from the relationships in Eq. 24, which derive from Eq. 23. Again, the coefficients $b_{yx\cdot z}$ and $b_{yz\cdot x}$ refer to the observed coefficients, while $\beta_{y^*x\cdot z}$ and $\beta_{y^*z\cdot x}$ refer to the true coefficients.

$$b_{yx\cdot z} = \frac{\beta_{y^*x\cdot z}}{\sigma_u}; b_{yz\cdot x} = \frac{\beta_{y^*z\cdot x}}{\sigma_u} \qquad (24)$$

In Appendix S1, we present results for one additional simulation study #7, in which the

Table 2. Simulation studies #5 and #6: input parameters and recovered raw parameter values.

Model	Input value (β or ε)	Recovered value (*b* or *e*)
Latent regression example: simulation study #5		
Full model		
$\beta_{yx\cdot z}$ or $b_{yx\cdot z}$	**0.60**	**0.205**
$\beta_{yz\cdot x}$ or $b_{yz\cdot x}$	**0.80**	**0.271**
$\sigma_{\varepsilon y}$ or σ_{ey}	**5.00**	**1.814**
Single-predictor models		
β_{yx} or b_{yx}	**0.60**	**0.151**
$\sigma_{\varepsilon y}$ or σ_{ey}	**5.00**	**1.814**
β_{yz} or b_{yz}	**0.80**	**0.251**
$\sigma_{\varepsilon y}$ or σ_{ey}	**5.00**	**1.814**
Latent path model example: simulation study #6		
Full model		
β_{zx} or b_{zx}	0.08	0.080
$\beta_{yx\cdot z}$ or $b_{yx\cdot z}$	**0.20**	**0.463**
$\beta_{yz\cdot x}$ or $b_{yz\cdot x}$	**2.00**	**4.841**
$\sigma_{\varepsilon z}$ or σ_{ez}	0.25	0.245
$\sigma_{\varepsilon y}$ or σ_{ey}	**0.75**	**1.814**
Path relations		
Direct effect: $\beta_{yx\cdot z}$ or $b_{yx\cdot z}$	**0.20**	**0.463**
Indirect effect: $\beta_{zx} \times \beta_{yz\cdot x}$ or $b_{zx} \times b_{yz\cdot x}$	**0.08 × 2.0 = 0.16**	**0.080 × 4.841 = 0.385**
Total effect = direct + indirect	**0.36**	**0.848**
Reduced-form model		
β_{yx} or b_{yx}	0.36	**0.689**

Notes: Full models and parameter labels are shown in Fig. 3. For these simulations, binary responses were generated from a continuous underlying propensity using a cutpoint threshold = mean(y^*). Analysis models were of the form glm($y \sim x + z$, family = binomial, link = "logit"). Input and recovered values that are markedly different are indicated in bold. For additional detail, see Appendix S1.

input coefficients (the βs) are doubled, while at the same time, the input error standard deviation σ_ε is doubled. The returned coefficients (*b*s) are identical to those found in simulation study #6 even though the input coefficients are twofold different, demonstrating again, but in a different fashion, that the true parameters are unidentified under the latent propensity model.

Finally, the situation can be seen to be worse for path models where mediators are binary. Here, there will be separate unknown scaling factors for both the binary mediator and binary terminal response variables. Computed indirect effects (IE) will now be scaled by the products of two unknown scaling factors, as shown in Eq. 25 (here, z^* denotes the fact that true z is assumed to be latent in this case). The result will be that the direct effect of x on y cannot be compared to its indirect effect mediated through z even though y is a common response for both effects in a single model (the limited case where some have suggested relative comparisons of unstandardized effects can be made). As we will show below, standardization of coefficients can provide a basis for coefficient comparison in this case.

$$\mathrm{IE}_{yx} = b_{zx} \times b_{yz\cdot x} = \frac{\beta_{z^*x}}{\sigma_\omega} \times \frac{\beta_{y^*z\cdot x}}{\sigma_u} = \frac{\beta_{z^*x} \times \beta_{y^*z\cdot x}}{\sigma_\omega \times \sigma_u} \quad (25)$$

Performance of Standardized Coefficients in Ecological Applications

Comparisons of interest

As stated earlier, the goal of standardizing coefficients is to make them comparable. Comparability is sought regardless of whether responses are Gaussian or binary and irrespective of the distributional forms of predictors. There will also be cases where there is interest in comparing coefficients across models estimated based on different samples.

Standardization methods used in this evaluation include (1) no standardization, (2) latent-theoretical standardization, (3) observed-empirical standardization, (4) relevant-range standardization under the latent-theoretical framework, and (5) relevant-range standardization under the observed-empirical framework. To keep the presentation manageable, we will forego providing illustrations of x-only and y-only standardizations. It is our hope that explicit considerations of interpretations for raw coefficients and completely standardized coefficients will allow the reader to infer the interpretations that would apply to x-only and y-only versions of the standardizations. Here, we provide distilled results for several worked examples reflecting a range of comparisons and standardization methods. R code for these examples is in Appendix S2 and data provided in the Supporting Data file.

One limitation to basing conclusions about the supported comparisons based on empirical examples is the relatively small number of samples associated with those examples. It seemed useful to confirm that the observed-empirical (OE) method of standardization can produce coefficients that can be used for comparing total effects of a full mediation model with the net effects from a reduced-form model. Simulation study #4 provided us with such data, so we applied the OE standardization method to those data. Details of this analysis are given at the end of Appendix S2. Analysis of the simulated data showed excellent comparability between total and net effects estimates.

Examples

In the following section, we provide several worked examples. The data and code to reproduce the examples can be found in the supplementary material.

Example #1: Exotic bird invasion success—logistic regression with standardized comparisons between predictors.—This illustration is derived from the work described in Veltman et al. (1996). Their original study examined the correlates of introduction success for exotic birds in New Zealand. Data were obtained for 79 species involved in 496 introduction events. Status (locally extant or extinct) at the time of the study was used as a measure of success. Predictors of success examined included both traits of the species, such as whether they were migratory or not, and characteristics of the introduction effort, such as number of individuals released. The illustration presented here involves predicting invasion success (status) as a function of whether species frequent uplands (upland), whether the species is migratory or not (migratory), and the minimum number of individuals introduced (indiv).

Additional details of the example are in Veltman et al. (1996). Main results are presented in Table 3.

Unsurprisingly, the raw coefficients provide little information regarding the relative importance of factors predicting invasion success. Standardized coefficients, however, provide for some comparability of results. The rank order of effect strengths (ignoring sign of effect) is the same across all standardizations. Results suggest the number of individuals released is by far the most important predictor, while migratory status is the least important. Latent-theoretical coefficients were consistently larger (in absolute values) than observed-empirical ones, which we expect to be generally true based on their computations.

Construction of relevant ranges was an instructive process. Upland use is a binary variable (0,1) and was given a relevant range = 1. Migratory behavior is a three-level categorical variable (1,2,3) and was given a relevant range of 2. Individuals released ranged from 2 to 1539. Based on inspection of distribution (highly skewed) and data plots, this predictor was given a relevant range = 1537, which is the empirical difference in the ranges. Ratios of relevant range to standard deviations varied from 4.7 to 2.6, nearly a twofold difference. For comparison, we would expect a ratio of approximately six for a Gaussian variable. For LT and OE methods, individuals released were twice as important as upland use. Use of relevant ranges shifted that to three times as important. In general, we feel there is conceptual merit to the relative range method in this case as it solved the problems of making comparisons based on standard deviations for binary, categorical, and continuous predictors.

Table 3. Results for example 1: logistic regression of exotic bird invasion success (see Appendix S2).

Standardization method	Upland use	Migratory	Individuals released
None	−5.72	−1.53	0.012
Latent-theoretical	−0.39	−0.24	0.80
Observed-empirical	−0.31	−0.19	0.64
Relevant-range, latent-theoretical	−0.20	−0.11	0.62
Relevant-range, observed-empirical	−0.16	−0.09	0.50
Relevant range-to-std. dev. ratio	3.1	2.6	4.7

Example #2: Wildlife hotspots in the Serengeti of Africa—comparing the strengths of direct and indirect paths in a mediation model.—While the Serengeti in Africa is famous for the vast migrating herds that follow the seasonal rain (often featured in nature films), less well known are local wildlife hotspots or concentrations of non-migrating herbivores. Theories attempting to explain these local concentrations of grazers mostly relate to either features required for predator avoidance (low-statured grasses and associated high visibility of approaching predators) and those providing for high-quality forage (short-statured vegetation with associated high leaf nitrogen concentration). With this dataset, we wish to ask "How much of the influence of low-statured vegetation on hotspot location results from the associated higher leaf nitrogen concentration, as opposed to the avoidance of predators due to better visibility in low grass?"

For this illustration, we use four variables from the original study by Anderson et al. (2010), (1) herbaceous biomass in kg/m^2 (biomass), (2) leaf nitrogen concentration (LeafN), (3) position in the local landscape (LocLand), and (4) hotspot designation, yes = 1, no = 0 (Hotspot). With reference to the generic mediation model in Fig. 3B, in this example Biomass is the x variable, LeafN is the z variable, and Hotspot is the y variable. LocLand was included as a covariate variable c that points at Hotspot. Including LocLand in the model controls for local conditions that are favorable for predators (proximity to patches of woodlands and sources of water) that grazers avoid. Including the LocLand control variable in the model clarifies the other signals significantly. Relevant ranges in this example were, as in the previous example, based on the observed ranges of values. The main results of the analysis are given in Table 4.

Here, the direct effect is stronger than the indirect effect, regardless of the standardization method. This implies the influence of prey being able to see predators over long distances is more important than leaf nitrogen content in the probability of a location being a wildlife hotspot. All coefficient types support this interpretation. The reduced-form model coefficient (net effect) matched the total effect computed for the full mediation model closely. This supports the conclusion drawn from the simulation studies that all forms of standardization examined permit

Table 4. Results for hotspots mediation path model.

Standardization method	B→H	N→H	L→H	B→N	DE	IE	TE	NE
None	−7.78	6.69	1.36	−0.49	−7.78	−3.28	−11.1	−8.41
Latent-theoretical	−0.41	0.34	0.63	−0.50	−0.41	−0.17	−0.58	−0.59
Observed-empirical	−0.31	0.26	0.49	−0.50	−0.31	−0.13	−0.44	−0.44
Rr latent-theo.	−0.32	0.25	0.42	−0.55	−0.32	−0.14	−0.46	−0.47
Rr observed-emp.	−0.25	0.19	0.33	−0.55	−0.25	−0.10	−0.35	−0.35

Notes: H, hotspot (0/1); B, biomass; N, LeafN; L, location; DE, direct effect; IE, indirect effect; TE, total effect; NE, net effect; RR, relevant range. (For details, see Appendix S2). The ratio of relevant range to standard deviation was 4.73 for biomass and 4.33 for nitrogen.

comparisons across models built on a single dataset. Coefficient values were again higher from the LT method than from the OE method.

Ratios of relevant ranges to standard deviations were 4.73 and 4.33 for biomass and leaf nitrogen. The use of relevant ranges in place of standard deviations changes our perceptions of the relative importance and even the rank order of component effects. For the LT standardization, the L→H effect was the strongest in the model, while for the OE standardization, L→H and B→N are equivalent. Using relevant ranges reduced the relative importance of B→H, N→H, and L→H effects relative to the B→N effect. Again, we feel the relevant-range-standardized coefficients are more defensible in this example, as standardization using standard deviations inflated coefficients due to their platykurtic shapes.

Example #3: Environmental requirements for the population viability of a species of bats—binary path model with binary mediator.—Path models composed entirely from binary variables are uncommon in the natural sciences in our experience. It would perhaps be most likely that we would encounter such models treated within the framework of Bayesian networks (Pearl 1988, Pourret et al. 2008). Bayesian networks are well suited for modeling probabilistic relations among networks of discrete variables. However, because they summarize interrelations using discrete conditional probability tables, they are challenged when dealing with data involving continuous variables. In this example, we consider whether path models made up of binary variables can be analyzed using a structural equation approach and their effects represented using the standardization approaches demonstrated in this paper.

The Townsend's big-eared bat is a species of conservation concern in the Columbia River Basin in the NW USA. Marcot et al. (2001) assembled a database of conditional relationships between expert assessments of potential population responses by the bats and a host of environmental and management factors. Here, we develop an example three-variable model inspired by their study. Because their model was based on summaries of data represented as conditional probabilities, we simulated binary data possessing similar properties and interrelations for this illustration. The variables included (1) population responses by bats (viability), (2) the existence of caves in the area (caves), and (3) suitable hibernation locations (partially dependent on presence of caves, but with other requirements as well; Hibernicula). For all variables, a value of 1 refers to either a positive population response or conditions suitable for a positive response. With reference to Fig. 3B, the *x* variable in this example is caves, the *z* variable (mediator) is hibernacula, and the *y* variable is viability.

Results for this example (Table 5) indicate that the role of caves as places for developing hibernacula (suitable hibernation locations) is not sufficient to explain the total importance of caves to bats. Marcot et al. (2001 and references therein) discuss the additional roles of caves for this species. Here, we point out that our perception of the importance of the hibernaculum mechanism varies depending on coefficient standardization method. When we use relevant ranges, all effects are perceived to be smaller. The reason behind that difference is instructive, as this is an example where regular standardization using standard deviations would produce coefficients with upward bias. This is the case because the standard deviations of binary variables are large relative to their ranges (~0.5) compared to Gaussian variables (~0.15). Both the LT and OE methods

Table 5. Results for Example 3: bat population viability mediation model involving all binary variables.

Standardization method	C→V	H→V	C→H	DE	IE	TE	NE
None	1.12	1.37	1.89	1.12	2.59	3.71	1.46
Latent-theoretical	0.26	0.25	0.44	0.26	0.11	0.38	0.37
Observed-empirical	0.22	0.22	0.37	0.22	0.08	0.31	0.33
Rr latent-theo.	0.09	0.11	0.15	0.09	0.02	0.11	0.12
Rr observed-emp.	0.08	0.09	0.13	0.08	0.01	0.09	0.11

Notes: V, viability (0/1); C, caves (0/1); H, hibernacula (0/1); DE, direct effect; IE, indirect effect; TE, total effect; NE, net effect; RR, relevant range. (For details, see Appendix S2). The ratio of relevant range to standard deviation was 2.00 for caves and 2.26 for nitrogen.

imply standard deviations for the *y* variables consistent with Gaussian continuous variables. Thus, the combination of binary predictors and implied continuous values for *y* variables inflates coefficients when adjustments are based on standard deviations.

Evaluation of internal consistency for standardization methods.—Discussions of the merits and standards for standardization methods frequently emphasize the desirability of a consistent ranking of the relative importance of different effects (Breen et al. 2013). To consider the results obtained from our assessments, we computed scaled coefficients such that the largest coefficient in a model was assigned a value of 1 and the other coefficients were adjusted proportionally. Using this adjustment, we can see in Table 6 that for Example 1, logistic regression of bird invasion success, the differences we see between LT and OE coefficients largely disappear when expressed as proportions of the maximum value. However, for Examples 2 and 3, where coefficients are not all of the same raw type (some in logits, others in linear unit changes), LT and OE coefficients showed differences beyond those that could be attributed to simple scaling effects. That said, rank order among coefficients within a model was the same across standardization methods in all cases except one (in Example 2, LT differed from the others).

Comparing coefficients across groups—Example #4: environmental controls of plant diversity.—Perhaps one of the most challenging situations for comparing standardized coefficients is when comparing across groups. The fundamental problem is that variances, and therefore standard deviations, will not be the same for each group. This erodes the comparability of coefficients because the quantities used for standardization are not comparable. It is possible for this case to develop relevant ranges that can be used to compare range-standardized coefficients across groups. We present such an example in Appendix S4 to illustrate the method.

Discussion

It is perhaps surprising how difficult it can be to interpret the coefficients derived from binary response models. Ultimately, the challenge arises because there is a loss of information. Absent a continuous distribution of observed values, the task of summarizing our understanding is shifted from building a model for a continuum of predicted values to building a model for the conditional probabilities of ones and zeros. This

Table 6. Comparison of standardization methods based on the proportion of the largest standardized coefficient value within a model.

	Example 1			Example 2				Example 3		
Standard. method	Upland use	Migratory	Indivs. released	B→H	N→H	L→H	B→N	C→V	H→V	C→H
Latent-theoretical	0.49	0.30	1.00	0.65	0.54	1.00	0.79	0.59	0.65	1.00
Observed-empirical	0.48	0.30	1.00	0.62	0.52	0.98	1.00	0.70	0.76	1.00
Rr latent-theoretical	0.32	0.18	1.00	0.58	0.45	0.76	1.00	0.60	0.73	1.00
Rr observed-empirical	0.32	0.18	1.00	0.45	0.35	0.60	1.00	0.73	0.82	1.00

Notes: Labels for Example 1 are from Table 3, labels for Example 2 are from Table 4, and labels for Example 3 are from Table 5. Values presented are absolute (ignoring sign).

change produces a model form in which predictor effects on observation responses are inherently nonlinear. Despite this, scientists have the same questions as when dealing with linear Gaussian models. Multivariate models containing a mixture of response types, which are increasing in use, pose the greatest challenges for interpretations.

In this paper, we have provided a variety of ecological examples in illustrating the points being made. We do this because analyses of the literature suggest that scientists best see the value of a statistical approach when presented with examples similar to their own studies (Grace 2015). The literature also suggests that the recommended approaches for interpreting BRMs vary substantially depending on the scientific discipline. Most notable, there is evidence of the historical isolation between the social and natural sciences evident in cross-citation analyses (http://www.eigenfactor.org/map/maps.php). There also exist distinctly different practices among disciplines within the social sciences. In this paper, we have tried to draw on methodological perspectives from all disciplines, with the purpose of providing ecologists with a broad view of the possibilities. For our summary of the take-home points, we introduce one final ecological example appropriate for illustrating a wide variety of comparisons.

Example 5: Resilience of prairie grasslands to invasion by exotic trees

Grassland conversion to shrubland and forest is a global issue that has been widely discussed. To support our discussion, we consider the invasion of tallgrass prairie in the southern United States by the non-native tree known as Chinese tallow (Grace et al. 2005, Siemann and Rogers 2006). The basic situation is that the grasslands in this example have historically existed as prairie in large part because of frequent fires. When tallow invades, it is initially susceptible to control by fire, but as the trees get larger, they are both capable of resprouting after fire and also capable of suppressing the herbaceous fuel beneath their crowns. Once tallow becomes sufficiently established, the fire-maintained prairie is converted to non-flammable forest. Thus, the prairie grassland can be judged to be resilient to tallow invasion as long as trees are effectively reduced by fire. Here, we utilize data from an experimental burning study (Grace et al. 2005) to model the dependence of prairie resilience on the combined effects of (1) tree size prior to burning, (2) herbaceous community biomass, which fuels the fires, and (3) whether fires occur during the growing or dormant season (Fig. 4). In this case, resilience is a binary response (Y/N), tree size and community biomass are zero-limited continuous variables, and both season of burn and whether a fire is complete or not are binary variables. Coefficients of various types are given in Table 7.

Raw vs. standardized coefficients

As stated earlier in the paper, it is sometimes recommended that scientists interpret the coefficients from BRMs in the scale of the log odds ratios because these represent the linear effects of predictors in logistic regression. Scientists who are charged with drawing substantive interpretations are frequently uncomfortable with the log odds ratio scale. Of course, we can, through exponentiation, convert log odds ratios to probabilities. However, now the coefficients represent nonlinear effects of predictors and are therefore not constant, but, instead, dependent on the value of the xs. This presents an additional challenge for comparing effects across β_1–β_3 and between β_4 and β_5 because the shape of the nonlinear relationship underlying each coefficient

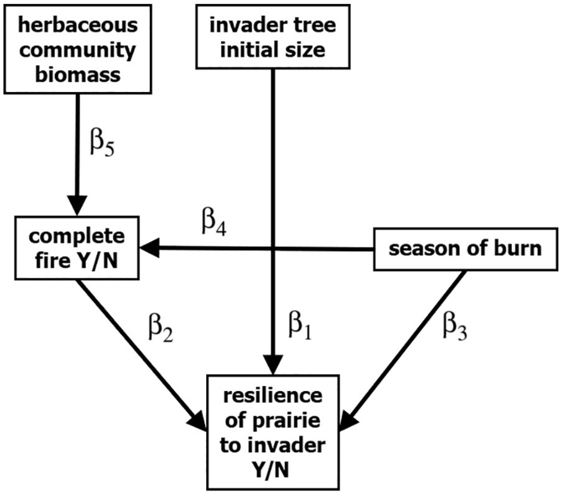
herbaceous community biomass
invader tree initial size
β5
complete fire Y/N
β4
season of burn
β2
β1
β3
resilience of prairie to invader Y/N

Fig. 4. Path model for prairie resilience example (Example #5, created from Grace et al. 2005).

Table 7. Coefficients of various types for Example 5: prairie grassland resilience as a function of invader size, fire completeness, and season of burn.

Coefficient type	β_1	β_2	β_3	β_4	β_5
Raw coefficients					
Estimates	−0.570	2.738	1.493	−1.364	0.524
Std. error	0.349	0.337	0.298	0.234	0.148
P-value	<0.001	<0.001	<0.001	<0.001	<0.001
Standardized coefficients					
Latent-theoretical	−0.40	0.54	0.30	−0.34	0.22
Observed-empirical	−0.36	0.48	0.27	−0.30	0.20
Relevant-range, latent-theo.	−0.29	0.19	0.10	−0.11	0.15
Relevant-range, observed-emp.	−0.26	0.17	0.09	−0.10	0.13

Note: Symbols are shown in Fig. 4.

can be unique. In our invasive tree example (Fig. 4), both endogenous variables, resilience and burn completeness, were binary; therefore, all raw coefficients represent unit changes in log odds ratios predicted for unit changes in predictors. In principle, β_1–β_3 are inter-comparable because they involve a common response variable (and therefore a common ratio between assumed error variance and true error variance). However, because log odds ratios are not comparable across submodels with different response variables, comparing β_1–β_3 to β_4 and β_5 would be improper, despite the fact that the units of y are technically the same.

Additional challenges for interpreting raw coefficients are generated by differences in scales of the predictors. For two of the predictors in the prairie resilience example, fire completeness and season of burn, unit change equals the full range of possible values (from zero to one). Tree size, in contrast, ranges from 0.12 to 7.4 meters tall, so a single unit of change represents only 13.5% of the observed range. Thus, we cannot meaningfully compare β_1–β_3 without at least standardizing the x variables. Similarly, for burn completeness, its two predictors are in very different units. To us, it seems that comparing coefficients across effects in this model requires complete standardization to a common scale.

Latent-theoretical vs. observed-empirical standardization

As our evaluations have led us to conclude that there is a pressing need for coefficient standardization, at least for comparing coefficients, we are then left with considering which approach or approaches to recommend. Various strategies for standardizing the response components of models have been discussed. In the literature, it seems there has been a preference for either assuming an underlying continuous propensity and using a latent-theoretical (LT) standardization method or treating responses as discrete quantities and using an observed-empirical (OE) standardization approach.

The idea behind the LT approach, that there is an underlying continuous propensity or potential that is linearly related to the predictors, is conceptually appealing. For example, it is easy to imagine that prairie resilience is a continuous latent capacity of systems that we learn about through binary state changes. In fact, the raw response in this example is actually a continuous measure of the difference between post-fire tree heights and pre-fire heights. The binary sign of the difference then provides a criterion that is used to judge resilience. The other binary response variable in the model, burn completeness, is also a classification variable, motivated by the strongly bimodal distribution of continuous observations (plots tended to burn either completely or not at all, with only a few intermediate results).

Some authors object to taking the idea of a continuous underlying linear propensity too far and applying it universally for discontinuous responses (Menard 2010). Some binary properties are not so obviously underlain by linear continuous propensities (and more importantly, we rarely know whether the form of the underlying propensity is linear, nonlinear, or discontinuous). The OE method provides for an alternative approach to standardization that is free of assumptions about continuous underlying states, though not necessarily as broadly relevant. For this method, assuming an approximately linear relation between predicted scores/probabilities and observed values allows for the computation of empirical error variance in the usual OLS fashion.

Our comparisons between LT and OE coefficients (Table 6) suggest that (1) OE coefficients are consistently smaller than LT coefficients and (2) the relative importance rankings are generally the same. The reason OE values are lower would seem to derive from the fact that this method is based on a linear approximation of a nonlinear relationship between y and $\hat{y}$ and therefore understates the predictive efficiency of the coefficients if a link function was used. Aside from that difference, both LT and OE standardizations produce values with useful properties. In fact, our results demonstrate that they possess many of the same properties as standardized coefficients in Gaussian models. In particular, results for Example 2, wildlife hotspots (Table 4), and Example 3, bat population viability (Table 5), demonstrate the property of decomposability that inspired Sewall Wright (1921) to first partition observed associations into pathways representing different hypothesized mechanisms. This property does not hold for raw coefficients under the latent-linear assumption (Table 2).

Standard deviations vs. relevant ranges

We have attempted to add to the options available to the investigator by illustrating select applications of the use of relevant ranges in place of standard deviations. We feel this approach deserves serious consideration for several reasons. Shifting from the arbitrary use of standard deviations to the use of range-related quantities selected by the investigator has the consequence of providing a means for introducing control over interpretations. We have pointed out that the ratio of standard deviation to empirical range for binary predictors is roughly 0.5, while for Gaussian variables, the number is in the realm of 0.20–0.15 depending on sample size. Simulations involving skewed distributions can generate ratios down to 0.05. Making the extreme comparison between binary and highly skewed variables, there can be a 10-fold difference in how many standard deviations can be needed to cover the full range of values upon which a coefficient is computed. Variable distribution shape can thus be seen as a potential source of bias in estimating the relative importance of standardized coefficients, with binary predictors creating upward bias and skewed variables creating downward bias. The added benefit of providing a means to place comparisons among groups with different variances on a common scale is one more way relevant ranges can support coefficient comparison. Of course, transparency is the accompanying ingredient that is required when using relevant ranges. One must report the raw coefficients, standard deviations, the relevant ranges used, and associated justifications.

For the example of prairie resilience, we can see that coefficients based on relevant ranges are not of the same rank order in importance as those based on standard deviations (Table 7). At play here is that the use of standard deviations exaggerate the importance of binary predictors. Thus, standard deviation-based coefficients imply fire completeness is the most important regulator of resilience and that season of burn is more important in regulating fire completeness than is community biomass. However, those comparisons are not on an equal basis from the perspective of relevant ranges. If we use coefficients standardized based on relevant ranges, we can see that tree size can have a substantially greater effect on resilience than either burn completeness or season. Further, burn completeness can be altered to a greater degree by changes in community biomass than by changes in season, though this would not be seen if standard deviations were used.

Summary recommendations

Individuals will likely vary in their preferences, depending on philosophy and circumstances. There are certainly plenty of situations where investigators will wish to focus on interpreting raw coefficients and/or computing the corresponding probabilities. These situations will likely be cases where models are small and coefficient comparisons are limited. For more complex models, and especially ones that include path relations (such as indirect vs. direct effects), standardized coefficients will be essential for comparisons of effect strengths.

We have observed that adopting the assumption of a underlying latent propensity is a choice frequently made by scientists in many fields for binary phenomena. Therefore, we think the latent-theoretical (LT) standardization can be justified for use if it makes sense to the investigator in their situation. To support this assessment, we point out that as far as we are aware, most widely used software that reports standardized coefficients for binary response models relies solely on the LT method. The observed-empirical (OE) standardization we present gives the investigator an option for cases where they wish to treat a response as strictly discrete. However, we must point out that the OE method uses a linear approximation for an inherently nonlinear relationship between predictor and probability of outcome, which means it will commonly underestimate the true relationship. Our comparative results indicate that this underperformance is noticeable; thus, there is a price to be paid (damping of signal) for using the OE method.

We feel that standardizing using relevant ranges instead of using standard deviations provides a useful alternative. When predictors are either roughly Gaussian or of a common distributional shape, the two approaches give very similar results. When predictors have a mix of contrasting distributions, however, the relevant-range formula is more easily justified. The use of standard deviations for standardization when predictors include binary or skewed variables has been criticized for many decades. In fact, many statisticians recommend against standardization entirely because of this problem. We oppose any abolition of standardization, as it would be an impediment to scientific interpretations in a wide range of circumstances. Our results do show, however, that for a model that contains binary and skewed predictors, as much as an order of magnitude bias can result. Careful selection of relevant ranges for use in standardization puts the investigator in conscious control over the interpretive meaning of the standardized coefficients. We understand that this is a new practice and also that it requires scientific judgments to be made that influence the conclusions; however, such a practice avoids unconsciously biasing interpretations and resolves a long-recognized problem with standardization.

Concluding thoughts

This paper was motivated in part by the increase in interest by ecologists in structural equation modeling over the past two decades, coupled with increased use of GLMs in a wide variety of circumstances. There are a number of motivations for using SEM. The most obvious is interest in evaluating causal hypotheses (Shipley 2000). However, there is also an increasingly recognized need to work with complex models that correspond more closely with system-level questions (Mitchell 2001, Forister et al. 2011, Lamb et al. 2014, Eisenhauer et al. 2015, Lefcheck et al. 2015, Fan et al. 2016). These more complex models often contain numerous multivariable pathways of scientific interest and place a premium on broad, general questions about the relative importance of different ecological flows, such as top-down vs. bottom-up control in ecosystems (Du et al. 2015), the relative importance of environmental vs. biotic drivers in community organization (Laughlin and Abella 2007, Byrnes et al. 2011, Cronin et al. 2014, Gama-Rodrigues et al. 2014, Jing et al. 2015), or the relative importance of different biotic processes (Cardinale et al. 2009, Scherber et al. 2010, Mazancourt et al. 2013, Lefcheck and Duffy 2015). For applications such as these, standardization methods that provide clear comparability among coefficients and also that can summarize indirect and total effects become of great importance. We hope that these methods, all of which we have now implemented in the *piecewiseSEM* package (Lefcheck 2016) for the open-source statistical software R (R Core Team, 2017), prove useful for the ecological community in addressing questions of broad interest in a more rigorous fashion.

Acknowledgments

We thank Scott Menard for reviewing an earlier version of the manuscript and especially our treatment of Menard's observed-empirical standardization method. Ken Bollen pointed us to some relevant literature and inspired us to develop detailed explications for the paper. Bob Dorazio and two anonymous reviewers provided helpful comments on and earlier version of the manuscript. JG was supported by the USGS Land Change Science and Ecosystems Programs. Any use of trade, firm, or product names is for descriptive purposes only and does not imply endorsement by the U.S. Government.

Literature Cited

Agresti, A. 2013. Categorical data analysis. Third edition. John Wiley & Sons, Hoboken, New Jersey, USA.

Allison, P. D. 1999. Comparing logit and probit coefficients across groups. Sociological Methods and Research 28:186–208.

Allison, P. D. 2012. Logistic regression using SAS: theory and application. SAS Institute Inc., Cary, North Carolina, USA.

Anderson, T. M., J. G. C. Hopcraft, S. Eby, M. Ritchie, J. B. Grace, and H. Olff. 2010. Landscape-scale analyses suggest both nutrient and antipredator advantages to Serengeti herbivore hotspots. Ecology 91:1519–1529.

Bolker, B. M. 2008. Ecological models and data in R. Princeton University Press, Princeton, New Jersey, USA.

Boos, D. D., and L. A. Stefanski. 2013. Essential statistical inference: theory and methods. Springer Science & Business Media, Berlin, Germany.

Breen, R., K. B. Karlson, and A. Holm. 2013. Total, direct, and indirect effects in logit and probit models. Sociological Methods and Research 42:164–191.

Buckley, Y. M. 2015. Generalised linear models. Pages 131–148 *in* G. A. Fox, S. Negrete-Yankelevich, and V. J. Sosa, editors. Ecological statistics. Oxford University Press, Oxford, UK.

Byrnes, J. E., D. C. Reed, B. J. Cardinale, K. C. Cavanaugh, S. J. Holbrook, and R. J. Schmitt. 2011. Climate-driven increases in storm frequency simplify kelp forest food webs. Global Change Biology 17:2513–2524.

Cardinale, B. J., H. Hillebrand, W. Harpole, K. Gross, and R. Ptacnik. 2009. Separating the influence of resource 'availability' from resource 'imbalance' on productivity–diversity relationships. Ecology Letters 12:475–487.

Cronin, J. P., M. A. Rúa, and C. E. Mitchell. 2014. Why is living fast dangerous? Disentangling the roles of resistance and tolerance of disease. American Naturalist 184:172–187.

Du, X., E. García-Berthou, Q. Wang, J. Liu, T. Zhang, and Z. Li. 2015. Analyzing the importance of top-down and bottom-up controls in food webs of Chinese lakes through structural equation modeling. Aquatic Ecology 49:199–210.

Eisenhauer, N., M. A. Bowker, J. B. Grace, and J. R. Powell. 2015. From patterns to causal understanding: structural equation modeling (SEM) in soil ecology. Pedobiologia 58:65–72.

Fan, Y., J. Chen, G. Shirkey, R. John, S. R. Wu, H. Park, and C. Shao. 2016. Applications of structural equation modeling (SEM) in ecological studies: an updated review. Ecological Processes 5:19.

Faraway, J. J. 2014. Linear models with R. CRC Press, Boca Raton, Florida, USA.

Floyd, T. 2001. Logit modelling and logistic regression: aphids, ants and plants. Pages 197–216 *in* S. M. Scheiner and J. Gurevitch, editors. Design and analysis of ecological experiments. Oxford University Press, Oxford, UK.

Flynn, D. F., N. Mirotchnick, M. Jain, M. I. Palmer, and S. Naeem. 2011. Functional and phylogenetic diversity as predictors of biodiversity–ecosystem-function relationships. Ecology 92:1573–1581.

Forister, M. L., J. A. Fordyce, A. C. McCall, and A. M. Shapiro. 2011. A complete record from colonization to extinction reveals density dependence and the importance of winter conditions for a population of the silvery blue, *Glaucopsyche lygdamus*. Journal of Insect Science 11:art130.

Fox, J. 2016. Applied regression analysis and generalized linear models. Sage Publications, Los Angeles, California, USA.

Gama-Rodrigues, A., M. Sales, P. Silva, N. Comerford, W. Cropper, and E. Gama-Rodrigues. 2014. An exploratory analysis of phosphorus transformations in tropical soils using structural equation modeling. Biogeochemistry 118:453–469.

Gelman, A., and J. Hill. 2007. Data analysis using regression and multilevel/hierarchical models. Cambridge University Press, Cambridge, UK.

Gotelli, N. J., and A. M. Ellison. 2004. Primer of ecological statistics. Sinauer, Sunderland, Massachusetts, USA.

Grace, J. B. 2015. Taking a systems approach to ecological systems. Journal of Vegetation Science 26:1025–1027.

Grace, J. B., P. B. Adler, S. W. Harpole, E. T. Borer, and E. W. Seabloom. 2014. Causal networks clarify productivity–richness interrelations, bivariate plots do not. Functional Ecology 28:787–798.

Grace, J. B., L. K. Allain, H. Q. Baldwin, A. G. Billock, W. R. Eddleman, A. M. Given, C. W. Jeske, and R. Moss. 2005. Effects of prescribed fire in the coastal prairies of Texas. USGS Open-File Report 2005-1287. U.S. Geological Survey, Reston, Virginia, USA.

Grace, J. B., and K. A. Bollen. 2005. Interpreting the results from multiple regression and structural equation models. Bulletin of the Ecological Society of America 86:283–295.

Grace, J. B., D. R. Schoolmaster Jr., G. R. Guntenspergen, A. M. Little, B. R. Mitchell, K. M. Miller, and E. W. Schweiger. 2012. Guidelines for a graph-theoretic implementation of structural equation modeling. Ecosphere 3:art73.

Greene, W. H. 2012. Econometric analysis. Pearson Education, New York, New York, USA.

Greenland, S., M. Maclure, J. J. Schlesselman, C. Poole, and H. Morgenstern. 1991. Standardized regression coefficients: a further critique and review of some alternatives. Epidemiology 2:387–392.

Greenland, S., J. M. Robins, and J. Pearl. 1999. Confounding and collapsibility in causal inference. Statistical Science 14:29–46.

Hellevik, O. 2009. Linear versus logistic regression when dependent variable is a dichotomy. Quality & Quantity 43:59–74.

Hilborn, R., and M. Mangel. 1997. The ecological detective: confronting models with data. Princeton University Press, Princeton, New Jersey, USA.

Jing, X., N. J. Sanders, Y. Shi, H. Chu, A. T. Classen, K. Zhao, L. Chen, Y. Shi, Y. Jiang, and J. S. He. 2015. The links between ecosystem multifunctionality and above-and belowground biodiversity are mediated by climate. Nature Communications 6: p.ncomms9159.

Kéry, M., and M. Schaub. 2012. Bayesian population analysis using WinBUGS. Academic Press, New York, New York, USA.

Lamb, E. G., K. L. Mengersen, K. J. Stewart, U. Attanayake, and S. D. Siciliano. 2014. Spatially explicit structural equation modeling. Ecology 95:2434–2442.

Laughlin, D. C., and S. R. Abella. 2007. Abiotic and biotic factors explain independent gradients of plant community composition in ponderosa pine forests. Ecological Modelling 205:231–240.

Lefcheck, J. S. 2016. piecewiseSEM: piecewise structural equation modelling in r for ecology, evolution, and systematics. Methods in Ecology and Evolution 7:573–579.

Lefcheck, J. S., J. E. Byrnes, F. Isbell, L. Gamfeldt, J. N. Griffin, N. Eisenhauer, M. J. Hensel, A. Hector, B. J. Cardinale, and J. E. Duffy. 2015. Biodiversity enhances ecosystem multifunctionality across trophic levels and habitats. Nature Communications 6:6936.

Lefcheck, J. S., and J. E. Duffy. 2015. Multitrophic functional diversity predicts ecosystem functioning in experimental assemblages of estuarine consumers. Ecology 96:2973–2983.

Legendre, P., and L. F. Legendre. 2012. Numerical ecology. Elsevier, New York, New York, USA.

Lindley, D. V. 2002. Seeing and doing: the concept of causation. International Statistical Review 70:191–214.

Long, J. S. 1997. Regression models for categorical and limited dependent variables. Sage Publications, Los Angeles, California, USA.

MacKinnon, D. P., C. M. Lockwood, C. H. Brown, W. Wang, and J. M. Hoffman. 2007. The intermediate endpoint effect in logistic and probit regression. Clinical Trials 4:499–513.

Maddala, G. 1983. Limited dependent and qualitative variables in econometrics. Cambridge University Press, Cambridge, UK.

Marcot, B. G., R. S. Holthausen, M. G. Raphael, M. M. Rowland, and M. J. Wisdom. 2001. Using Bayesian belief networks to evaluate fish and wildlife population viability under land management alternatives from an environmental impact statement. Forest Ecology and Management 153:29–42.

Matson, P. A., and M. D. Hunter. 1992. Special feature: the relative contributions to top-down and bottom-up forces in population and community ecology. Ecology 73:723.

Mazancourt, C., F. Isbell, A. Larocque, F. Berendse, E. Luca, J. B. Grace, B. Haegeman, H. Wayne Polley, C. Roscher, and B. Schmid. 2013. Predicting ecosystem stability from community composition and biodiversity. Ecology Letters 16:617–625.

McCullagh, P., and J. Nelder. 1989. Generalised linear modelling. Chapman and Hall, New York, New York, USA.

McKelvey, R. D., and W. Zavoina. 1975. A statistical model for the analysis of ordinal level dependent variables. Journal of Mathematical Sociology 4:103–120.

Menard, S. 1995. Applied logistic regression analysis: SAGE university series on quantitative applications in the social sciences. Sage Publications, Los Angeles, California, USA.

Menard, S. 2010. Logistic regression: from introductory to advanced concepts and applications. Sage Publications, Los Angeles, California, USA.

Menard, S. 2011. Standards for standardized logistic regression coefficients. Social Forces 89:1409–1428.

Mitchell, R. J. 2001. Path analysis. Pages 217–234 *in* S. M. Scheiner and J. Gurevitch, editors. Design and analysis of ecological experiments. Oxford University Press, Oxford, UK.

Mood, C. 2010. Logistic regression: why we cannot do what we think we can do, and what we can do about it. European Sociological Review 26:67–82.

Nelder, J. A., and R. W. M. Wedderburn. 1972. Generalized linear models. Journal of the Royal Statistical Society Series A 135:370–384.

Pearl, J. 1988. Probabilistic reasoning in intelligent systems. Morgan Kaufmann Publishers, San Francisco, California, USA.

Pearl, J. 2009. Causality. Cambridge University Press, Cambridge, UK.

Pedhazur, E. 1997. Multiple regression in behavioral research: explanation and prediction. Third edition. Wadsworth Publishing, Belmont, California, USA.

Pourret, O., P. Naïm, and B. Marcot. 2008. Bayesian networks: a practical guide to applications. John Wiley & Sons, New York, New York, USA.

Quinn, G. P., and M. J. Keough. 2002. Experimental design and data analysis for biologists. Cambridge University Press, Cambridge, UK.

R Core Team. 2017. R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Royle, J. A., and R. M. Dorazio. 2008. Hierarchical modeling and inference in ecology: the analysis of data from populations, metapopulations and communities. Academic Press, New York, New York, USA.

Scherber, C., N. Eisenhauer, W. W. Weisser, B. Schmid, W. Voigt, M. Fischer, E.-D. Schulze, C. Roscher, A. Weigelt, and E. Allan. 2010. Bottom-up effects of plant diversity on multitrophic interactions in a biodiversity experiment. Nature 468:553–556.

Shipley, B. 2000. Cause and correlation in biology: a user's guide to path analysis, structural equations and causal inference. Cambridge University Press, Cambridge, UK.

Siemann, E., and W. E. Rogers. 2006. Recruitment limitation, seedling performance and persistence of exotic tree monocultures. Biological Invasions 8: 979–991.

Tukey, J. W. 1954. Causation, regression, and path analysis. Chapter 3. Pages 35–66 *in* O. Kempthorne, T. A. Bancroft, J. W. Gowen, and J. L. Lush, editors. Statistics and mathematics in biology. Iowa State College Press, Ames, Iowa, USA.

Turner, M. E., and C. D. Stevens. 1959. The regression analysis of causal paths. Biometrics 15: 236–258.

Veltman, C. J., S. Nee, and M. J. Crawley. 1996. Correlates of introduction success in exotic New Zealand birds. American Naturalist 147:542–557.

Winship, C., and R. D. Mare. 1984. Regression models with ordinal variables. American Sociological Review 49:512–525.

Wright, S. 1921. Correlation and causation. Journal of Agricultural Research 20:557–585.

Zuur, A., E. N. Ieno, and G. M. Smith. 2007. Analyzing ecological data. Springer Science & Business Media, New York, New York, USA.

Supporting Information

Additional Supporting Information may be found online at: http://onlinelibrary.wiley.com/doi/10.1002/ecs2.2283/full